

**MICRO-REVISÕES GERADAS POR USUÁRIOS:  
CARACTERIZAÇÃO E PREDIÇÃO DE  
POPULARIDADE**



MARISA AFFONSO VASCONCELOS

**MICRO-REVISÕES GERADAS POR USUÁRIOS:  
CARACTERIZAÇÃO E PREDIÇÃO DE  
POPULARIDADE**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: JUSSARA MARQUES DE ALMEIDA GONÇALVES

Belo Horizonte  
Fevereiro de 2015



MARISA AFFONSO VASCONCELOS

**USER GENERATED MICRO REVIEWS:  
CHARACTERIZATION AND POPULARITY  
PREDICTION**

Thesis presented to the Graduate Program  
in Computer Science of the Universidade  
Federal de Minas Gerais in partial fulfill-  
ment of the requirements for the degree of  
Doctor in Computer Science.

ADVISOR: JUSSARA MARQUES DE ALMEIDA GONÇALVES

Belo Horizonte

February 2015

© 2015, Marisa Affonso Vasconcelos.  
Todos os direitos reservados.

Vasconcelos, Marisa Affonso

V331u User generated micro reviews: characterization and popularity prediction / Marisa Affonso Vasconcelos. — Belo Horizonte, 2015  
xxv, 166 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais — Departamento de Ciência da Computação

Orientadora: Jussara Marques de Almeida  
Gonçalves

1. Computação - Teses. 2. Redes sociais on-line - Teses. 3. Predição (Lógica) - Teses. 4. Comportamento do consumidor - Teses. I. Orientadora. II. Título.

CDU 519.6\*04(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS EXATAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

## FOLHA DE APROVAÇÃO

User generated micro-reviews: characterization and popularity prediction

### MARISA AFFONSO VASCONCELOS

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

*Jussara Marques de Almeida Gonçalves*  
PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora  
Departamento de Ciência da Computação - UFMG

*Fabício Benevenuto de Souza*  
PROF. FABRÍCIO BENEVENUTO DE SOUZA  
Departamento de Ciência da Computação - UFMG

*Luis da Cunha Lamb*  
PROF. LUIS DA CUNHA LAMB  
Departamento de Informática Teórica - UFRGS

*Marcos André Gonçalves*  
PROF. MARCOS ANDRÉ GONÇALVES  
Departamento de Ciência da Computação - UFMG

*Marco Antônio Pinheiro de Cristo*  
PROF. MARCO ANTÔNIO PINHEIRO DE CRISTO  
Departamento de Ciência da Computação - UFAM

*Virgílio Augusto Fernandes Almeida*  
PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA  
Departamento de Ciência da Computação - UFMG

*Wagner Meira Júnior*  
PROF. WAGNER MEIRA JÚNIOR  
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 02 de fevereiro de 2015.





*Aos meus pais, Maria Aparecida e Antônio, aos meus irmãos, Mariana e Daniel, a minha avó Hercília, a amiga Vanessa Vidal e a todos aqueles que acreditaram nesse trabalho.*

*“Reconhece a queda.  
E não desanima.  
Levanta, sacode a poeira.  
E dá a volta por cima.”  
Paulo Vanzolini, “Volta por Cima”*



# Acknowledgments

À Prof. Jussara Almeida, orientadora da tese, agradeço o seu apoio, disponibilidade, paciência e dedicação. Obrigada pelos seus ensinamentos, pela confiança e amizade ao longo deste período. Ter sido sua orientada foi uma experiência extremamente rica. Ao Prof. Marcos Gonçalves pela colaboração, dedicação e incentivo.

Eu gostaria de agradecer a todos os professores do PPGCC, em particular aos professores Fabrício Benevenuto, Virgílio Almeida e Wagner Meira Jr. por estar sempre dispostos a me ouvir e ajudar. Eu gostaria de agradecer aos vários funcionários e alunos do PPGCC que sempre estiveram disponíveis para me auxiliar. Tenho muito orgulho de ter feito parte de um programa de pós-graduação de tamanha excelência e qualidade.

Aos amigos do laboratório CAMPS que sempre estiveram disponíveis para me auxiliar no que eu precisasse e em tornar essa caminhada mais divertida. Obrigada Giovanni Comarela, Gabriel Magno, Tiago Rodrigues, Geraldo Franciscani, João Pesce, Rafael Ottoni, Matheus Santos, Evandro Cunha, Diego Las Casas, Gustavo Rauber, Emanuel Vianna, Diego Saez-Trumper e Felipe Moraes. À amiga Tatiana Pontes pela colaboração e por estar sempre ao meu lado compartilhando as dificuldades e conquistas dessa caminhada. Ao amigo Saulo Ricci por ser imprecidível para a realização desse trabalho e pela amizade. Ao amigo Daniel Hasan pelas discussões que contribuíram para o sucesso desse trabalho.

À amiga Vanessa Vidal que insistiu que eu seguisse as minhas intuições e meus talentos no desenvolvimento desse trabalho.

Aos meus pais, irmãos e minha avó que torceram pelo meu sucesso. Obrigada pelo amor, carinho e incentivo dado durante todo o processo.

À FAPEMIG e ao CNPq pelo apoio financeiro.

A todos aqueles que de alguma forma colaboraram para a realização deste trabalho.



*“Life can only be understood backwards; but it must be lived forwards.”*

(Søren Kierkegaard)



# Resumo

Desde a popularização da Web 2.0, as pessoas se tornaram cada vez mais engajadas ao expressar suas opiniões através de revisões sobre produtos e serviços. Como outros tipos de conteúdo gerado pelo usuário, revisões on-line vêm em várias formas, tamanhos e qualidades. Tal variabilidade na qualidade é particularmente notória em revisões textuais produzidas em aplicativos móveis, geralmente chamadas de micro-revisões ou *tips*, devido à sua concisão inerente. Em um ambiente abundante de conteúdo, ser capaz de estimar a utilidade de uma (micro-) revisão on-line, e finalmente, prever a sua popularidade futura entre os usuários, com precisão e o mais cedo possível, pode ser muito benéfico para os métodos de filtragem e recomendação de conteúdo, auxiliando os usuários a encontrar revisões valiosas e fornecendo um *feedback* rápido a empresários e futuros clientes.

Nesse contexto, investigamos como os usuários exploram micro-revisões, focando, particularmente, nas *tips* do Foursquare, um tipo cada vez mais popular de revisão cujo elevado grau de informalidade e concisão oferece dificuldades extras na concepção de métodos de predição efetivos. Usando dados coletados do Foursquare, investigamos também como a popularidade da *tip*, estimada pelo número de vezes que a *tip* recebeu um *like* de um usuário, evolui ao longo do tempo e quais os fatores que podem ser combinados para desenvolver um modelo para prever a popularidade da *tip* em um dado instante no futuro. Finalmente, desenvolvemos soluções para duas diferentes tarefas de predição: predição do ranking de popularidade de um conjunto de *tips* e a predição do nível de popularidade que uma *tip* em particular irá alcançar. Resultados experimentais mostram que um conjunto multidimensional de variáveis previsoras, que considera atributos do usuário que postou a *tip* e do *venue* onde ela foi postada, leva a resultados mais precisos do que a utilização de cada um desses conjuntos isoladamente. Além disso, os modelos, quando aplicados às *tips* do Foursquare, são também mais robustos que os modelos do estado da arte para predição de popularidade, já que nossos modelos podem ser aplicados em qualquer *tip*, no momento da postagem ou após dela.

**Palavras-chave:** Micro-revisões, popularidade, predição, redes sociais, comportamento do usuário.



# Abstract

Since the popularization of the Web 2.0, people are becoming increasingly engaged expressing their opinions with reviews about products and services. As any other type of user-generated content, online reviews come in various forms, sizes and qualities. Such quality variability is particularly prominent in textual reviews produced on mobile apps, often called micro-reviews or tips, due to their inherent conciseness. In such content abundant environment, being able to estimate the helpfulness of an online (micro-)review, and ultimately predict its future popularity among users as accurately and early as possible, can greatly benefit content filtering and recommendation methods, helping users find valuable reviews and providing quick feedback to business owners and future customers.

In this context, we investigate how users exploit micro-reviews, focusing particularly on Foursquare tips, an increasingly popular type of review whose high degree of informality and brevity offers extra difficulties to the design of effective prediction methods. Using collected data from Foursquare, we also investigate how tip popularity, given by the number of times the tip received a “like” from a user, evolves over time and which factors impact this popularity evolution. Then, we explore how these factors can be combined to develop models to predict tip popularity at a given point in time in the future. We develop solutions to two different prediction tasks: predicting the popularity ranking of a set of tips and predicting the popularity level a particular tip will achieve. Our experimental results show that a multidimensional set of predictor variables, which considers features of both the user who posted the tip and the venue where it was posted, leads to more accurate results than using each set of features in isolation. Our models, when applied to Foursquare tips, are also more robust than state-of-the-art popularity prediction methods, as they can be applied to any tip, at or after posting time.

**Keywords:** micro-reviews, popularity, prediction, social networks, user behavior.



# List of Figures

3.1	Screenshot of a Foursquare Venue Page. . . . .	35
4.1	User Tipping Activity on Foursquare. . . . .	42
4.2	Number of Friends and Followers and Number of Mayorships per User. . .	44
4.3	Fraction of Likes Received from the User's Social Network (Friends and Followers). . . . .	45
4.4	Visiting and Tipping Activities per Venue. . . . .	46
4.5	Distributions per Venue Category. . . . .	47
4.6	Content Features of Foursquare Tips and Yelp Reviews. . . . .	48
4.7	Correlation between User Attributes (top 3% users with largest percentages of tips with links). . . . .	51
4.8	User Profiles: Attribute Distributions. . . . .	53
4.9	Venue Category Distributions. . . . .	54
4.10	Words Commonly Used in Users' Tips. . . . .	56
4.11	Correlation between User Attributes (only users with at least 10 tips). . .	57
4.12	Degree Distribution of the User Network (log scale). . . . .	60
4.13	Distribution of Tip Popularity over Time. . . . .	68
4.14	Distribution of Percentage of Likes Received During the First Month after Posting Time. . . . .	70
4.15	Distribution of Time Until x% of Total Likes are Received for the Most Popular Tips (G1) . . . . .	71
4.16	Social vs. Non Social Likes: Distribution of Percentage of Likes Received over Time. . . . .	72
4.17	Cumulative Distributions of Popularity Peak for Most Popular Tips (G1). .	74
5.1	Monitoring Time Scheme. . . . .	80
5.2	Temporal Data Split into Train and Test Sets. . . . .	89

5.3	Correlations between the top-10 most popular tips at time $t_r$ and at time $t_r + \delta$ ( $\delta$ in months).	92
5.4	Effectiveness of Ranking for Varying Target Time $t_r + \delta$ : NY Scenario (Avg and 95% Confidence Intervals).	93
5.5	Tips Ranking Example.	94
5.6	Effectiveness of Ranking for Varying Target Time $t_r + \delta$ : NY Food Scenario (Avg and 95% Confidence Intervals)	95
5.7	Effectiveness of Ranking when Removing One Feature at a Time: NY Scenario (Avg and 95% Confidence Intervals for All Considered Days)	97
5.8	Effectiveness of Ranking When Using Only 4 Features for $\delta = 1$ month (Avg and 95% Confidence Intervals)	98
6.1	Monitoring Time Scheme.	104
6.2	Chronological Split of Training and Test Sets: Sliding Windows Over Time.	112
6.3	Macro-Average Results for Two Popularity Levels.	115
6.4	Macro-Average Results for Three Popularity Levels.	116
6.5	Results for Tips in the Low Popularity Category.	118
6.6	Results for Tips in the High Popularity Category.	119
6.7	Distribution of Most Important User Feature for Predicting a Tip's Popularity Level.	122
6.8	Distributions of the Most Important Venue Features for Predicting a Tip's Popularity Level.	123
6.9	Distribution of the Most Important Content Feature for Predicting a Tip's Popularity Level.	123
6.10	Macro-Average Precision and Recall for OLS using one feature at time	124
6.11	Macro-Average Results for OLS After Removing Each Collinear Feature.	126
6.12	Recall for OLS When Removing One Feature at a Time.	127
6.13	Macro-Average Results for Various Monitoring Times $\varepsilon$ ( $\delta = 1$ month).	129
6.14	Macro-Average Results for Various Target Times $\delta$ ( $\varepsilon=0$ ).	132
6.15	Model Accuracy in the Training Set (Each point is a Sector in the 10-Dimensional Space Defined by the Top-10 Features).	140
6.16	Model Accuracy in the Testing Set (Each Point is a Sector in the 10-Dimensional Space Defined by the Top-10 Features).	141

# List of Tables

3.1	Summary of Our Venue Dataset. . . . .	38
3.2	Summary of Our User Dataset . . . . .	38
4.1	Summary of Users Tipping Activities. . . . .	43
4.2	Summary of Visiting and Tipping Activities at Venues. . . . .	45
4.3	Summary of Tip Textual Characteristics. . . . .	48
4.4	Summary of User Attributes Across Clusters. . . . .	52
4.5	Results of the Manual Inspection of a Sample of Users from Each Cluster. . . . .	56
4.6	Summary Statistics for User Influence Networks per Venue Category as well as for All Categories (General). . . . .	59
4.7	Kendall $\tau$ Correlation Values Between Rankings Lists. . . . .	64
4.8	Top-5 Most Influential Users Overall and per Venue Category According to Each Method. . . . .	65
4.9	Distribution of Likes for Groups of Tips . . . . .	68
4.10	Rich-get-Richer Analysis: Coefficients $\alpha$ (and 95% Confidence Intervals) and $R^2$ of Linear Regressions from (log) Popularity in $t_r$ to (log) Popularity $t_r + \delta$ . . . . .	75
5.1	Tip's Syntactic Content Features. . . . .	86
5.2	Complete Set of Features for Tip Popularity Prediction . . . . .	88
5.3	Overview of Datasets and Scenarios of Evaluation . . . . .	89
5.4	Features Ranked by Information Gain . . . . .	96
6.1	Distribution of Candidates for Prediction Across Different Popularity Levels	103
6.2	Complete Set of Features for Tip Popularity Level Prediction . . . . .	111
6.3	Confusion Matrix for a Three-Class Classification Task . . . . .	113
6.4	Examples of Confusion Matrices for a Two-Class Classification Task . . . . .	117
6.5	Features Ranked by Information Gain. . . . .	121
6.6	Features with High Collinearity with at Least One Other Feature. . . . .	125

6.7	Macro-Average Results of Models that Use Early Popularity Measurements (only tips with at least 1 like, $\varepsilon=168$ hours, $\delta=1$ month). . . . .	130
6.8	Geographical Model Specialization: Macro-Average Results . . . . .	134
6.9	Categorical Model Specialization: Macro-Average Results. . . . .	136

# Contents

<b>Acknowledgments</b>	<b>xi</b>
<b>Resumo</b>	<b>xv</b>
<b>Abstract</b>	<b>xvii</b>
<b>List of Figures</b>	<b>xix</b>
<b>List of Tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Basic Concepts . . . . .	4
1.2 Dissertation Goals and Contributions . . . . .	4
1.3 Challenges . . . . .	8
1.4 Organization of this Dissertation . . . . .	9
<b>2 Literature Review</b>	<b>11</b>
2.1 Information Credibility . . . . .	12
2.2 Predicting the Quality of User Generated Content . . . . .	13
2.2.1 Helpfulness of Online Reviews . . . . .	14
2.2.2 Opinion Mining . . . . .	18
2.2.3 Spam Detection . . . . .	19
2.3 Analysis of Online Content Popularity . . . . .	20
2.3.1 Popularity Prediction Models . . . . .	21
2.3.2 Information Propagation and Social Influence Models . . . . .	24
2.4 Analyses of Location-Based Social Networks . . . . .	27
2.5 Summary . . . . .	30
<b>3 Foursquare: Case Study</b>	<b>33</b>
3.1 Foursquare: Key Elements and Features . . . . .	33

3.2	Measurement Methodology . . . . .	36
3.2.1	Crawling Methodology . . . . .	36
3.2.2	Venue Dataset (Dataset 1) . . . . .	37
3.2.3	User Dataset (Dataset 2) . . . . .	38
3.3	Summary . . . . .	40
<b>4</b>	<b>Tipping Activity on Foursquare: Characterization and User Influence</b>	<b>41</b>
4.1	Characterization of Tipping Activity . . . . .	41
4.1.1	User Analysis . . . . .	42
4.1.2	Venue Analysis . . . . .	45
4.1.3	Tip Analysis . . . . .	47
4.2	User Profiles . . . . .	49
4.2.1	Suspicious Behavior . . . . .	49
4.2.2	Uncovering User Profiles . . . . .	51
4.3	User Influence . . . . .	57
4.3.1	User Behavioral Patterns . . . . .	57
4.3.2	User Influence Network . . . . .	58
4.3.3	Measuring User Influence . . . . .	60
4.4	Dynamics of Tip Popularity Evolution . . . . .	67
4.4.1	Popularity Evolution . . . . .	69
4.4.2	The Role of the Social Network . . . . .	71
4.4.3	Popularity Peak . . . . .	72
4.4.4	The Rich-Get-Richer Phenomenon . . . . .	73
4.5	Summary . . . . .	76
<b>5</b>	<b>Predicting the Popularity Ranking of a Set of Tips</b>	<b>79</b>
5.1	Popularity Prediction Task . . . . .	80
5.2	Ranking Strategies . . . . .	81
5.3	Tip Features . . . . .	82
5.3.1	User Features . . . . .	83
5.3.2	Venue Features . . . . .	84
5.3.3	Tip's Content Features . . . . .	86
5.4	Experimental Setup . . . . .	87
5.5	Experimental Results . . . . .	91
5.5.1	Ranking Stability . . . . .	91
5.5.2	Prediction Results . . . . .	92
5.5.3	Experiments Removing Features . . . . .	94



5.6	Summary . . . . .	98
<b>6</b>	<b>Predicting the Popularity Level of a Tip</b>	<b>101</b>
6.1	Popularity Levels . . . . .	102
6.2	Tip Popularity Prediction: Formal Definition . . . . .	103
6.3	Prediction Methods . . . . .	104
6.3.1	Support Vector Machines (SVM) . . . . .	105
6.3.2	Ordinary Least Square Regression (OLS) . . . . .	106
6.3.3	Support Vector Regression (SVR) . . . . .	106
6.4	Tip Features . . . . .	106
6.5	Methodology Evaluation . . . . .	110
6.5.1	Experimental Setup . . . . .	110
6.5.2	Evaluation Metrics . . . . .	113
6.6	Experimental Results: Predictions at Posting Time . . . . .	114
6.6.1	Analysis of the Groups of Features . . . . .	114
6.6.2	Feature Importance . . . . .	120
6.7	Experimental Results: Other Prediction Scenarios . . . . .	127
6.7.1	Prediction Results Varying the Monitoring Period $\varepsilon$ . . . . .	128
6.7.2	Prediction Results Varying Target Prediction Window $\delta$ . . . . .	131
6.8	Model Specialization . . . . .	133
6.8.1	City-Based Model Specialization . . . . .	134
6.8.2	Category-Based Model Specialization . . . . .	135
6.9	Summary . . . . .	138
<b>7</b>	<b>Conclusions and Future Work</b>	<b>143</b>
7.1	Main Conclusions . . . . .	144
7.2	Directions for Future Work . . . . .	145
7.3	Publications . . . . .	148
	<b>Bibliography</b>	<b>149</b>



# Chapter 1

## Introduction

In the last years, we have seen an increasing amount of data, especially on personal interests and activities, shared on the Web. This was possible thanks to the success of online social networks (OSNs) which not only have enhanced the connectivity among people but also have allowed the dissemination and visibility of user-generated content (UGC), previously restricted to some niches. Particularly, users are no longer only the target of information on products and services published in advertising campaigns but also often act as media and content producers, commenting on and evaluating previous experiences. Usually, these recommendations are posted as online reviews of products, services, and businesses. Previously, people used to share their opinions by word-of-mouth, orally to each other or through anonymous comments deposited in suggestion boxes. Nowadays, the social Web allows people to interact and freely share opinions on products, services or companies in real-time and in large scale.

In fact, the number of product and service reviews available online has been increasing at many retailer websites such as Amazon<sup>1</sup>, and Walmart<sup>2</sup>, as well as on specialized review websites such as Epinions<sup>3</sup>, TripAdvisor<sup>4</sup> and Yelp<sup>5</sup>. More and more people base their buying decisions on online reviews written by others [Chen et al., 2004]. Indeed, several studies and surveys have found evidence that online reviews affect product sales [Chevalier and Mayzlin, 2006; Ante, 2009; Li and Hitt, 2008]. Moreover, these user generated reviews nourish the relationship between customers and real businesses, offering constructive criticism and possibly a competitive advantage to business owners. Reviews also work as a benchmark of their offered products and

---

<sup>1</sup><http://www.amazon.com>

<sup>2</sup><http://www.walmart.com>

<sup>3</sup><http://www.epinions.com>

<sup>4</sup><http://www.tripadvisor.com/>

<sup>5</sup><http://www.yelp.com>

services, and may be used to drive more effective marketing strategies [Chen and Xie, 2008].

The number of reviews on a single product or service available may be large, varying greatly in quality. Some reviews may contain spam or misleading and fake information [Lappas, 2012; Lin et al., 2014], which may make it hard for users to find helpful reviews. To support that task, many websites allow users to evaluate reviews, by voting in their helpfulness. Unfortunately, this feedback is usually very sparse [O’Mahony and Smyth, 2009]. Moreover, ranking reviews based solely on the helpfulness votes received may not be useful for promoting recently posted reviews, with few or no votes, which, regardless of their potential helpfulness, are doomed to be outranked by older reviews that have already received more votes. Thus, those reviews may never gain visibility.

This problem has already inspired a series of studies attempting to automatically predict the quality and helpfulness of a review [Kim et al., 2006; Liu et al., 2007; T.Ngo-Ye and Sinha, 2014], estimated by the number of people who found the review useful. The same metric can be also used as a measure of the review’s popularity as it provides a lower bound on the number of people who actually read the review. In order to compare with the state-of-the-art models, we here use the terms quality, utility, helpfulness and popularity interchangeably.

Both user and business owners can benefit from such predictions of review popularity, as they can drive the design of automatic review filtering and recommendation schemes as well as review ranking methods, which in turn can help users find potentially more valuable reviews (or reviews that are likely to draw more attention in the future). Predicting the potential popularity of a review can also stimulate reviewers to post higher quality reviews as a rapid feedback can be provided to review authors. Similarly, such predictions can also offer valuable feedback to business owners who are able to more quickly identify (and fix) aspects of their services or products that may affect revenues most, since potentially more popular reviews may contain information about how a product is seen by a larger fraction of the customers.

However, previous efforts to predict the popularity of online reviews focused on longer, more verbose and formally structured reviews, such as those present in systems like Amazon and TripAdvisor, often exploiting textual and content related features (e.g., review length, readability) [Kim et al., 2006; Liu et al., 2007; T.Ngo-Ye and Sinha, 2014]. Yet, with the diffusion of smartphones, new services were created targeting mainly social networking users who spend most of their time accessing information through mobile apps. In this environment, the communication is usually briefer mainly because of the limited amount of information that can be displayed on the mobile

screen. This limitation may have also influenced the creation of new review services (Foursquare<sup>1</sup>, Google+ Local<sup>2</sup>), and the expansion of traditional desktop services to the mobile environment (e.g., Yelp, Trip Advisor). In these services, users write *micro-reviews* or *tips*, which are typically much more concise (e.g., up to 200 characters), often written while the information is still fresh in the user’s mind, and may contain much more subjective and informal content, varying from a narrow recommendation (“You must try the apple pie.”) to a general warning (“Stay away from this place”). In this dissertation, we focus on this special type of review, the micro-review (or tip)<sup>3</sup>, popularized by Foursquare.

Unlike several traditional reviews systems, where users are allowed to assign unhelpfulness signals to a review, tips are rated by other users by simply clicking on a “like” mark. The number of “likes” received by a tip can then be seen as an estimate of its helpfulness or popularity. However, the lack of a “like” does not imply that a tip was not helpful or interesting, as it may not have been seen by any user. Moreover, utility is an abstract concept that can be more comprehensive than the fact that a user has given a “like”. This further contributes to make the automatic prediction of tip popularity much harder than in systems that offer a rating scale (e.g., 1 to 5), such as Yelp and Epinions. Moreover, tips are also different from product reviews since they may be active for much longer periods (e.g., a restaurant or an airport tip) and some product reviews may become inactive when a new version of a the same product is released.

Finally, the problem of predicting tip popularity is also inherently different from other efforts to forecast the attention to other user-generated content, such as tweets [Suh et al., 2010; Hong et al., 2011], news posts [Bandari et al., 2012], videos [Borghol et al., 2012; Brodersen et al., 2012; Figueiredo et al., 2014b] or questions in a forum [Anderson et al., 2012; Li et al., 2012] which exploited mainly aspects related to the user who posted the information or the content itself (e.g., category). Unlike tweets, news, videos and questions, tips are associated with specific venues, and tend to be less ephemeral (particularly compared to news and tweets), as they remain associated with the venue (and thus visible to users) for a longer time.

---

<sup>1</sup><https://foursquare.com/>

<sup>2</sup><http://www.google.com/+/learnmore/local/>

<sup>3</sup>We will use the word tip and micro-review interchangeably in the text.

## 1.1 Basic Concepts

This research focus on the Foursquare micro-reviews, known as *tips*. Foursquare is the most popular location-based social network (LBSN) where users can share their current location with friends and followers through check-ins. Check-ins are performed by users via devices with GPS (Global Positioning System) when they are close to a specific (physical) location, named *venue*, which has an associated page in the system. Thus, venues are virtual places that are grouped into a large variety of categories such as airports, monuments or squares that represent real locations like John Kennedy International Airport (New York, United States), Taj Mahal (Agra, India) or Eiffel Tower (Paris, France).

In addition to check-ins, users can post tips about a given location on the corresponding venue’s page. Foursquare tips are limited to 200 characters and may contain more informal or subjective content. than longer reviews. They can be informative (“This place opened in 2002”), contain some recommendation (“Try the Fettuccine”) or even a report of users’ experiences (“Best place for tacos” or “Avoid lunch time”). Unlike Yelp reviews, for example, which are much more extensive and often written “after the fact”, Foursquare tips are usually written “during the moment” using a mobile device, and thus tend to be more brief and direct, avoiding many details about specific characteristics of the venue.

When visiting a venue’s page, users may assign a “like” mark to a previously posted tip in sign of agreement with the tip’s content and/or intention to visit the (physical) location with which the tip is associated. The aggregate number of likes is an estimation of the tip popularity and the same metric is used by Foursquare for ranking the tips in the venue’s page.

## 1.2 Dissertation Goals and Contributions

In this dissertation, we investigate how users exploit micro-reviews focusing particularly on how the popularity of such pieces of content evolves over time and which factors impact this evolution. Then, we explore how these factors can be combined to develop popularity prediction models.

Towards that end, we focus our study on micro-reviews on Foursquare, also called tips, which was responsible for the popularization of this feature [Vasconcelos et al., 2012b], and is currently one of the most popular systems that support micro-reviewing. We first characterize user interactions through tips, uncovering relevant user behavioral patterns which impacts tip popularity. The understanding of such patterns is important

as they can provide useful insights into which features should be exploited in the design of the prediction models as well as a better comprehension about the prediction results. We then develop regression and classification methods that exploit the most important factors to predict the future popularity of a tip as soon as it is posted, or at most, a short period afterwards. We develop solutions to two different prediction tasks: (1) predicting the popularity ranking of a set of tips, and (2) predicting the popularity *level* of a tip, whereas the former focuses on the relative (future) popularity of a group of tips, the latter tackles whether a particular tip will achieve a certain popularity level in the future.

A ranking of the most popular tips can be used to summarize a large set of tips focusing on the most popular ones for a scenario of interest. For example, a list of tips with a greater potential to become popular posted at any venue in the user home city. Moreover, to estimate the popularity of a single tip can benefit both users and venue owners. For instance, the system can offer different filtering strategies to the users based on the prediction, while venue owners can quickly react to opinions that may have a greater impact on decision making.

Specifically, our investigation tackles the following three questions:

1. *What are the most common user behavior and interaction patterns while using micro-reviews? How does the popularity of a tip evolve over time? How is it affected by the social network of the tip's author? To which extent does the rich-get-richer phenomenon impact the popularity evolution of tips?* (Chapter 4)

First, we present a characterization of user behavior on Foursquare. Our analyses were performed over a collected Foursquare dataset consisting of more than 1,5 million users, more than 6 million tips, and 5 million likes. This study consisted of two main phases. First, we characterized venues and users with respect to number of tips, number of likes and to-dos<sup>1</sup> as well as percentage of tips containing links (i.e., URLs or email addresses). We have also identified four groups of users with different tipping behavior, including one that is consistent with spamming.

Using a larger Foursquare dataset, containing over 10 million tips and 9 million likes posted by over 13,5 million users, we modeled the user interactions through tips and likes using a graph to identify the most influential users. To that end, we proposed a variation of the PageRank algorithm in which each arc of the graph is weighted by the number of tips posted by each node (user). Using the modified PageRank, we were able to identify users that were influential by consistently

---

<sup>1</sup>Users can also save tips in to-do lists.

receiving feedback in their tips. Moreover, we found users that were influential in a given venue category, which suggests that the category of the venue must be taken into account in the tip popularity prediction task.

Furthermore, we characterized how the popularity of different sets of tips evolves over time, and how it is affected by the social network of the user who posted the tip (its author). We observed that tips experience a very slow popularity evolution, compared to other types of user-generated content (UCG), such as news articles and photos. Moreover, the social network of the tip's author has an important influence on the tip popularity throughout its lifetime, but especially in earlier periods after posting. Compared to other types of UCG, such as YouTube videos, we observe a weaker presence of the rich-get-richer phenomenon in the popularity evolution of tips, suggesting that other factors, but the current popularity, may significantly impact the tip's future popularity.

2. *Which are the most important factors for predicting the popularity of Foursquare tips? How can we tackle the problem of predicting the future popularity of tips?* (Chapter 6.4 and 6)

We identified three important entities related to the Foursquare system that may impact a tip's popularity: the user who posted the tip, the venue where it was posted, and its content. We investigated the potential benefits from exploiting these aspects to predict the popularity of a tip (or a group of tips) will achieve at a future time.

To that end, we considered two different tasks. The first prediction task aimed at ranking a group of tips based on their predicted popularity at a given future time. We exploited a regression model using the aspects from the three most important entities as predictors. Moreover, we evaluated the stability of the tip popularity ranking over time, assessing to which extent the current popularity ranking of a set of tips can be used to predict their popularity ranking at a future time. We then found that the set of features used in our model can improve the prediction accuracy, given that enough training data is available.

The second prediction task is more challenging since it tackles the problem of predicting the popularity *level* of a single tip<sup>1</sup>. We addressed this problem by formalizing it as a classification task. Since over 80% of the tips received no like at all, great part of this dissertation is focused on predicting the popularity

---

<sup>1</sup>The tip popularity level is defined according to a range of values which will be defined in Chapter



of a tip at posting time or when there is no information about the tip current popularity. For that goal, we employed classification and regression methods along with an extended set of features related to the tip author, venue and content as predictors. We investigated the relative importance of each predictor variable, finding that features extracted from both the user and the venue are among the most important ones on Foursquare.

3. *To which extent can we improve prediction by monitoring the tip for a short period after posting? How do the prediction models behave as we predict further into the future? Can we improve prediction accuracy by building specialized models?* (Chapter 6)

The slow popularity evolution of a Foursquare tip also raises a question as to how robust our solutions are to long-term predictions. By monitoring the tips for a certain time after their creation, we expect to add to our models information about how its popularity is evolving, which may contribute to improve predictions. We also investigated how far into the future we can predict tip popularity with reasonable accuracy, that is, we analyzed how robust our prediction models are when we perform long-term predictions. Our intent is to analyze when our models become less accurate since we expect that the accuracy drops as the prediction are performed further into the future. We found significant improvements in prediction accuracy as we extend the initial monitoring time, although prediction accuracy may drop as we predict for more than 2 months ahead in the future.

We also investigated whether factors related to a specific geographic region (e.g., city) or category of venue impact how the popularity of a tip evolves over time. To that end, we built specialized prediction models using only tips posted in a specific city or in a specific venue category, and compared to which extent such models improve over the single general model. We found that model specialization does bring some (limited) improvements if performed at the city-level, whereas category-based specialization does not bring clear and consistent gains.

In summary, the key contributions of this dissertation are: (1) a solid understanding of how users exploit micro-reviews and the factors that impact the popularity of such content on Foursquare; (2) the design of cost-effective methods to predict the future popularity of micro-reviews. For content providers and system administrators such prediction methods can be exploited to improve their systems in, for example, the design of automatic review filtering, recommendation strategies and identification

of malicious behavior (e.g., spammers or detractors that can harm a user or business reputation). For marketers or advertisers the popularity prediction is valuable since a popular review may be tied directly to a product or service revenue which can be estimated ahead in time and negotiated by all parties involved. Furthermore, the system can provide incentives to users whose contributions increase the overall value of content on the site. More broadly, the knowledge uncovered in our study may help understanding how the dynamics of the user community in the target application works.

### 1.3 Challenges

The prediction of the popularity of micro-reviews poses several challenges.

- *New content type*: tips have inherent characteristics that distinguish them from other types of content and that might impact their popularity evolution. For example, tips are associated with specific venues, and thus are visible to all users that visit the venue, including those that are drawn to it by other reasons (e.g., other tips). Also, tips usually contain opinions that might interest others for much longer periods of time than other types of content such as news and tweets. Thus, tips may remain live in the system, attracting attention (and likes), for longer periods.
- *Content analysis*: most of the previous proposed models to automatically estimate the popularity or helpfulness of a review formulate the problem as classification or regression problem using observed features, i.e. textual or social features. Textual features are usually related to the structure, syntactic, readability and sentiment of the review's content. However, most of tips do not follow any formal structure (e.g., capitalizing the first letter, punctuations, truncated sentences, etc.) and they may present an informal vocabulary (word abbreviations, emoticons, slangs, etc.) [Thurlow and Brown, 2003; Grinter and Eldridge, 2003; Thelwall et al., 2010]. These variations cause problems since most content features rely on readability metrics that requires a good text structure. Moreover, typical sentiment analysis algorithms assume that the text is written on standard spelling and grammar, so the current algorithms are unlikely to work well on that scenario.
- *Feature selection*: popularity can be affected not only by the differences in tip content, but also because of a multitude of factors with complex and unknown interactions such as users, venues, the system interface and the rich-get-richer

effect. Thus, one of our main challenges is to formalize these factors and assess the impact that they have on tip popularity.

- *Data sparsity*: both tip and user interactions are extremely sparse, as will be shown in Chapter 4. Our analyses of the features related to the main entities – user, venue and tip’s content – revealed that most of them exhibit very large variability, with great concentration on few users, venues and tips. For instance, according to our analyses, 49% of the tips were posted by only 10% of the users. Such very skewed distribution of number of likes per tip brings technical challenges to the prediction task modeling [He and Garcia, 2009; Liu et al., 2009] (e.g., severe class imbalance when predicting the popularity level of tips). Moreover, over 80% of the tips used in our experiments received no like at all, which means that these tips have no information about their popularity. Thus, this limits the effectiveness of the state-of-the-art methods that uses early measurements, since they cannot be applied to these tips and shows the robustness of our model since we are able to perform predictions at posting time.
- *Model evaluation*: to evaluate our prediction models, our dataset has been chronologically split into training and test sets. However, this splitting schema has more restrictions than the cross-validation schema which may also cause the class imbalance problem.

## 1.4 Organization of this Dissertation

This dissertation is organized as follows. In Chapter 2, we survey the literature on four topics closely related to our present study, namely information credibility, prediction of review helpfulness and of popularity of online content as well as location-based social network. Chapter 3 introduces the main elements and features of Foursquare. It also describes our crawling methodology as well as a summary of the collected datasets. Next in Chapter 4, we present a characterization of tips usage aiming to identify relevant user profiles and obtaining valuable insights into prediction models. We also analyze the dynamics of tip popularity in this chapter. Chapter 6.4 describes how we addressed the first prediction task (the ranking task), and the features selected for the ranking experiments. Chapter 6 describes our investigation of the second prediction task (the classification task), and the models developed to predict the popularity level of a tip. We also analyze the impact in the accuracy of our models when varying monitoring time and prediction target time as well as the impact of model specialization.

Finally, Chapter 7 concludes this dissertation, presenting some directions for future work.

# Chapter 2

## Literature Review

In this dissertation, we study the popularity of micro-reviews. We use the number of likes received as our measure of the tip’s popularity. Moreover, the same measure can be seen as an estimation of the tip helpfulness or quality since it reflects the number of people who found a tip useful and indirectly assessed its quality. Thus, there are two groups of studies related to our problem: one aims at assessing the helpfulness or quality of reviews, whereas the other tackles the prediction of the popularity of online content.

Most previous work on assessing the helpfulness of reviews has typically focused on automatically determining the quality, helpfulness or utility of reviews using textual features. However, such features are more appropriate for longer and formally structured reviews. Micro-reviews are shorter than traditional reviews, usually having length constrained to around 200 characters in order to be published and read on a variety of platforms. This size constraint has led users to write reviews using non-standard textual artifacts (e.g., emoticons) and informal language [Bermingham and Smeaton, 2010].

Moreover, in some micro-reviews systems, such as Foursquare, the micro-reviews, known as tips, are rated by other users only by simply marking them as “liked”, as opposed to in other review systems where reviews are rated through rating or helpfulness votes. Likes are not as informative as ratings, since the lack of these marks cannot be seen as an unhelpfulness signal. These aforementioned factors make the prediction of the helpfulness or popularity of micro-reviews a challenging task. To the best of our knowledge, no previous study has tackled the popularity of micro-reviews, but there are several threads of related research which we will review in this chapter that have guided us during our model design.

We start by briefly discussing studies on information credibility in Section 2.1.

We conjecture that the helpfulness or popularity of a micro-review can be influenced by the perceived credibility of the reviewer. Based on this conjecture, some of our proposed features (see Chapter 6) are inspired by some of these credibility directives. Next, in Section 2.2, we discuss previous analyzes of the quality of various types of user generated content with particular focus on the assessment of the quality of online reviews. In this section we also briefly review other related efforts towards automatically detecting the polarity (positive, neutral, negative) of online reviews and detecting spam of fake reviews. Such studies can be considered complementary to this dissertation. For example, we use sentiment scores as content features exploited as input to our popularity prediction models (Chapter 6). Moreover, as a result of our characterization of user tipping activity, we were the first to uncover evidence of spamming activity on Foursquare (Chapter 4).

From the perspective of predicting the popularity of online content, we also survey recent work on popularity prediction models and on information propagation and social influence in Section 2.3. These proposed models are highly influenced by the target application and by the type of data (e.g., number of video views, number of retweets, number of digg votes) which makes unfeasible the creation of a generic prediction model [Tatar et al., 2011]. However, some of our analyses and proposed features are inspired and/or adapted from these previous work. Finally, we present previous analyses of location-based social networks (LBSNs) in Section 2.4.

## 2.1 Information Credibility

The Web 2.0 has empowered users to express their opinions by interacting with others through social networks and by publishing a wide variety of user-generated content such as blogs, online forums, product or service reviews, among others. However, not all information available on the Web is credible or comes from reputable sources. Credibility affects how customers perceive the quality of online services and influences their decision-making processes.

In one of the first studies on the source of information credibility, Hovland and Weiss [1951] designed an experiment in which news stories with identical content were presented to volunteers as coming from two different sources (i.e., high-credibility or low-credibility). They identified the perceived “expertise” and “trustworthiness” of the source as factors impacting credibility. Fogg et al. [2001] defined credibility as believability, applying the two dominant factors as expertise and trustworthiness to identify credibility on websites. They designed a questionnaire to determine which

factors have the greatest impact on user’s perceptions of credibility. They found that the evaluated web site attributes fall into seven dimensions: five of them increase perceptions of credibility (real-world feel, ease of use, expertise, trustworthiness, and tailoring) while the other two contribute with negative perceptions of credibility (the commercial implications of the site, and amateurism). This study was performed for websites, but we use these dimensions to guide our choice of some features used in our popularity prediction model.

Credibility has also been analyzed in the social media domain. For the task of exploring trending topics on Twitter, Castillo et al. [2011] studied the information credibility of topics defined by a set of tweets. They used Amazon Mechanical Turk to gather user judgments about the credibility of a tweet, and extracted the most relevant features from each topic. They defined a complex set of features over messages, users, topics and propagations which were used to build a classifier to automatically assess the level of credibility of the topic. Based on a credibility framework for blog post retrieval proposed by Rubin and Liddy [2006], Weerkamp and de Rijke [2012] defined two groups of credibility indicators, namely post-level (e.g., spelling, timeless, post length) and blog-level (e.g. regularity, expertise, comments) indicators. Concerning online reviews, several studies were developed to automatically assess their helpfulness considering various credibility indicators. Those studies are discussed in the next Section.

## 2.2 Predicting the Quality of User Generated Content

We start by discussing quality in different contexts, and then focus on online reviews, our target domain. Our research is inspired by several previous studies that focused on analyzing the quality of socially generated content, including the quality of Wikipedia articles [Dalip et al., 2011, 2014], video or news comments [Siersdorfer et al., 2010; Hsu et al., 2009; Chen et al., 2011], and user-contributed answers on community question answering (CQA) forums [Anderson et al., 2012; Li et al., 2012].

Dalip et al. [2011] used Support Vector Regression (SVR)[Drucker et al., 1997] to estimate the quality of articles in collaborative digital libraries (e.g., Wikipedia) using features related to the text structure, citation network and article revision history. The same authors also studied the impact of feature selection on a multi-view algorithm for assessing quality in collaborative encyclopaedias [Dalip et al., 2014]. In [Siersdorfer et al., 2014], the authors proposed a Support Vector Machine (SVM) based model to predict the acceptance by the user community of a comment posted on YouTube and

Yahoo! News. The proposed model uses a term-based representation of comments (TF-IDF or term frequency – inverse document frequency) to automatically classify them as likely to obtain a high overall rating or not. With a similar goal, Hsu et al. [2009] proposed an SVR-based model to rank comments posted by users on Digg based on their quality. They exploited features such as the comment posting time, the number of articles submitted, and comment length. Chen et al. [2011] focused on user reputation in comment rating environments (Yahoo! News and Yahoo! Buzz). They showed that the quality of a comment judged editorially is almost uncorrelated with the ratings that it receives, but can be predicted using standard text features (e.g., length, spelling, and readability scores). Closely related to our target problem of estimating the popularity of a (micro-)review is the problem of predicting if a question will have long lasting value. With that particular goal, Anderson et al. [2012] demonstrated that features that mapped the user activity related to a question (e.g., pageviews) within a short interval after it was posted can help predict the number of page views that a question will receive. Li et al. [2012] investigated the quality of questions in CQA services, defined by a combination of the following features: the number of tags-of-interest (reflecting the attractiveness of a question), the number of answers, and the amount of time for getting the best answer. They also proposed a mutual reinforcement-based label propagation algorithm to predict the quality of a question using features of the question’s text and of the askers profile. Finally, Momeni et al. [2013] developed a classifier for predicting useful comments on YouTube and Flickr exploiting, for that task, not only textual features, but also features that describe the author’s posting and social behavior, such as the number of links posted and the size of the author’s social network.

In this dissertation, we also apply regression methods used in some of those studies, particularly the SVR method. However, we applied these techniques in a novel context, using other sets of features, to automatically predict the popularity of micro-reviews.

### 2.2.1 Helpfulness of Online Reviews

We now turn to previous efforts to predict the quality (helpfulness or utility) of online reviews, which are more closely related to our work.

The quality of reviews can have a significant impact on purchase decisions for future customers. Many currently popular websites, such as Amazon, Epinions and Trip Advisor, provide mechanisms for users to give some feedback (or review) about products or services provided. The large amount and wide variability of quality of



the reviews available in some websites motivate the use of filtering, reputation and personalized recommendation mechanisms to help users to find useful reviews [Kim et al., 2006; Hsu et al., 2009; O’Mahony and Smyth, 2009]. Indeed, some websites, such as Amazon, allow users to indicate whether he/she finds a review helpful. These meta ratings help users filter relevant reviews more efficiently [Siersdorfer et al., 2010], and summarize a general opinion about a product or service. However, such type of feedback is still sparse, with many reviews, especially the most recent ones, failing to attract any feedback. This problem has inspired several research studies about automatic prediction of the quality of reviews. The task of assessing the quality [Liu et al., 2007; Lu et al., 2010; Yu et al., 2010], utility [Zhang and Varadarajan, 2006; Liu, 2010] or helpfulness [Kim et al., 2006; Zhang and Tran, 2008; O’Mahony and Smyth, 2009; Tsur and Rappoport, 2009; Korfiatis et al., 2012; Ngo-Ye and Sinha, 2012] of a review, is typically addressed by employing classification or regression-based solutions using a set of observed features, often textual features, as predictors, and the users’ votes as ground-truth.

Liu et al. [2007] identified three types of biases in the Amazon review ranking system. The first type was observed through an imbalance voting pattern, where users tend to value others’ opinions positively more often rather than negatively. The second one, the *rich-get-richer* effect, named as winner circle bias by the authors, was characterized by a larger amount of votes accumulated by the top reviews, while the third type (early bird bias) observed a clear trend that the earlier a review is posted, the more votes it will get. Moreover, the authors proposed an SVM-based approach to detect low-quality reviews, based on a manually determined ground-truth in accordance with proposed set of specifications for judging the quality of a review. However, the proposed model is based only on features suitable for longer (e.g., more verbose) and structured reviews (e.g., number of positive sentences, number of product features or brand names in the review).

Similarly, Danescu-Niculescu-Mizil et al. [2009] found that the perceived helpfulness of a review depends not only on its content, but also on the relation of its score to other scores. The authors investigated the dependency between helpfulness of product reviews from Amazon users, and concluded that users tend to consider reviews that agree with the average item rating as helpful. We attempt to capture similar trend in our models by using features related to the specific venue where the tip was posted, including characteristics of previously posted tips.

O’Mahony and Smyth [2009] proposed a classification-based approach to recommend helpful reviews in TripAdvisor, using features related to the user reviewing history, as well as the scores previously assigned to the hotels by the users. As an

extension of that work, the same authors considered structural features (e.g. ratio of uppercase characters, number of words, etc.) and readability features (e.g. scores indicating the difficulty in reading the text) to develop a classification technique to automatically identify the most helpful reviews [O’Mahony and Smyth, 2010]. Korfiatis et al. [2012] observed that review readability has a greater effect on the helpfulness ratio of a review than its length. We make use of some of the features defined by these three studies in our model, but we extend them by using social network information and features capturing the sentiment (or polarity) of the micro-reviews.

Other studies have considered estimating the helpfulness of a review using regression models. Basically, these studies aim at ranking the reviews by their helpfulness score (defined by the fraction between positive and negative votes) or at estimating their average rating, which is usually a real value between zero and five. Ghose and Ipeirotis [2007] studied the economic impact of online reviews using product reviews from Amazon. They proposed two mechanisms for ranking product reviews: a consumer oriented ranking mechanism which ranks the reviews according to their expected helpfulness, and a manufacturer-oriented ranking mechanism which ranks the reviews according to their expected effect on sales. Their experimental results showed that subjectivity analysis can give useful clues about the helpfulness of a review and about its impact on sales.

Zhang and Varadarajan [2006] found that syntactic features, such as number of proper nouns, comparatives and modal verbs extracted from the text reviews, are the most effective predictors for SVR and linear regression to predict utility of a product review. They observed that the perceived utility of a product review highly depends on its linguistic style. Kim et al. [2006] also used SVR to rank reviews according to their helpfulness, exploiting textual features such as length and the unigrams<sup>1</sup>, and the rating score given by the reviewers. They concluded that the review length and the number of stars in product rating were the most useful features for the regression model. Liu et al. [2008] proposed a non-linear regression model that incorporated the reviewers’ expertise, the review timeliness, and its writing style for predicting the helpfulness of movie reviews. They found that timeliness was a good predictor as the general helpfulness of a movie review declines for older reviews. Reviewer expertise was also found to be a useful feature, motivating the exploration of features that effectively describe user preference. The authors also used their proposed regression model as a classifier to retrieve only reviews having a predicted helpfulness higher than a certain threshold. In this dissertation, we also use regression methods to classify a tip into

---

<sup>1</sup>Unigrams are defined as the TF-IDF statistic of each word occurring in the review.

multiple levels of popularity based on their *predicted* popularity. However, the textual features proposed in [Liu et al., 2008] are once again more suitable for longer reviews and the timeliness factor is not observed in our Foursquare dataset.

T.Ngo-Ye and Sinha [2014] compared several text regression models for predicting the number of people who would find a review helpful, using datasets from Amazon and Yelp. Their proposed models exploit the words extracted from reviews and the reviewer engagement characteristics such as reputation, commitment and current activity as input features. The authors found that incorporating features capturing reviewer’s engagement and using a subset of unique review words selected by a dimension reduction method (Correlation-based Feature Selection) help predict review helpfulness. We used some of these reviewer’s engagement features such as frequency (number of reviews written before the current review) and monetary value (average number of helpfulness votes received for all her previous posted reviews) as predictors in our model. We also exploit the words micro-review’s content by taking their sentiment as predictors.

Tsur and Rappoport [2009] proposed an unsupervised method to rank book reviews according to their helpfulness. Their method works in stages: first, the algorithm identifies the terms that are less frequent, but contributes more information that is relevant to a specific product (dominant terms). These terms will constitute the core of a virtual optimal review. The reviews are then converted to a feature vector representation defined by the terms in the virtual core, and ranked according to their distances from the core. Martin and Pu [2014] developed a method to predict the helpfulness of a review using emotion features extract from the review text. The authors based their study on three product review datasets (Yelp, TripAdvisor, and Amazon) and used a general lexicon of emotion words (GALC [Scherer, 2005]) to extract words that convey emotions to the readers. They applied supervised classification algorithms (SVM, Naïve Bayes, and Random Forest) to estimate if a given review is helpful or not. The authors’ framework showed an improvement up to 9% when compared to models that use only text statistics or readability features.

Lu et al. [2010] exploited contextual information about the authors’ identities and social networks for improving the prediction of review quality in Ciao, a community review web site. They proposed a generic framework for incorporating social context information by adding regularization constraints to a text-based predictor. Their results show that adding social context as additional features can improve predictions significantly over text-based predictions, but none of them outperforms the combined model (using both textual and social features) when there is sufficient amount of training data. Hong et al. [2012] built a classification system to automatically assess review helpfulness based not only on textual features but also on features that represent the

user preferences. Such features capture if a review has attributes that the user prefers to know, whether the user who wrote the reviews was a buyer of the product, and the divergence of the polarity of the review from the mainstream opinion. Moghaddam et al. [2012] proposed a series of probabilistic factorization models to address the problem of personalized review quality prediction. Their models are based on the assumption that the observed review ratings depend on latent features of the reviews, reviewers, raters, and products. Lee and Choeh [2014] used neural networks to predict helpfulness of Amazon reviews. They found that characteristics of the product such as its list price, its sales rank, and textual characteristics of reviews such as the average number of words in a sentence, the number of words, and the number of one-letter words in the review are important for estimating helpfulness. Tang et al. [2013] analyzed various types of social context (i.e. author, rater, connection and preference contexts) to predict unknown helpfulness ratings of reviews using matrix factorization based methods. As in Moghaddam et al. [2012], the authors claim that the helpfulness of a review is not necessarily the same for all users. Moreover, the dual role of a user (i.e. author and rater) must be considered as separated contexts. We adapted some of these features to our domain.

In sum, those prior studies are based mostly on content features, which are suitable for more verbose and objective reviews, and thus may not be adequate for predicting the popularity of tips, which tend to be more concise and subjective. Moreover, previous studies did not address how helpfulness as perceived by users (or popularity) of reviews evolve over time, as we do in this dissertation.

### 2.2.2 Opinion Mining

Other related studies on customer reviews focus on opinion mining, particularly on classifying a review as positive or negative based on sentiments of the reviewers captured by the textual content. For example, Pang et al. [2002] analyzed several supervised classification approaches using different sets of features including unigrams, bigrams, adjectives and part-of-speech tags to classify the sentiment of movie reviews. Bermingham and Smeaton [2010] compared the performance of supervised classifiers (Naïve Bayes, SVM) and an unsupervised lexicon-based classifier on microblog data and reviews. These two studies motivated the use of the part-of-speech tags and the same lexicon (SentiWordNet) used in [Bermingham and Smeaton, 2010] as features in our popularity prediction models. Moraes et al. [2013b] evaluated the effectiveness of four methods for automatic polarity detection of Foursquare tips: SVM, Naïve Bayes, Maximum Entropy and one unsupervised method based on SentiWordNet. The ex-

perimental results showed that the unsupervised approach produced results that were statistically tied to those of the best supervised method (Naïve Bayes) without the cost of labeling. We use the same steps of the unsupervised approach to generate the scores used in our sentiment features. Gonçalves et al. [2013] compared eight popular sentiment analysis tools for social networks in terms of coverage (i.e., the fraction of messages whose sentiment is identified) and agreement (i.e., the fraction of identified sentiments that are in tune with ground truth). The authors found that the methods have varying degrees of coverage and agreement and that no single method is always best across different text sources. Moreover, the problem setting in these studies differs from ours as we use the sentiment of the review as a feature to predict its popularity rather than trying to predict the sentiment itself.

Nguyen et al. [2013] proposed a heuristic to select a small set of reviews that cover as many tips as possible with as few sentences as possible. By covering the tips, the authors expected to identify the review content that is more important to provide a summary of the content of the tips. They claimed that tips are good for quick zooming in on what is interesting about a item. However, when there is a large collection of tips, they may be repetitive and fragmented. Thus, the authors claimed that by selecting the reviews that cover the tips, they would obtain a readable, flowing text that would summarize and expand upon the tip content. The problem tackled by Nguyen et al. [2013] differs from our target problem as we are not aiming at selecting tip for summarization purposes.

### 2.2.3 Spam Detection

The identification of spam or fake reviews was also analyzed in a few prior studies. Spam reviews have distinct characteristics from low quality reviews [Li et al., 2011]. Low quality reviews may be biased and/or may be due to poor writing, but they reflect the user’s real opinion. A spam review, on the other hand, may be fraudulent, and is often added to the review system with a clear intention or goal to achieve [Ma and Li, 2012]. Jindal and Liu [2008] studied opinion spam in the context of product reviews. They identified three types of spam reviews: *untruthful* or *fake* reviews, which give undeserving positive reviews to promote some target objects or malicious negative reviews to defame some other object reputation; *reviews on brands only*, which do not comment on the products but only on the manufacturers or sellers of the products; and finally *non-reviews*, which can be advertisements or other irrelevant pieces of text containing questions, answers or random texts. The second and third types of spam reviews were detected using a supervised learning technique using manually labeled

training examples while the first spam type was detected by verifying whether reviews involved many opinions opposing to the majority of the other reviews. Lappas [2012] presented a study of fake reviews from the perspective of the attacker, formalizing the factors that determine the success of an attacker and exploring different attack strategies. Akoglu et al. [2013] proposed an unsupervised network-based framework to detect fraudulent users and fake reviews in online review networks using textual features and the reviewers' social networks as input features. Lin et al. [2014] proposed six features to find the spam based on the review content and reviewer behaviors. They applied supervised and unsupervised methods to identify the review spam as early as possible.

In this dissertation, our characterization study (Chapter 4) revealed the presence of spamming activity in Foursquare tips. Specifically, we revealed the existence of users who post tips whose contents are unrelated to the nature or domain of the venue where the tips were left [Vasconcelos et al., 2012b]. We discuss this further in Chapter 4. Indeed more recent studies analyzed this problem using machine learning techniques to detect user behavior related to tip spamming in LBSNs [Costa et al., 2013; Aggarwal et al., 2013].

## 2.3 Analysis of Online Content Popularity

Broadly related to our prediction task of popularity of micro-reviews is the work of assessing the popularity of online content. We review prior efforts in this direction by first describing, in Section 2.3.1, studies about popularity prediction models in several systems such as Twitter, YouTube, Digg, Boards.ie (community forums), and Facebook. Next, in Section 2.3.2, we discuss prior studies on information diffusion models using both explicit (created from users' contacts) [Leskovec et al., 2007; Bakshy et al., 2009] and implicit (created by the users' interactions) network links [Gruhl et al., 2004] as well as studies on the identification of influential or experts users [Zhang et al., 2007; Adamic et al., 2008; Agarwal et al., 2008; Cha et al., 2010; Bakshy et al., 2011].

Inspired by some of those prior studies, we investigate the properties of the implicit network built from the user interaction through tips, and propose a method to identify influential users on Foursquare (Chapter 4). In our context, influential users can be seen as reputable or experts users, who post high quality or helpful reviews and are highly rated by other users. Therefore, popularity here can be considered as an implicit measure of assessing credibility of users and tips [Abbasi and Liu, 2013].

### 2.3.1 Popularity Prediction Models

Several studies have addressed the problem of predicting the popularity of newly uploaded content. Most studies exploited textual features extracted from the messages (e.g., hashtags and URLs) or the topic of the message, as well as user related features, such as the number of followers and the source of the message (celebrities or organization) to predict content popularity in several systems.

For example, in the context of Twitter, Hong et al. [2011] tackled the problem of predicting the popularity of tweets as a classification task based on several types of features, including textual content, structural properties of the user graph, meta-data of users and messages (e.g., number of previous retweets), as well as temporal information. Suh et al. [2010] built a predictive retweet model using a generalized linear model with content and contextual features. They also identified that amongst content features, URLs and hashtags are strongly correlated with retweetability, while the number of followers and followees as well as the age of the user account are among the most important contextual features. Bandari et al. [2012] used regression and classification algorithms to predict the number of times a news URL was posted and shared on Twitter. They exploited features extracted from the news article, such as the source of the article, the category, subjectivity of the language, and the named entities mentioned in the article. Similarly to our work, Hong et al. [2011] and Bandari et al. [2012] also defined classes or levels of popularity, and developed solutions to predict which class a given tweet or article will belong to at a certain future time.

Borghol et al. [2012] developed and applied a methodology to assess the impact of various content-agnostic factors on the popularity of YouTube videos. They focused on analyzing differences among videos that have essentially the same content (clones) using a multi-linear regression model to determine which factors most influence video popularity. In that study, popularity was defined by the number of views during a given week. Our methodology has some similarities with the one adopted in [Borghol et al., 2012]. For example, we model several other factors such as the influence and activity of the user social network as well as specific Foursquare characteristics related to the venues which have no counterpart on YouTube.

Szabo and Huberman [2010] proposed a log-linear model for predicting long-term popularity of YouTube and Digg content based only on early measurements of user accesses. The authors found that long time popularity of a piece of content is correlated with its early measured popularity. However, according to Yin et al. [2012] and to what we will show in Chapter 6 prediction suffers inaccuracy if it is purely based on the number of early measurements. Finally, they concluded that social network does

not affect the content exposure, which contrast to our findings (Chapter 4). Pinto et al. [2013] extended the simple log-linear model proposed by Szabo and Huberman [2010] by building multiple linear regression models to predict the video popularity. Unlike the base model [Szabo and Huberman, 2010] which used the total number of views up to a reference date as single predictor, the proposed multivariable model uses daily views during the same period<sup>1</sup>. The authors also proposed a second model variant that includes, in addition to daily views, Radio Basis Functions (RBF) to capture the similarity between training and test data. The authors found that their RBF model leads to an accuracy gain of 71% over the model proposed by Szabo and Huberman [2010]. In Chapter 6, we used these two models, the log-linear model Szabo and Huberman [2010] and RBF models [Pinto et al., 2013] to compare with your proposed models.

Lerman and Hogg [2010] proposed a stochastic model to predict the popularity (number of votes) of user's post generated on Digg based also on early user reactions to this new content. Their model considers the complex interactions among content quality, the layout of the website, and the influence among users.

Yin et al. [2012] proposed a model to rank potentially popular items based on their early votes. The authors evaluated their model using a joke sharing application, where users can post jokes, and other users can vote if they like or dislike them. Their model assumes users tend to be conforming to the opinions of the majority in the user community (conformers) while some others exhibit contrary voting behavior (mavericks). Each person has different distributions of these two patterns, which can be learned according to the observed voting history. The authors pointed out that their model is more suitable for application in which people patterns' distributions tend to be stable for items without complex genres (e.g., jokes) as opposed to items with multiple genres (such as movies). This method may not work in our scenario since the users are not allowed to mark a tip as disliked on Foursquare. Moreover, the approach proposed by Yin et al. [2012] is based on ranks of new items while our goal is not only to rank but also to predict the potential popularity level that an individual item can achieve. Our proposed models explore other types of features and scenarios, in which the monitoring time is variable. In particular, we analyze the tip popularity at posting time, while the above studies require early votes to perform predictions.

Wagner et al. [2012] studied the patterns of user attention towards content shared within online communities, where attention was measured by the number of replies to a given post. One of their findings was that the purpose of a community may influence

---

<sup>1</sup>In this model, each variable represented the number of views during a given measured day.



how individual factors affect the attention pattern of that community. For example, posts from advice-seeking communities which contain many links in their contents are less likely to get replies, while posts from content sharing oriented communities which typically have a high number of links may have a positive impact and make posts more likely to attract the attention of such a community. They also concluded that the factors that impact whether a discussion starts tend to differ from factors that impact the length of the discussion.

Yu et al. [2011] analyzed the popularity of social marketing messages on Facebook. Using the number of likes to measure the popularity of a message, the authors evaluated the effectiveness of marketing strategies used by a number of messages from restaurants, analyzing only their textual content. The messages were grouped into “more popular” (number of likes above average) or “less popular”, and modeled using a bag-of-words representation. Two classification methods (SVM and Naïve Bayes) were used to separate messages into the two popularity classes, and rank the most discriminative features. There are some major differences between this work and our proposal. First, their method is limited only to textual content while we also make use of several features related to the user who posted the tip and the venue where the tip was posted. Second, their approach using bag-of-words may be not effective for short and informal messages such as Foursquare tips. Finally, the authors suggest to overcome the lack of dislike votes using the comments left by the users to disclose both positive and negative sentiment about the posted marketing message. In our work, we take an alternative approach and make use of the SentiWordNet scores of each tip term to capture the polarity of its content.

Tatar et al. [2014] addressed the problem of predicting the popularity of new articles based on user comments, formulating it as ranking problem. The authors compared the ranking effectiveness of two prediction methods proposed by Szabo and Huberman [2010], a linear regression model on a logarithmic scale and constant scaling model. These methods were compared with several baselines methods and with learning to rank algorithms. Their results indicate that the linear log popularity model is as effective as the learning to rank algorithms. We also approach our prediction problem as a ranking problem in Chapter 6.4. Other efforts focused in the popularity prediction exploring temporal trends and time series models. Radinsky et al. [2013] developed methods for modeling the temporal dynamics of queries and click behaviors seen in a large population of Web searchers. They explored several facets of the dynamics of Web search behavior, including the detection of trends, periodicities, and surprises by using current and past user behavioral data. Matsubara et al. [2012] proposed a unifying model for popularity evolution of blogs and tweets, showing that it can be used

for tail-part forecasts while Yang and Leskovec [2011] developed a clustering algorithm to uncover the temporal dynamics of Twitter hashtags. As future work, we intent to explore time series techniques in our popularity prediction problem.

All these previous efforts towards predicting the popularity of a online content have a similar goal compared to ours. However, our type of data, micro-reviews, does not fit completely in these models. As mentioned before, some of them have assumptions that hold only for contents with shorter life cycles (e.g., tweets or news) or features that do not have a counterpart on Foursquare, such as dislike votes.

### 2.3.2 Information Propagation and Social Influence Models

An important aspect that may affect the popularity of a content is the social influence or the fact that a individual may not make decisions independently, but rather are influenced by the behavior of other individuals. There are some contrasting theories or views about how an idea, a trend or an innovation can be spread or assimilated by people.

One of the views on social influence is based on a theory called “the two-step flow of communication”, in which ideas often flow from the mass media to a group of individuals (*opinion leaders*), who are very persuasive or well-connected, and from those to the social groups they belong to [Katz, 1957]. Since this seminal work, the opinion leaders have been the subject of several studies [Gladwell, 2002; Walther et al., 2010; Wu et al., 2011a]. Moreover, technology changes with the emergence of new forms of media such as blogs, online communities and social networks have caused the fragmentation of the mass audience into many smaller audiences. Nowadays, people can select the information which they want to be exposed to and in some cases people can generate new information themselves [Wu et al., 2011a]. The media fragmentation made the traditional adverting strategies less effective. Consequently, marketers have turned their attention to other marketing strategies (e.g., word-of-month, viral and buzz marketing) that focus on opinion leaders.

Recently, there have been several research efforts focused on analyzing the interplay between social structure and information dissemination in real networks. Gruhl et al. [2004] studied the dynamics of information propagation in weblogs. They investigated characteristics of this propagation considering the topic discussed in the posts and the individuals’ posting behavior. Leskovec et al. [2007] used a person-to-person recommendation network on a e-commerce website to study how individuals are influenced as a function of how many of their contacts have recommended a product. They also presented a model that identifies communities, products and pricing categories for

which viral marketing is effective. Bakshy et al. [2009] studied diffusion of “gestures” between friends in the social network of the Second Life virtual game. By examining the cascading trees, they found that roughly 48% of data transfers occur along the social graph. We observed similar results in the high percentage of liking activity coming from the tip author’s friends or followers, as discussed in Chapter 4.

Other studies explored algorithms using network structure to identify experts or influential users. For example, Zhang et al. [2007] discussed several approaches to identify and rank experts in a Java forum, using network-based algorithms, such as PageRank [Page et al., 1998] and HITS [Kleinberg, 1999]. Adamic et al. [2008] investigated Yahoo! Answer forums using network and textual analysis finding that user interactions vary depending on the forum topic. They also found that some categories of Yahoo! Answer forums are characterized by the presence of experts while others have a different dynamics. We employed some of the network analyses performed in that work in our characterization of tipping activity in Foursquare (Chapter 4) aiming at identifying influential users.

Agarwal et al. [2008] proposed a graph-based algorithm to identify influential bloggers based on their blog posts. The authors defined four features: recognition, activity, novelty, and eloquence which they expected to be present in an influential post. They weigh these four features to produce a combined score for each blogger. Cha et al. [2009] studied the spread of photo bookmarks on Flickr and found that social links are a dominant method for information propagation. Moreover, their results show that even popular photos do not spread neither widely nor rapidly through the Flickr network, contrary to viral marketing intuition.

Another view on social influence says the information diffusion process depends less on the properties of the influentials and more on the global conditions of the network. Watts and Dodds [2007] investigated this hypothesis using a series of computer simulations of several social network configurations. They found that, in general, most of the social changes are driven not by highly influential people but rather by easily influenced individuals who influence other easily influenced individuals, and so on. Thus, a epidemic outbreak of social influence satisfies what the individuals are asking for it (susceptibility) and not because someone is pushing for it [Smith, 2013]. Whether or not influence can spread widely depends mostly on the network structure. Therefore, if the network permits spread (global cascades), virtually anyone can start disseminating some information. The individuals who later seemed to be influential may simply be accidents of circumstances according to the author’s study. Recently, Figueiredo et al. [2014a] used an epidemic model to capture revisits by the same user and the impact of external events on the popularity evolution of objects in social media appli-

cations. Cheng et al. [2014] examined the problem of predicting the growth of cascades over social networks. Using a month of complete photo-resharing data from Facebook, the authors found that the temporal features (e.g., cascade speed) are predictive of a cascade’s eventual shape.

Other studies focused on user influence on Twitter. The Twitter graph, where the nodes represent users and edges represent following relationships, has been extensively used to predict influence on this social network [Bakshy et al., 2011; Weng et al., 2010; Cha et al., 2010; Quercia et al., 2011]. The studies have shown that influentials are not accidental, but they have some specific characteristics, such as the presence of homophily between the follower relationship [Weng et al., 2010] great personal involvement [Cha et al., 2010], the low degree of passivity of the followers [Romero et al., 2011], or specific linguistic qualities that reflect the person’s personality and mood [Quercia et al., 2011].

Bakshy et al. [2011] studied the distribution of retweet cascades on the propagation of influence, and explored various marketing strategies governed by the cost of identifying the influential users. Weng et al. [2010] proposed a new PageRank-like algorithm to measure the topic-sensitive influence of a Twitter user, comparing it against ranks based on the number of followers and based on the traditional Page Rank. Their algorithm was motivated by an observation of high reciprocity among follower relationships in their dataset, which they attributed to homophily. We also adopt a similar method based on PageRank algorithm to identify influential users on Foursquare based on a graph built from the user interaction through tips (Chapter 4). Cha et al. [2010] compared three metrics of influence, namely, number of followers, number of retweets and number of mentions, concluding that popular users with large number of followers are not necessarily influential in terms of spawning retweets or mentions. They also concluded that most influential users can hold significant influence over a variety of topics. Finally, Quercia et al. [2011] investigated the connection between the use of the language, in particular the expressed sentiment on the user’s tweets, and the influence of users on Twitter. They found that influential users have specific ways to structure linguistically their tweets, and tend to express negative sentiment in part of their tweets.

The aforementioned studies offer important insights into properties of user interactions in social networks. However, the location-based social networks have other characteristics or dimensions that have not been explored by those studies. In the specific context of predicting tip popularity, not only the social network influence, but also other entities (e.g, venues) should be taken in account as they may influence how the tip propagates through the user population. Next, we survey related work on

location-based social networks.

## 2.4 Analyses of Location-Based Social Networks

Prior studies of location-based social networks (LBSNs) tackled problems such as the identification of user mobility profiles [Li and Chen, 2009a; Noulas et al., 2011b; Rossi and Musolesi, 2014] and characterization of the use of the LBSNs [Scellato et al., 2010; Scellato and Mascolo, 2011; Noulas et al., 2011a; Scellato et al., 2011a], the design of human mobility models [Cho et al., 2011; Cheng et al., 2011b; Noulas et al., 2012; Silva et al., 2014], the development of mechanisms for recommending friends or places [Berjani and Strufe, 2011; Ye et al., 2010; Li and Chen, 2009b; Scellato et al., 2011b; Gionis et al., 2014], the design of location-based search methods [Cheng et al., 2011a], event prediction [Georgiev et al., 2014] as well as the investigation of privacy related issues [Lindqvist et al., 2011; Pontes et al., 2012b,a]. To our knowledge, we were the first to analyze the use of tips, likes and to-dos on Foursquare, as discussed in Chapter 4.

By the time we started investigating, we were aware of only two previous studies that aim at uncovering user profiles in LBSNs. In Li and Chen [2009a], the authors applied two different clustering approaches to identify user behavior patterns on BrightKite. One approach exploited the geographic position associated with user updates (i.e., check ins, photos, and notes) to classify them into four groups according to their mobility, namely, home, home-vacation, home-work, and other. The second approach clustered users based on multiple attributes such as total number of updates, social features, and mobility characteristics, and led to the identification of five groups, namely, inactive, normal, active, mobile, and trial (or non loyal) users. The second study was performed on Foursquare [Noulas et al., 2011b]. The authors used a spectral clustering algorithm to group users based on the categories of venues at which they had checked in, aiming at identifying communities and characterizing the type of activity in each region of a city. We here also identify user profiles on Foursquare (Chapter 4). Rossi and Musolesi [2014] proposed a trajectory-based approach where a user is identified simply considering the trajectory of spatio-temporal points given by his/her check-in activity. However, unlike these previous studies, we focus on user profiles in terms of their tipping activity, revealing relevant patterns (including some illegitimate or spamming activity). Wang et al. [2014] proposed a framework to discover overlapping and hierarchical communities of LBSN users considering both user-venue check-in and the attributes of users and venues (e.g. venue category).

Some other studies focused on the properties of the social networks in LBSNs. For example, Scellato et al. [2010] analyzed the social, geographic and geo-social properties of four social networks that provide location information about their users, namely BrightKite, Foursquare, LiveJournal and Twitter. They showed that LBSNs are characterized by short-distance spatial-clustered friendships, while in the other types of networks, such as Twitter and LiveJournal, users have heterogeneous connection lengths. An analysis of Gowalla users showed that the number of friends follows a double Pareto-like distribution, whereas the numbers of check ins and places are better described by log-normal distributions [Scellato and Mascolo, 2011]. The authors also analyzed the temporal variations of such distributions, observing that users tend to add new friends at a faster rate than they give check ins and go to new places. Noulas et al. [2011a] analyzed the dynamics of user check ins and the presence of spatio-temporal patterns on Foursquare. They found that user heterogeneity was observed with respect to the number of friends, average distance and social triads. Another study also pointed out that users with fewer friends tend to generate social triangles on a small geographic scale, while users with more friends tend to belong to geographically wider triangles [Scellato et al., 2011a].

Modeling human mobility requires access to spatial and temporal information about the places people visit. LBSNs are rich sources of this kind of data. Indeed, towards that goal, Noulas et al. [2012] studied the urban mobility patterns in several metropolitan cities around the world by analyzing a dataset containing check ins of Foursquare users. Human mobility patterns were also investigated on Gowalla, Brightkite and cell phones trace datasets [Cho et al., 2011], with the conclusion that humans experience a combination of strong short range spatially and temporally periodic movement that is not influenced by the social network structure, while the long-distance travel is more influenced by the social links. The authors proposed a mobility model that combines periodic daily movement patterns with the social movement effects caused by the friendship network. Finally, Cheng et al. [2011b] studied human mobility patterns by analyzing spatial, temporal, social and textual aspects associated with tweets containing check-ins. They observed that LBSN users follow reproducible patterns, and that social economical factors are related with mobility. Silva et al. [2014] investigated the potential of LBSNs to build participatory social networks and exploited them to study city dynamics.

The task of recommending friends or places to LBSN users, has also been tackled by previous work. Using collaborative filtering techniques, Berjani and Strufe [2011] proposed a personalized recommender for places in Gowalla based on the number of check ins at spots, whereas Ye et al. [2010] proposed a collaborative recommendation

algorithm that uses the number of Foursquare check ins of commonly visited places to perform recommendations. Li and Chen [2009b] proposed a three-layered recommender model using attributes of user profiles (preferences), social graphs (friendship), and mobility patterns (distance of visited places) to recommend friends in Brightkite. Another supervised learning framework to recommend places and friends was proposed by Scellato et al. [2011b] and evaluated in a longitudinal dataset collected from Gowalla. Using check in information regarding users who visited the same place and friends of friends, the authors were able to reduce the link prediction space and thus improve prediction accuracy. Both previous studies [Scellato et al., 2011b; Li and Chen, 2009b] concluded that the inclusion of information about location-based activity leads to better predictions if compared to when only social data is used. Complementary, Cheng et al. [2011a] used check in traffic patterns, extracted from the temporal dynamics of Foursquare check ins during a time period, to develop a traffic-driven location clustering algorithm, which in turn, was used to improve the recommendation of nearby places. Gionis et al. [2014] focused on the problem of recommending customized tours in urban settings using the same dataset as Cheng et al. [2011b]. Their proposed framework recommends tours considering the different types of venues, the order in which the user wants to visit each place, the budget constraints expressed in terms of distance, and the satisfaction that each venue can provide to the user. Finally, Georgiev et al. [2014] used data from Foursquare to analyze event patterns in three metropolitan cities (London, New York, and Chicago) to understand to which extent geospatial, temporal, and social factors influence users' preferences towards events.

As in other online social networks, information sharing also raises concerns about exposure of user private data. Another thread of research in LBSNs has been devoted to discuss privacy related issues. A large-scale study on inferring the home location of Foursquare users is presented by Pontes et al. [2012b,a]. The authors analyzed the potential of using publicly available features such as mayorships, tips and likes as sources of information leakage. Lindqvist et al. [2011] surveyed users about their motivation for using location sharing services as well as how they manage their privacy. The authors observed that the majority of interviewed users did not have privacy concerns since, reportedly, they can manage their privacy selecting what they are willing to share.

Although the aforementioned studies offer important insights into properties of user interactions in LBSNs, none of them addressed how users exploit tips and likes as a means to review or recommend a place or a service as well as to give feedback regarding previously posted tips. Our characterization study, reported in Chapter 4, aims at contributing to fill this gap.

## 2.5 Summary

In this dissertation, we attempt to understand how users exploit micro-reviews (tips) on Foursquare and we focus on the specific problem of predicting a tip's future popularity. Our popularity metric is based on the number of likes received by a tip. Besides reflecting the popularity of a tip at a given point in time, the same number can also be used as an estimate of the helpfulness or the quality of the same tip. Thus, studies on assessing the helpfulness or quality of reviews as well as studies on popularity prediction of online content are also related to our work.

We started this chapter describing some related work on information credibility in Section 2.1. Some of those studies pointed that the quality (or the popularity) of a content is related to the perceived credibility of the author or of the information itself. Some of the credibility directives, proposed in those studies, guided us in the design of features used in our prediction models.

Next, in Section 2.2, we discussed previous work focused on automatically determining the quality (or helpfulness, or utility) of reviews, which primarily made use of either content features and target longer pieces of content. In contrast to those reviews, Foursquare tips tend to be more concise, subjective and informal. Thus, we here exploit in addition to content features, attributes related to the user who posted the tip and the venue where it was posted. Moreover, most previously studied review systems are based on helpfulness scales (e.g., ratings 1 to 5 or votes of unhelpfulness), while Foursquare tips are evaluated only by like marks. The lack of a clear helpfulness scale makes the prediction task more complicated.

Other studies closely related to our prediction problem focus on predicting the popularity of online content (Section 2.3). However, the models proposed in these studies are very specific to certain scenarios or type of data (e.g., tweets, videos, and news), where they are evaluated. In addition, some models require early popularity measurements of the same content that is the target of the prediction. Our prediction models explore other types of features and scenarios in which we vary the monitoring time, including a scenario where prediction is performed at posting time (so no early popularity measurements is available). In the second part of the same section, we discussed relevant studies about information propagation and social influence models. Those studies guided us in some of our analyses described in Chapter 4. In particular we have adopted an approach already exploited by Weng et al. [2010] and by Adamic et al. [2008] to characterize and estimate user influence using the network built from the user interactions through tips.

Finally, in Section 2.4 we surveyed related work on location-based social networks.



However, most of those studies focus on the user check in dynamics, on the properties of the social graph, and on related geographical information. Thus, to the best of our knowledge, there is no previous analysis of how users exploit tips and no previous studies about assessing tip popularity. Next, we present the main elements and features of Foursquare and describe the methodology we adopted to crawl it, as well as a summary of the collected datasets.



# Chapter 3

## Foursquare: Case Study

Social networks have become increasingly present in our everyday habits. The advances in mobile communication and development of geographic information systems, such as the GPS (Global Positioning System) technology, allowed the design of a variety of context-aware applications. Sharing their current location (check in), uploading location-tagged photos, commenting in real-time events of services are examples of features available for users of a new type of online social network: the location-based social networks (LBSNs). Foursquare and Google+, the currently most popular LB-SNs<sup>1</sup>, and social networks that incorporated location-based services (e.g., Facebook and Yelp) have facilitated new types of relationships between users and places in the physical world that are registered in the application.

One of the contributions of this dissertation is a comprehensive characterization of one of the currently most popular location-based social networks, namely Foursquare, focusing on the user-generated micro-reviews (called tips) that users post on the system. In this chapter, we first review the main elements and features of Foursquare (Section 3.1). Next, we present the methodology we adopted to crawl the system, as well as the datasets collected (Section 3.2), which are used in our analyses in the following chapters.

### 3.1 Foursquare: Key Elements and Features

Foursquare is a prime example of an LBSN on which users share their locations with friends. As videos and images are the main objects on YouTube and Flickr, respectively, the main objects on Foursquare are venues. A **venue** represents any physical location,

---

<sup>1</sup>Gowalla and Brighkite are also examples of LBSNs, but they no longer exist.

like a store, a restaurant, an university, an airport or a monument, where users can check in. Users may check in at venues when they are physically close to those venues, using GPS-equipped mobile devices. Once users check in, they may choose to share their locations with friends. Every time a user checks in a venue, she collects points, namely **badges**, on Foursquare. If a user has more “check ins” in a certain venue than any other user in the past 60 days, she becomes the venues’ mayor. Venues are created by Foursquare users who become owners of that place. However, venues can be claimed by real business owners. In this case, venues are *verified* by Foursquare and, if approved, the real owners of the venue can start offering promotions and special deals to users who frequently check in at that venue. Foursquare also maintains a set of nine pre-defined venue categories, namely, “Arts & Entertainment”, “Colleges & Universities”, “Food”, “Great Outdoors”, “Nightlife Spots”, “Professional & Other Places”, “Travel & Transport”, “Shops & Services”, “Residences”.

Foursquare users are categorized into standard (or regular) users, celebrities, and brand pages<sup>1</sup>. Standard users automatically become celebrities when they achieve more than 1,000 friends in the system<sup>2</sup>; Brand pages, in turn, are users that represent companies or businesses (e.g., History Channel, Starbucks). Foursquare allows two types of social relationship among their users, namely, friendship and the follower-follower relationship. The type of relationship that a user may establish depends on the user’s category: standard users can only have friends and followees, brand pages can have only followers and followees, whereas celebrities can have friends, followees and followers.

In addition to check ins, users can also post *tips* (i.e., comments or reviews) at specific venues, commenting on their previous experiences when visiting the corresponding places. Tips can contain helpful information to guide others in their choices, such as suggestions, recommendations or disapprovals (e.g., “I love this apple pie”, “The bathrooms are not clean”), practical advices or directions (e.g., “where is the closest ATM machine from the museum?”) and factual comments that reveal fun and surprising facts about a location (“This place was founded in 1788”) [Sarah Best, 2012].

Foursquare tips are typically much more concise (the length is restricted to 200 characters) and often more informal and subjective than other reviewing systems. For example, on systems like TripAdvisor, Amazon and Yelp, reviews are often longer and more formally structured, and often carry very specific information about a product/service. Nevertheless, tips may nourish the relationship between users and real businesses, offering valuable feedback that business owners can benefit from to im-

---

<sup>1</sup><http://aboutfoursquare.com/user-type-comparison/>

<sup>2</sup><http://aboutfoursquare.com/foursquare-converts-most-popular-users-to-celebrity-accounts/>

prove their products. Moreover, tips may be key features to attract future visitors to both the venue and the corresponding physical place. We note that unlike user check ins, which are visible only to the user’s friends, tips are visible to everyone. Thus, tips have the potential to significantly impact online information sharing and business marketing.

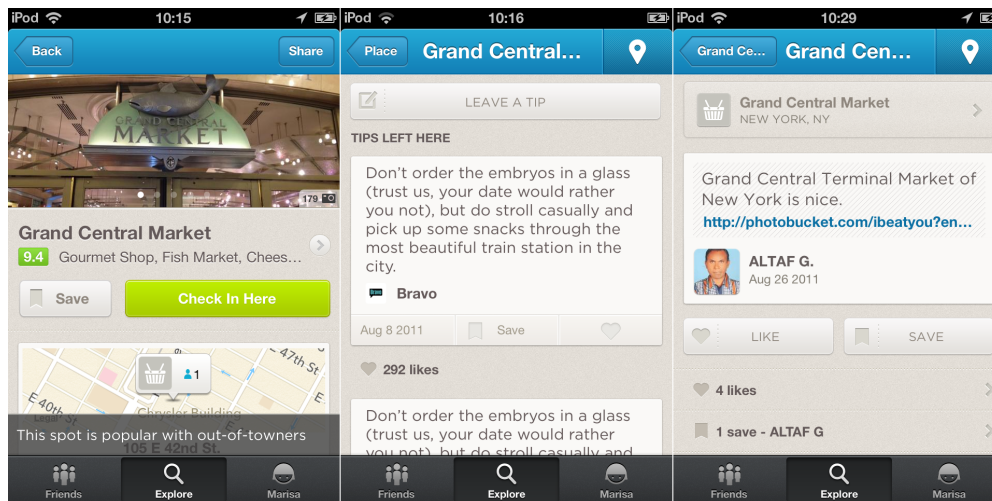


Figure 3.1: Screenshot of a Foursquare Venue Page.

Users can also evaluate previously posted tips by clicking on a “like” mark or saving them in a to-do list, in sign of agreement with the posted content or interest in the information provided in it. “Like” and to-do marks ultimately serve as feedback from other users with regards to the helpfulness or interestingness of the tip. Examples of the most popular tips of our dataset: “The park opened in 1971 & is the world’s largest and most visited recreational resort.”, “Go on a Sunday - it’s a nonstop party all day and make sure you get a pitcher of Mojitos all for yourself.” and “You can shop days and nights. They don’t sleep.”

A newly posted tip goes to the venue page, displayed sorted by the number of likes received or by posting time, but only the former is available in the mobile application (Figure 3.1). The order at which tips are displayed to the user may also vary depending on the authors of the tips. For example, tips posted by the user’s friends and followees are shown first to her. Moreover, the user also receives notifications when any friend or followee posts a tip<sup>1</sup>.

A number of factors may impact a tip’s future popularity, including: (1) the website layout which gives more visibility to already popular tips, thus contributing to the rich-get-richer effect [Liu et al., 2007]; (2) the popularity of the venue where the

<sup>1</sup>Likes are not notified to the user’s social network, as in Facebook.

review was posted; (3) characteristics of the user who posted it (including the number of active friends/followers); and (4) the content itself. Most of these factors are analyzed in Chapter 4, and mapped into features exploited by our popularity prediction models in Chapter 6.

## 3.2 Measurement Methodology

We study how users explore micro-reviews (tips) by observing certain properties of each entity that plays a central role in the problem – namely user, venue, and the tip’s textual content – in order to estimate the number of likes received by those during a given time. We use two datasets collected from Foursquare: one was collected from a set of venues, whereas the other was gathered from a set of users. In this section, we briefly describe the strategy adopted to crawl Foursquare in Section 3.2.1. We then summarize the two datasets in Section 3.2.2<sup>1</sup> and 3.2.3<sup>2</sup>.

### 3.2.1 Crawling Methodology

Our study is based on two datasets collected from Foursquare using the system API. We start describing how the venue dataset was collected. Our crawling strategy, which relies on a set of worker processes and a master process, exploits the fact that each venue in Foursquare receives a unique and sequential numeric identifier (ID)<sup>3</sup>. Given  $M$  (an estimate of) the largest ID assigned to a venue in the system, the master process randomly selects an ID according to a uniform distribution in the  $[0, M]$  range, and gives it to the next idle worker. We chose to perform a random selection of IDs, as opposed to sequentially selecting each possible value, to minimize the chance of a bias towards older user accounts (which, we conjecture, have smaller IDs). The worker then sends a request to the Foursquare API to gather information about the corresponding venue. In particular, for each collected venue, the crawler collects all its tips, the identifications of the users who posted each of them, the number of likes and to-dos each tip received, the number of users who checked in the venue, the venue category as well as its geographic coordinates.

A series of initial experiments consisting of sending HTTP GET requests to the pages of specific venues identified by their IDs was executed to verify their existence. We tried increasing values of IDs, starting with 0. The largest ID for which we did get

---

<sup>1</sup>We named this dataset as dataset 1.

<sup>2</sup>We named this dataset as dataset 2.

<sup>3</sup>We observe that Foursquare venues IDs are no longer sequentially assigned.

a response corresponding to a valid webpage was 20 million. We experimented with many IDs greater than that value, but in all those cases, the response was “Not Found”. Thus, we speculate that, at the time of our crawling, 20 million was the largest venue ID in the system. Thus, we set  $M$  equal to this value, and used it as input to our crawler. The venue crawling process ran from May 23<sup>rd</sup> to July 19<sup>th</sup> 2011, gathering data corresponding to more than 1.6 million venues as further described in Section 3.2.2.

After this crawling, we learned that the Foursquare API made available information about its users. We then decided to stop the venue collection, and started collecting the users (instead of the venues) from August to October 2011. This decision was made since the number of users on Foursquare (around 10 million users<sup>1</sup>) at that time was smaller than the number of venues, and we were interested only in venues with at least one tip. Thus, it would be most cost-effective to collect the users.

We used the same crawling methodology used for the venues to collect data about the users. We collected profile data for each user, including name, user type, home city, total number of check ins, list of friends, list of mayorships, and list of tips. For each tip posted by the user, the worker also collected the total number of likes received, the set of users who marked it as liked, as well as the tip’s content, timestamp and venue identifier where the tip was posted. Finally, for each venue associated with a tip, the crawler collected the total number of user check ins, total number of unique visitors, and its category. At the end of this period, we were able to collect more than 13 million users. We believe that this represents a large fraction of the total user population, since, as previously reported<sup>2</sup>. Next, we describe each collected dataset, the total number of registered users varied from 10 million in June 2011 to 15 million in December of the same year.

### 3.2.2 Venue Dataset (Dataset 1)

Our venue crawler ran from May 23<sup>rd</sup> to July 19<sup>th</sup> 2011, gathering data corresponding to more than 1.6 million venues. Table 3.1 summarizes the collected dataset. Associated with the crawled venues, we were able to identify almost 1 million tips, and more than half million unique users, out of whom 1,248 are brand users. Moreover, 3.8% of the venues are verified, while 18.5% of them had received, by the time of the crawling, at least one tip. Since our focus is on understanding how users exploit tips, likes and

---

<sup>1</sup><https://foursquare.com/infographics/10million>

<sup>2</sup><http://www.socialmedianews.com.au/foursquare-reaches-15-million-users/>

Table 3.1: Summary of Our Venue Dataset.

Number of venues	1,601,412
Number of venues with at least one tip	296,217
Number of verified venues	61,378
Number of users	526,651
Number of brand (pages) users	1,248
Number of tips	984,251
Total number of likes for all tips	1,407,835
Total number of to-dos for all tips	393,574

to-dos, our analyses in the Section 4.2 are performed over the venues with at least one tip.

### 3.2.3 User Dataset (Dataset 2)

Table 3.2: Summary of Our User Dataset

Number of tips	6,817,992
Number of tips in English	4,374,922
Number of likes <sup>1</sup>	5,740,954
Number of tips with at least one like	2,341,579
Number of venues with at least one tip	3,194,556
Number of users who posted at least one tip	1,831,747
Number of users who marked at least one tip as like	756,734
Number of users with at least one tip marked as like	910,486
Number of venues with at least one tip marked as like	1,254,843
Number of verified venues	219,418

Our complete user dataset, collected from August to October 2011, contains almost 16 million venues, and over 10 million tips. However, to avoid introducing biases towards tips that are either too old or too recent, we restricted our analyses to tips and likes created between January 1<sup>st</sup> 2010 to May 31<sup>st</sup> 2011. After applying this filter, we ended up with over 6 million tips, and over 5 million likes, posted in slightly more than 3 million venues by more than 1,8 million users. Table 3.2 provides some statistics about our analyzed user dataset. Note that around 34% of the tips received, during the considered period, at least one like. As shown in Table 3.2, more than 4

<sup>1</sup>Some likes were filtered since they had timestamps inconsistent (earlier) than the timestamp of the associated tips.



million tips in our user dataset are in English. To identify them, we used a Linux dictionary (myspell) to filter tips with fewer than 60% of the words in English.

### 3.3 Summary

In this chapter, we discussed the main elements and features of Foursquare, the currently most popular location-based social network. We also presented our crawling methodology and the summary of our two datasets. Next, we use these datasets to analyze how users interact using tips to understand what factors affect the tip popularity. These datasets capture the three dimensions (user, venue and tip textual content) that are relevant for our study.

## Chapter 4

# Tipping Activity on Foursquare: Characterization and User Influence

Understanding how users behave when they interact with each other through tips or related features (e.g., likes and to-dos) is important to derive insights into which factors impact tip popularity as well as insights that can help the interpretation of the prediction results. In this chapter, we first study several factors impacting the popularity of a tip, including attributes related to the three entities that play central role in the tip popularity prediction problem, namely, users (Section 4.1.1), venues (Section 4.1.2), and tips (Section 4.1.3). These factors are incorporated as features in our prediction models, introduced in Chapter 6.4. Next, we identify four groups of users with very different behaviors regarding their usage tips and likes (Section 4.2). One of the identified group consists of potential spammers, as those users post tips that are unrelated to the venue. Towards better understanding user interactions, notably the presence of influential (or expert) users, we discuss methods to automatically infer a user’s influence level on Foursquare in Section 4.3. Finally, in Section 4.4, we analyze the dynamics of tip popularity on Foursquare.

### 4.1 Characterization of Tipping Activity

In this section, we analyze how users exploit the tips and likes on Foursquare using the larger user dataset (dataset 2) presented in Section 3.2.3<sup>1</sup>. We start by focusing on the users, and analyze how they interact through tips and how their tips are evaluated (Section 4.1.1). Next, we discuss how users interact with venues by posting tips and

---

<sup>1</sup>Since dataset 1 is a subset of dataset 2, the findings of this section are also valid for dataset 1.

marking them as liked (Section 4.1.2). Finally, we analyze features extracted from the tips' contents (Section 4.1.3).

### 4.1.1 User Analysis

We start our characterization by focusing on the users and analyzing the total number of tips posted as well as the total number of likes received and given by each user. The number of likes received by a user refers to the total number of times that any tip posted by the user was marked as liked by others. Thus, it reflects the popularity of the collection of tips posted by the user. The number of likes given by a user reflects the feedback given by her on other users' tips. Our characterization is performed over all users in the dataset 2 (see description in Section 3.2.3). We note that these are, in the vast majority (99.8%), standard users.

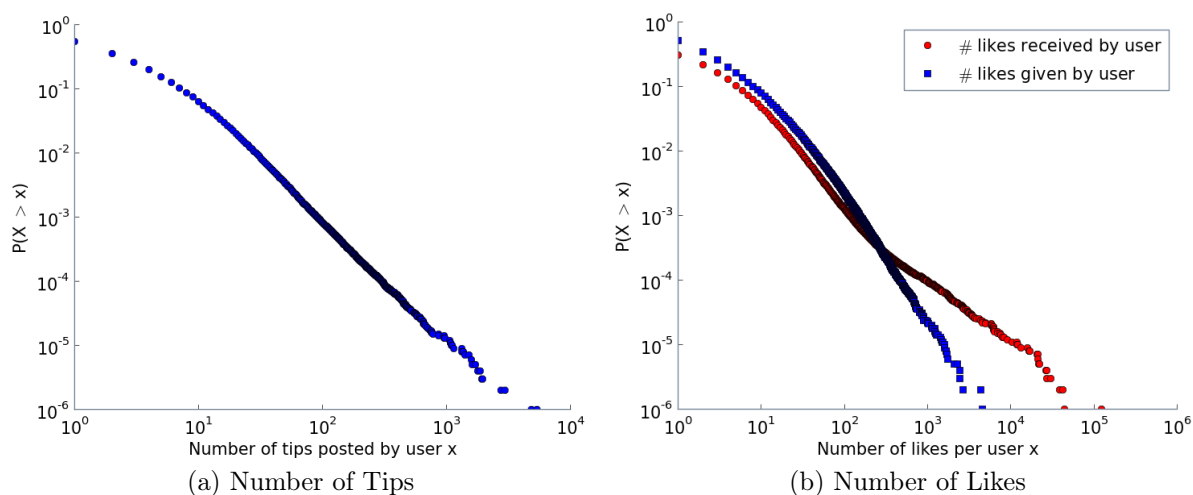


Figure 4.1: User Tipping Activity on Foursquare.

The complementary cumulative distribution functions (CCDF) of these measures are shown in Figure 4.1, with both axes in logarithmic scale. Complementary data is also shown in Table 4.1, which provides maximum, mean, and median values as well as coefficient of variation (CV)<sup>1</sup> for these three metrics<sup>2</sup>. Clearly all three distributions are heavy-tailed: most users posted very few tips and/or received/gave a few likes, while a small fraction of the users posted a lot of tips and/or received/gave a lot of feedback on previously posted tips. For instance, 46% of the users post only one tip, whereas 19% and 48% of them receive and give only one like, respectively. In contrast,

<sup>1</sup>Ratio between standard deviation to the mean.

<sup>2</sup>Recall that, since we are analyzing the filtered user dataset, all users have at least one tip.

1,499, 2,318 and 2,935 users posted more than 100 tips, received more than 100 likes, and gave more than 100 likes, respectively. The maximum number of tips posted by a single user is 5,791, whereas a single user received almost 209 thousand likes, as shown in Table 4.1. The heavy tail nature of these distributions indicates that most likes are concentrated on tips posted by few users, suggesting that the tips posted by such users may experiment the rich-get-richer phenomenon, which suggests that a tip will attract new likes at a rate proportional to the number of likes already acquired [Borghol et al., 2012]. We further analyze this in Section 4.4. Note also that the median number of likes per user is, on average, only 0.48, which implies that many users have this feature equal to 0. This will impact our prediction results, as discussed in Chapter 6.

We also analyze whether users who post more tips tend to give/receive more likes. Figure 4.1b shows that there are some users who tend to receive a proportionally much larger number of likes in their tips, indicating high popularity of their tips. Correlation between variables was assessed by the non-parametric Spearman’s rank correlation coefficient ( $\rho$ ) test [Zwillinger and Kokoska, 2000], defined as :

$$\rho = 1 - \frac{6 \sum (d_i)^2}{n(n^2 - 1)} \quad (4.1)$$

where  $n$  is the number of paired ranks, and  $d_i$  is the difference between the paired ranks. The correlation ( $\rho$ ) computed over the number of tips posted and the number of likes received is moderate (0.54), while between the number of tips and the number of likes given by each user is lower (0.37). We also analyzed the correlation between the number of likes received and number of likes given by each user, finding an even lower correlation (0.35) between them. Thus, in general, users who tip more do not necessarily receive more likes, and users who give more likes do not always receive more likes.

Table 4.1: Summary of Users Tipping Activities.

Metric	Maximum	Mean	Median	CV
# of tips per user	5,791	3.72	2.0	3.25
# of likes received per user	208,619	3.13	0	63.40
# of likes given per user	14,090	4.38	2	4.50
Median of the # of likes per user	657	0.48	0	2.77
Mean the # of likes per user	858	0.58	0	2.70
Std of the # of likes per user	632.81	0.34	0	3.88
# of friends/followers per user	318,890	44.79	17	15.95
# of mayorships per user	325	1.21	0	2.42

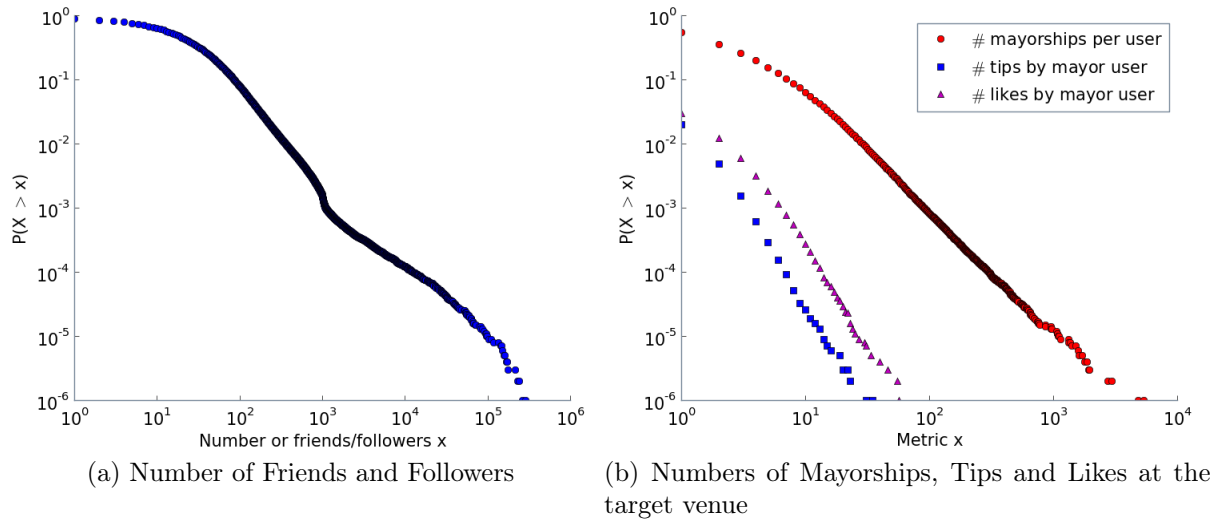


Figure 4.2: Number of Friends and Followers and Number of Mayorships per User.

Next, we analyze the total number of friends or followers of the user, the total number of mayorships won by each user, as well as the number of tips or likes posted or received by a user at the same venue where she was a mayor. Figure 4.2 shows the distributions of these measures, with both axes in logarithmic scale.

The distribution of the number of friends and followers indicates that the social network is quite sparse among users who post tips: a user has on average only 44 friends or followers, whereas 37% of the users have at most 10 friends/followers, although the maximum reaches 318,890 (*MTV* user). Moreover, we analyze the user familiarity with the venue through two points of view: number of mayorships accomplished, and if the user was a mayor in the venue where he posted a tip. We conjecture that a user who frequently visits the same venue will have a higher probability of writing more interesting/helpful/popular tips about the place. For instance, 41% of the users who post a tip had at least one mayorship in the same venue. In contrast, 37 users are mayors of more than 100 venues, 47 users have more than 10 posted tips and 32 users have more than 20 received likes in venues where they were mayors. We note that users who post tips have at least one mayorship on average but the correlation between the number of tips posted and number of mayorships is relatively low (0.36). Similarly, the correlation between the number of likes received and the number of mayorships won by each user is also relatively low (0.20). Thus, our conjecture is not as strong as we expected. Yet, it is not negligible, thus, we capture them in some of the features exploited by our popularity prediction models (Chapter 6).

Finally, we analyze the cumulative distribution of the fraction of likes received by

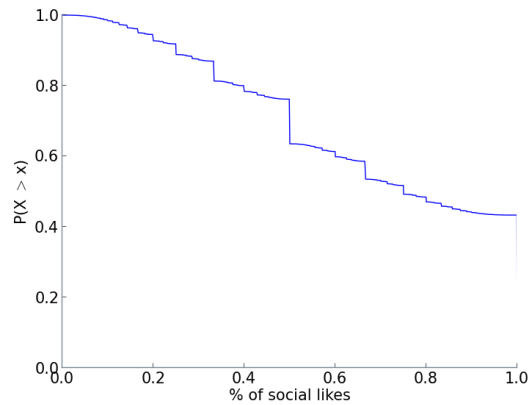


Figure 4.3: Fraction of Likes Received from the User’s Social Network (Friends and Followers).

a user that comes from her social network, that is, likes that were given by her friends or followers. Figure 4.3 shows that, for 70% of the users, at most 50% of the likes come from their friends and followers. In other words, the social network has influence on the popularity of the tips posted by a user.

### 4.1.2 Venue Analysis

We now turn to the characterization of the venues. Figure 4.4a shows the CCDFs of the number of posted tips and the number of likes accumulated by all tips per venue, while Figure 4.4b shows the distributions of the total number of check-ins and the number of unique visitors. Complementary data is shown in Table 4.2, which provides maximum, mean and median values for these three metrics. Recall that, since we are analyzing the filtered dataset, all venues have at least one tip.

Table 4.2: Summary of Visiting and Tipping Activities at Venues.

Metric	Maximum	Mean	Median	CV
# of tips per venue	2,419	2.13	1	2.23
# of likes per venue	7,103	1.80	0	11.60
Median # of likes per venue	390	0.45	0	2.81
Mean # of likes per venue	390	0.52	0	2.79
Standard deviation of # of likes per venue	774.44	0.26	0	8.27
# of check ins per venue	484,683	217.33	48	6.18
# of unique visitors per venue	167,125	87.35	15	5.87

Once again, Figure 4.4 shows that the four distributions are heavy tailed: most venues, when tipped, receive only a few likes as most likes are concentrated on tips

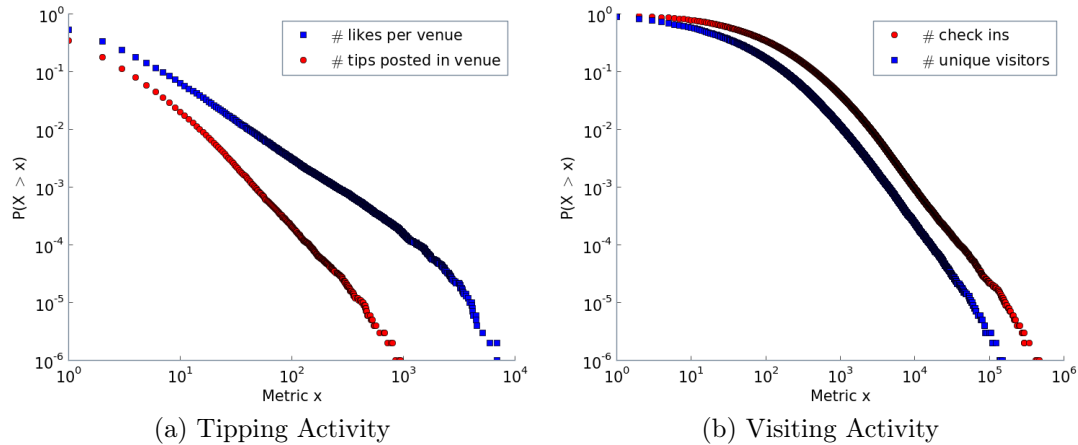


Figure 4.4: Visiting and Tipping Activities per Venue.

posted at few venues. Note that the maximum number of likes per venue exceeds 7,000, but is, on average, only 1.80. The Spearman’s rank correlation coefficient computed between the total number of tips and the total number of likes per venue is moderate ( $\rho = 0.50$ ), implying that tipping, in general, tends to be somewhat effective in attracting visibility to a venue: the larger the number of tips, the higher the amount of user feedback the venue tends to receive.

Some venues also concentrate most of the check-ins, visitors and tips. For instance, while the median number of check ins per venue is only 48, the mean exceeds 217. Moreover, around 70 venues have more than 100,000 check-ins each, and one venue (i.e., Los Angeles International Airport) has almost half million check ins. Similarly, around 10 venues have more than 100,000 unique visitors. Regarding the total number of tips per venue, we find that 66% of the analyzed venues have only one tip, while 667 of them receive more than 100 tips each. One venue (i.e., Super Bowl Arlington) in particular receives more than 2,000 tips. The correlation between the total number of check ins and number of unique visitors is 0.86, revealing a very strong correlation, as one might expect. Moreover, we find that the correlation between either the number of check ins or the number unique visitors and total number of tips posted in the venue is moderate to high (around 0.52) indicating that, to some extent, popular venues do tend to attract more tips. The same moderate to high correlation is also observed between the total number of likes and the number of check ins ( $\rho = 0.50$ ) and the number of unique visitors ( $\rho = 0.46$ ). As temporal information related to check ins are not available in our dataset, we are not able to tell whether the larger number of check ins is due to the larger amount of feedback received by the tips posted at the venue or vice-versa. However, regarding the median number of likes per venue, we found a low



correlation ( $\rho = 0.3$ ) with the total number of check ins, which means that not all tips posted at the same popular venue receive comparable number of likes. Whereas this is probably due to various levels of interestingness of tips posted at the same venue, it might also reflect the rich-get-richer effect.

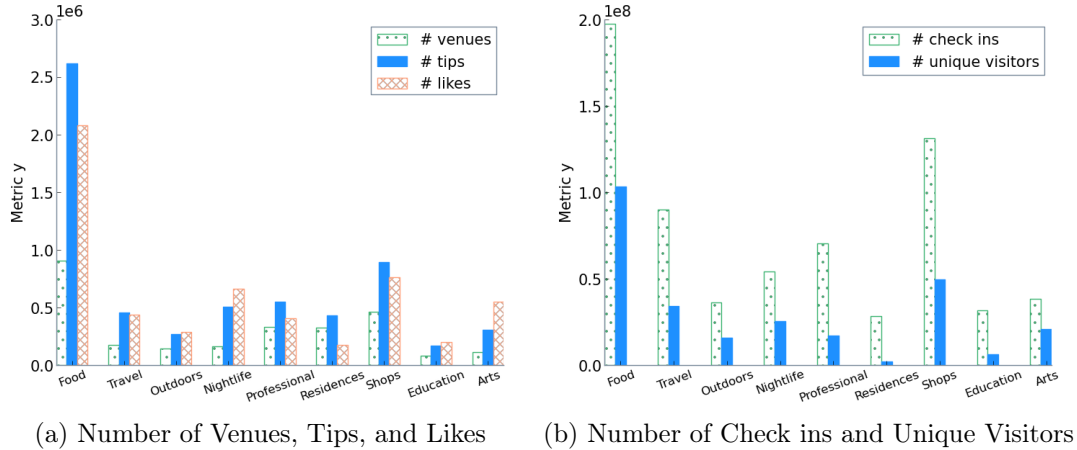


Figure 4.5: Distributions per Venue Category.

Foursquare also maintains a set of nine pre-defined venue categories, namely, “Food”, “Travel & Transport” (Travel), “Great Outdoors” (Outdoors), “Nightlife Spots” (Nightlife), “Professional & Other Places” (Professional), “Residences”, “Shops & Services” (Shops), “Colleges & Universities” (Education), “Arts & Entertainment” (Entertainment). Figure 4.5 shows histograms of the number of venues, tips, likes, check ins and unique visitors in each category, in our dataset.

Approximately 43% of the venues in our filtered dataset (i.e., venues with at least one tip) are from the *Food* and *Shops* categories. These categories are also the ones that attract most tips and receive most feedback on their tips. For instance, the two venues that receive the largest number of tips are the *Super Bowl Sunday* event and Soekarno-Hatta International Airport in Jakarta, whereas the venues that have the largest number of likes in their tips are the Madison Square Garden and the Disneyworld’s Magic Kingdom Park. In contrast, the categories that receive the smallest number of tips and likes are *Residences* and *Colleges and Universities*.

### 4.1.3 Tip Analysis

Finally, we analyze the tips, characterizing not only the total number of likes received by each tip but also three properties of the tip’s content, namely, the number of characters, the number of words, and the number of URLs and e-mail addresses. Figure 4.6a shows

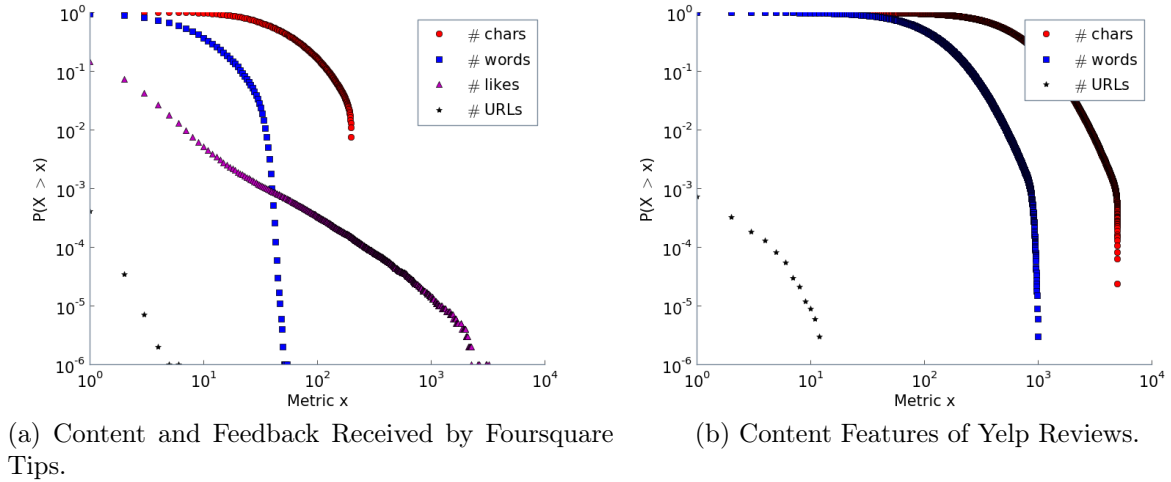


Figure 4.6: Content Features of Foursquare Tips and Yelp Reviews.

that, as expected, most tips receive a very small number of likes (66% of tips receive no like at all), whereas some tips have a high popularity among the users. For instance, 95 tips receive more than 1,000 likes. One such example is a tip posted at the Magic Kingdom Park venue giving historical facts about the place.

We also observe that most tips are very short, with, on average, approximately 60 characters and 10 words. The maximum numbers of characters and words are, respectively, 200 (limit imposed by the application) and 66, as shown in Table 4.3. Also, the vast majority (98%) of the tips carry no URL or e-mail address. Moreover, we find no strong correlation between the size of the tip and the number of likes received by it ( $\rho$  under 0.07).

This is a preliminary analysis, other features that capture other textual properties such as readability, informativeness, part-of-speech tagging of each tip word, and polarity of the tip’s sentiment (positive, neutral, negative) will be exploited in Chapter 6.4 and 6.

Table 4.3: Summary of Tip Textual Characteristics.

Metric	Maximum	Mean	Median	CV
# of words per tip	66	10.25	8	0.78
# of characters per tip	200	59.78	46	0.75
# of URLs per tip	9	0.02	0	8.27
# of likes per tip	5352	0.84	0	10.74

Finally, we also observe that Foursquare tips are much shorter than reviews in

Yelp (Figure 4.6b), which confirms our hypothesis about the nature of tips being different from other types of online reviews previously studied<sup>1</sup>.

In this section, we observed that most users posted very few tips and/or received few likes while most tips and likes are concentrated on a small fraction of the users (Figure 4.1). Next, we further analyze user tipping behavioral patterns by first identifying and characterizing relevant and typical user profiles (Section 4.2), and then characterizing the properties of the network that emerges from user interactions through tips and can be used to assess user influence (Section 4.3).

## 4.2 User Profiles

In the previous section, we found that the correlation between the number of tips and the total number of likes received by the tips of a user is only moderate (0.54). Thus, users who tip more do not necessarily receive more feedback on them. In other words, if we take the total number of likes received by all tips posted by a user as an estimate of the user influence in the system, such influence is only moderately correlated with her degree of tipping activity, seeming much more related to the tipped venue<sup>2</sup>.

In this section, we discussed various user behavior patterns observed in our Foursquare dataset, with respect to the use of tips, likes and to-dos. We start analyzing users who have suspicious behavior, in particular those who post tips with links (Section 4.2.1). Next, based on the patterns of users tipping activity, we classify users into four tipping profiles in Section 4.2.2. The analysis made in this Section were performed over the dataset 1 described in Section 3.2.2<sup>3</sup>. The results of this Section were published in [Vasconcelos et al., 2012b]<sup>4</sup>.

### 4.2.1 Suspicious Behavior

According to the Foursquare’s terms of service, the introduction of links to unrelated sites across various venues is considered *spamming*, and users who are caught doing it should have their accounts deactivated [Foursquare, 2011]. We found some users who had the majority of their tips containing links, i.e., URLs and email addresses. This finding raises a concern about suspicious behavior, particularly because only links

---

<sup>1</sup>This analysis was performed using a Yelp dataset published in [http://www.yelp.com/academic\\_dataset](http://www.yelp.com/academic_dataset)

<sup>2</sup>We define in Section 4.3 other ways to measure influence.

<sup>3</sup>The total number of to-dos were also used as a tip feedback since they were public available in that dataset.

<sup>4</sup>Likes were referred to as dones in that work.

included in the tip’s text were accounted for. In other words, links placed in the “More info” field, expected to be directly related to the target venue (e.g., the venue’s website), as well as related pictures, placed in a separate field, were disregarded.

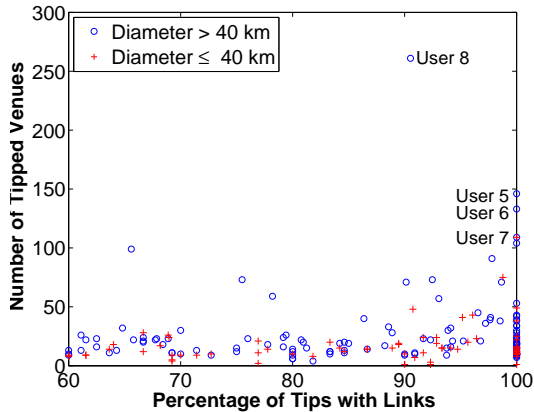
To delve deeper into this issue, we selected users with at least 10 tips and at least 60% of their tips containing links for further analysis. This selection corresponds to 3% of all users who posted at least 10 tips. Figure 4.7a plots the percentage of tips with links versus the number of tipped venues for these users, whereas Figure 4.7b shows the percentage of tips with links versus the total number of likes and to-dos. In this plot, users are grouped into two sets based on the maximum geographical distance between any pair of venues tipped by them. We refer to such distance as the *diameter* of the venues tipped by the user, and use it to assess the scale (local or global) of the user’s tipping activity.

The graphs show no clear correlation between the percentage of tips with links and the total number of likes and to-dos or the number of tipped venues. Indeed, the Spearman coefficients are -0.17 and 0.13, respectively. In other words, there are many users with a large percentage of tips with links who posted tips at only a few venues, which cannot necessarily be configured as spamming, according to Foursquare’s rules. Moreover, there are also users who, despite the large percentage of tips with links, did receive a large number of likes and to-dos (see discussion below).

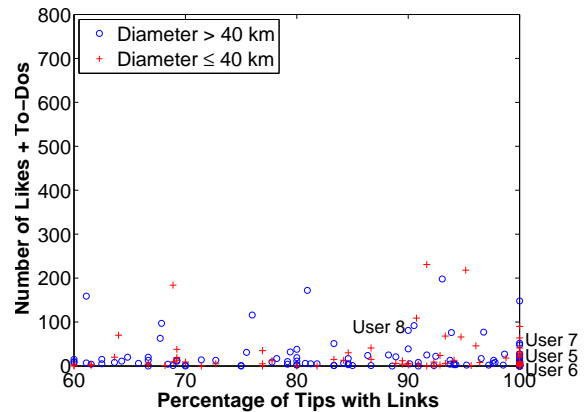
However, Figure 4.7a also shows that several of the selected users post tips with links at a large number of different venues. Take for instance “User 5”, “User 6” and “User 7” in that figure. They post tips at more than 100 different venues, and all tips contain links. These numbers reveal a behavior pattern that is consistent with spamming, and violates Foursquare’s terms of service. Moreover, the total numbers of likes and to-dos for “User 5” and “User 7” are only 12 and 32, respectively, whereas “User 6” receives no feedback. Interestingly, we found no clear correlation between suspicious behavior and diameter of the tipped venues. In other words, our results reveal potential spamming activity both locally and globally.

We note that not all users who posted tips with links at many venues are necessarily engaged in spamming activity, as the linked webpage might be somewhat related to the tipped venue. Take, for instance, “User 8” in Figure 4.7a, who posts 286 tips, 90% of which contains links, at 261 different venues. Despite the large number of tips with links, those tips receive a total of 92 likes and to-dos. We manually investigated this user, finding that it corresponds to a large business chain that placed the same tip to all of its stores advertising a promotion. The tip contains a link to an external webpage that should be visited by those interested in participating to learn more about it. A reasonably large number of users (92) marked the tip as liked or added it to their

to-do lists. In this case, the link pointed to a content that was related to the tipped venue. Indeed, as we further discuss in Section 4.2.2.2, some users who post many tips containing links at many venues aiming at spamming might still receive many likes and to-dos from others. In other words, they might succeed in triggering the interest of many users.



(a) # Tipped Venues vs. % of Tips with Links.



(b) # Likes and To-Dos vs. % of Tips with Links.

Figure 4.7: Correlation between User Attributes (top 3% users with largest percentages of tips with links).

## 4.2.2 Uncovering User Profiles

In the previous sections, we discussed various user behavior patterns observed in our Foursquare datasets, with respect to the use of tips and likes. We here go one step further, and identify user profiles. We do so by applying a clustering algorithm to group users based on three attributes, namely the number of tipped venues, the total number of likes and to-dos, and the percentage of tips with links.

We selected the Expectation-Maximization (EM) clustering algorithm, which is a well-known algorithm used for clustering in the context of mixture models [Dempster et al., 1977]. We ran the EM implementation in Weka [Weka Machine Learning Project, 2012], which has a built-in iterative mechanism to determine the number of clusters. The mechanism is based on ten-fold cross-validation: for each candidate number of clusters, it breaks the data into 10 folds, 9 are used as training sets and 1 as testing set. It builds the clusters on the training sets and, given those clusters, computes the log-likelihood for each instance in the testing set. The log-likelihood values are summed up and then averaged over all 10 folds. The number of clusters selected is the one with maximum (average) log-likelihood.

Table 4.4: Summary of User Attributes Across Clusters.

Attribute	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	Avg	CV	Avg	CV	Avg	CV	Avg	CV
Number of Venues	21.99	0.94	1.97	0.52	13.23	0.52	43.81	1.41
Percentage of Tips with Links	83.11	0.20	3.88	2.35	0.62	5.21	7.02	1.71
Number of Likes and To-Dos	20.41	1.82	7.35	1.52	29.53	2.09	1350.58	5.48
Number of Users	222		190		5660		477	

Because of the large variability observed in the values of the user attributes, particularly number of tipped venues and total number of likes and to-dos, which makes the clustering task harder, we converted all values to a log scale, and normalized the results afterwards. Next, we first present the clustering results (Section 4.2.2.1), and then discuss some findings of a manual inspection of selected users (Section 4.2.2.2).

#### 4.2.2.1 Clustering Results

We applied the EM clustering algorithm over *all users with at least 10 tips*. The algorithm identified 4 clusters, referred to as, throughout this section, clusters 0, 1, 2 and 3. Table 4.4 shows, for each cluster, averages and CVs for each user attribute. It also shows the number of users in each cluster. Complementarily, Figure 4.8 shows, for each cluster, the cumulative distribution function (CDF) of each attribute.

Cluster 0, which includes around 3% of all clustered users, is characterized by a much larger percentage of tips with links (83% on average). This is consistent across most users of the cluster, as shown in Figure 4.8c and summarized by the low CV. Indeed, this attribute clearly distinguishes these users from the others. The number of tipped venues also tends to be large, though smaller than for users in cluster 3. These patterns are consistent with the suspicious, potential spamming, behavior discussed in Section 4.2.1. Moreover, in general, users in cluster 0 do not tend to receive a large number of likes and to-dos.

Cluster 1 consists of focused users who are neither very active nor influential: they tend to post tips at only a few venues, and do not receive many likes and to-dos from others. These are mostly occasional users<sup>1</sup>. Users of cluster 2, on the contrary, are much more active: they tend to leave tips at a larger number of venues, mostly with no links, getting many more likes and to-dos in return. Around 86% of all clustered users are in this group.

<sup>1</sup>We did observe, among users of cluster 1, many tips posted at the same venue within a very short time. For instance, around 10% of users of cluster 1 had an average inter-tip time below 2 seconds, suggesting that the user might have posted the same tip multiple times without knowing. This might reflect lack of experience with the application.

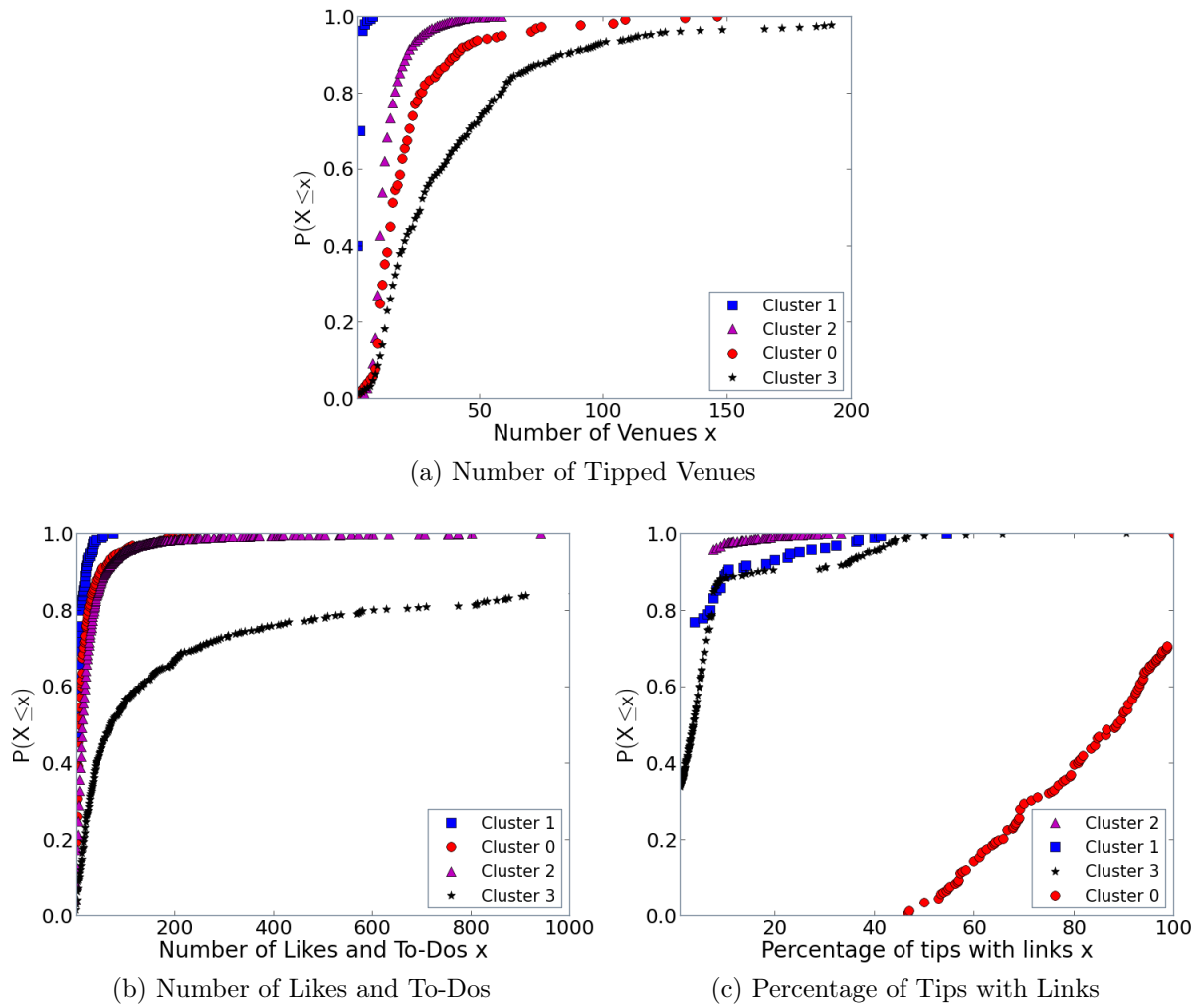


Figure 4.8: User Profiles: Attribute Distributions.

Finally, cluster 3, containing around 7% of the considered users, is characterized by the largest total number of likes and to-dos. These users also tend to post tips at a large number of venues. Therefore, we expect that most very influential users who target a large number of venues fall into this cluster. Moreover, as shown in Figure 4.8a, this cluster also contains users who post tips at only a few venues, indicating that this cluster also contains some very influential but focused users.

We also analyzed the distribution of venues tipped by users in each cluster across the various venue categories maintained by Foursquare. Figure 4.9 shows the distributions. The fractions of venues tipped by users from both clusters 0 and 3 vary only slightly across categories, except for “Colleges & Universities” (Education), which clearly attracts much fewer tips than the other categories. Indeed, it is the least popular category among users of all four clusters. In other words, neither users who seem

engaged in suspicious activity (cluster 0) nor those who tend to be very influential in the system (cluster 3) are focused, collectively, on specific categories. In contrast, the venues tipped by the occasional users (cluster 1) are more concentrated in the “Food” category. The same can be said, to a lesser extent, for users of cluster 2.

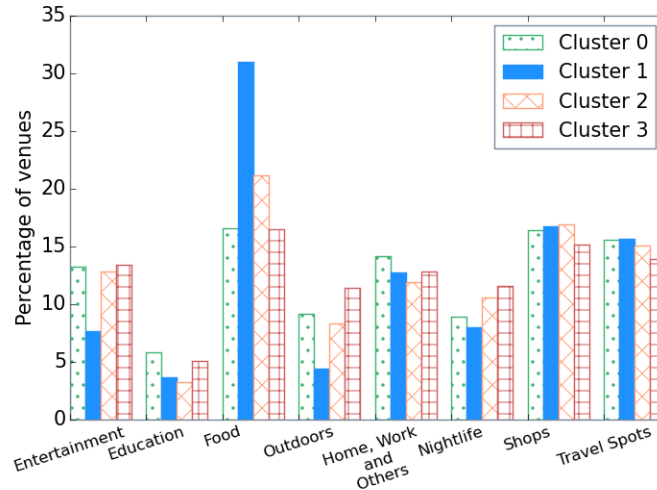


Figure 4.9: Venue Category Distributions.

#### 4.2.2.2 Manual Inspection

As discussed in Section 4.2.1, we did find evidence of suspicious behavior, consistent with spamming activity. Note that this evidence was based only on the percentage of tips posted by a user containing links to external sites and email addresses and, to a lesser degree, on the feedback those tips received from other users. However, an interested spammer may find other ways of reaching users. For instance, a spammer interested in selling a product may write a text advertising it, and post it as a tip. One such case was a tip posted to various venues, including a Japanese restaurant and an university, whose contents advertised a fitness center. The tip’s text is completely unrelated to the nature and business domain of the target venue.

*We here consider as spam a tip whose content is unrelated to the tipped venue, typically an advertisement for a product that is, in nature, unrelated to that kind of venue.* Given this definition, we note that, some spam tips might indeed be successful in getting a large number of likes and to-dos, since many users may find the advertised product interesting, despite it being unrelated to the venue where the tip was posted.

To further investigate the existence of tip spamming in Foursquare, we manually inspected a sample of users from each cluster. For each sampled user, we inspected the



contents of her tips and the venues at which those tips were posted. In case the tip contained a link, the contents of the page pointed to by the link were also inspected.

Each sampled user was inspected independently by three volunteers, who labeled the user as either spammer or not. A user was labeled spammer if the contents of at least 50% of her tips were not related with the nature or domain of the tipped venues<sup>1</sup>. The volunteers were instructed to be conservative: in doubt, they should label the user as *not spammer*. Majority voting was used for final classification, although the volunteers agreed in the vast majority (93%) of the cases. The volunteers also counted down the number of inspected users who are brand users.

Table 4.5 presents our results, showing, for each cluster, the number of inspected users, the number and percentage of them labeled as spammers by the volunteers, and the number and percentage of them identified as brands. The sample size for each cluster was defined so as to have a maximum error in our estimates of 5% with 95% confidence [Jain, 1991].

Most of the users labeled as spammers are, as expected, in the cluster 0 sample. Indeed, all users sampled from that cluster were labeled as spammers, mostly because they posted tips with links pointing to unrelated content. However, our results also show the presence of spammers in clusters 1, 2 and 3. Most of them were classified as such because the text of their tips advertised a product unrelated to the nature of the venue. Moreover, some users of cluster 3 labeled as spammers receive a large number of likes and to-dos, indicating that, despite posting unrelated content, they trigger the interest of many users<sup>2</sup>.

Table 4.5 also shows that there are brand users in all four clusters, although the vast majority of them are in cluster 3. As discussed in Chapter 3, brand users are special Foursquare users who are expected to use the system to reach their followers and potential customers. The brand users of cluster 3 indeed succeeded in promoting themselves in Foursquare: they are very influential in the system, receiving a large number of likes and to-dos. Interestingly, we also found 4 brand users among the 127 users sampled from cluster 0 who were labeled as spammers.

We also analyzed the words commonly used by users labeled as spammers in their tips, contrasting them with the vocabulary of the remaining users. Figure 4.10 shows the word clouds of both user groups. Note that words related to service or product

---

<sup>1</sup>Recall that, given our data collection methodology for the dataset 1, our analysis is based only on a subset of all tips posted by each user.

<sup>2</sup>We note that most of these users receive only a couple of likes and to-dos in each tip. However, because they posted a large number of tips at various venues, collectively, those tips ended up attracting a large number of users.

Table 4.5: Results of the Manual Inspection of a Sample of Users from Each Cluster.

Cluster	Number of Sampled Users	Number of Spammers	Number of Brands
0	127	127 (100%)	4 (3.2%)
1	181	13 (7.2%)	2 (1.1%)
2	237	37 (15.6%)	7 (2.9%)
3	203	41 (20.2%)	58 (28.6%)



Figure 4.10: Words Commonly Used in Users' Tips.

advertisement, such as *business*, *apartment*, *cellular*, *franchise*, *gadget* and, *iphone* are more frequently used by users labeled as spammers.

Finally, we note that the concept of spamming in Foursquare is a very subjective and thus controversial matter, possibly even more than in other systems. What one classifies as spam, another person might interpret as creative marketing strategy. Take for instance the case, reported in Catalyst Marketers Blog [2010], of a tip advertising a product sold by a certain venue  $v_i$ , posted at another venue  $v_j$ , which is indeed a business competitor of  $v_i$ . When customers check the tips left at  $v_j$ , they will see the advertisement and might indeed be drawn towards its competitor. In other words, the tip might contribute to steal business away from the venue at which it was left<sup>1</sup>. Moreover, obviously of any marketing game, users might find the tip interesting, and mark it as a to-do or as liked. This might happen even for tips advertising products completely unrelated to the tipped venues, as discussed above. Thus, identifying and dealing with tip spamming and spammers is a hard task. Some recent efforts to address this problem were reported by Costa et al. [2013]; Aggarwal et al. [2013].

In this section, we observed the presence of four user profiles defined based on their tipping activity. In the next section, we investigate the properties of the network built from user interactions through tips.

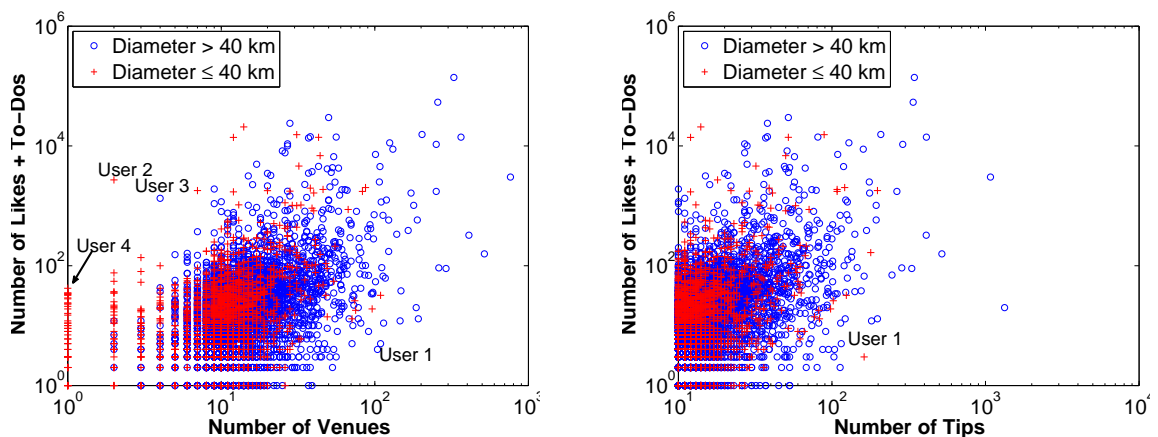
<sup>1</sup>During our manual inspection, we took a conservative approach, and *did not* consider such cases as spamming.

## 4.3 User Influence

We now further focus on user interactions established among users through the use of tips and likes. Recall that user influence is important for estimating the popularity of the tips posted by the user. To that end, we present an empirical analysis of user influence patterns on Foursquare. We start analyzing influence as the number of likes and to-dos received by the tips of each user. Focusing on the geographical location of the venue where the tips were posted, we present a preliminary analysis about how influence is geographically distributed (Section 4.3.1). Next, we present other ways to measure influence, defining a graph composed by users who interact to each other either posting tips or marking them as like (Section 4.3.2). Thus, we proposed a method based on PageRank to rank influential users in Section 4.3.3. In this analysis we used the dataset 2.

### 4.3.1 User Behavioral Patterns

Figure 4.11a shows the number of tipped venues versus the total number of likes and to-dos received by each user. Note the logarithm scale in both axes. As in the previous sections, we group users into those with diameter shorter than or larger than 40 kilometers. Figure 4.11b shows a similar graph, plotting the number of tips versus the total number of likes and to-dos of each user. To improve graph readability, both figures only show users with at least 10 tips. Note that, given the strong correlation between number of tips and number of tipped venues per user, the two graphs are very similar.



(a) Number of Venues vs. Number of Likes and To-Dos.

(b) Number of Tips vs. Number of Likes and To-Dos.

Figure 4.11: Correlation between User Attributes (only users with at least 10 tips).

Both graphs show the existence of different classes of users. On one hand, we find users who, despite posting a large number of tips and/or posting tips at a large number of venues, receive only a comparatively small number of likes and to-dos. One such user, marked as “User 1” in both graphs, posted approximately 104 tips at 100 different venues, but received only 4 to-dos and likes. By manually inspecting a sample of those users, located in the right bottom corner of the graphs, we found that they tend to be ordinary users who, in spite of being very active in the system and contributing with a lot of tips regarding many different venues, do not get a lot of feedback from others.

In contrast, the right top corner of the plots show users who not only post a large number of tips at a large number of different venues but also receive a lot of feedback on them. Those users, many of which are famous brands such as “Bravo”, “History Channel”, “The Wall Street Journal” and “National Post”, seem engaged in providing many recommendations (tips) on a large variety of venues, and clearly succeed in reaching and attracting the attention of many users. Clearly, those users are very influential in the system. Note that, interestingly, the graphs show users with large numbers of tips (and tipped venues) as well as large numbers of likes and to-dos in both user sets, suggesting the presence of local and global activity.

Both graphs also show some users who, despite posting only a few tips at a few venues, receive a comparatively very large amount of feedback, and thus can also be considered very influential. Examples are “User 2” and “User 3” in Figure 4.11a, who received 2708 and 1337 likes and to-dos, respectively, despite targeting only a couple of venues with a few tips. Some of these highly focused influential users are brands, such as “Six Flags”, while others are ordinary users (e.g., “User 4”). Once again, we found focused users (i.e. user who post tips at the same venues) with strong activity both locally and globally.

As a side note, we point out that, as expected, both graphs show a slight trend for users with larger numbers of tips and tipped venues also having larger diameter. Perhaps more interesting is the presence of some very active users, with tens to hundreds of tips and tipped venues, who are focused on venues in a local region (diameter under 40 kilometers).

### 4.3.2 User Influence Network

We now further focus on user influence analyzing the interactions established among users through the use of tips and likes<sup>1</sup>. To that end, we focus on venues of each category separately, and build a user network representing the relationships established when a

---

<sup>1</sup>We disregard to-do marks here.

user likes a tip posted by another user at a venue of a specific category. More precisely, we build a directed weighted graph where each node represents a user, and an arc from a node  $u_i$  to a node  $u_j$  indicates that  $u_i$  liked a tip posted by  $u_j$ . The weight of an arc indicates the number of likes from  $u_i$  to tips posted by  $u_j$  at venues of the category. We also build a user network considering tips and likes at venues of all categories, referring to it as General.

Table 4.6 summarizes the main characteristics of the graph built for each venue category and for the general graph. Some basic graph properties such as the number of nodes, i.e. the number of users who received or gave at least one like, as well as the number of arcs in each graph are reported. The table also shows the average node degree, the number of reciprocal arcs (two users who liked each other’s tips), and the size of the largest strong connected component (SCC) for each graph. The SCC represents the set of nodes that can be reached from any other, following the arcs from the user who liked a tip to the tip’s author. A larger SCC indicates the presence of a community where users interact posting and liking each others’ tips. Complementary data is also shown in Figure 4.12, which provides the indegree and outdegree distributions for the user network built for each venue category. The degree distribution is a function describing the number of users in the network with a given degree (number of neighbors).

Table 4.6: Summary Statistics for User Influence Networks per Venue Category as well as for All Categories (General).

Category	# Nodes	# Arcs	Avg degree	Reciprocal arcs	SCC
General	1,143,914	2,283,949	2.0	61,233 (2.75%)	147,282
Arts	177,876	222,952	1.25	1,288 (0.60%)	116
Education	95,274	94,608	0.99	1,766 (1.90%)	16
Food	650,543	964,012	1.48	17,497 (1.85%)	34,117
Outdoors	140,479	140,559	1.00	979 (0.70%)	28
Nightlife	253,631	309,081	1.22	5,421 (1.79%)	1,559
Professional	221,091	199,578	0.90	4,104 (2.10%)	26
Residences	138,320	98,319	0.71	5,336 (5.74%)	9
Shops	321,049	362,310	1.00	4,094 (1.14%)	31
Travel	186,970	223,881	1.20	1,451 (0.65%)	361

As we observe in Figure 4.1b, a few users receive a proportionally large number of likes in their tips (high indegree) and a few users give a significant number of likes (high outdegree). Although all categories exhibit heavy tailed distributions, we see that the *Food* graph is the largest graph in terms of nodes, arcs and node average degree, which shows that this category has the highest level of user activity in terms of likes. Indeed,

analyzing the indegree distributions, we observe that *Food* and *Nightlife*, which have the distributions with the longest tails, are categories where the tip authors receive the most amount of feedback, i.e., the largest number of likes in their tips. Moreover, the outdegree distributions suggest that Foursquare users have more interest in tips from the *Entertainment* and *Food* categories.

On the contrast, the *Residences* category has one of the smallest graphs, confirmed by the shorter tails of its indegree and outdegree distributions (Figure 4.12) and a small numbers of nodes and arcs. This reflects the lower user activity in this category, and can be explained by the fact that the venues in such category represent private locations, usually the user residence (its geographical location is only revealed to the user’s friends). Even though the tips are publicly available to all users, the contents of the tips in this category may not be of great interest to other users, which is also reflected by the small size of the strong connected component and the relatively large fraction of reciprocal arcs.

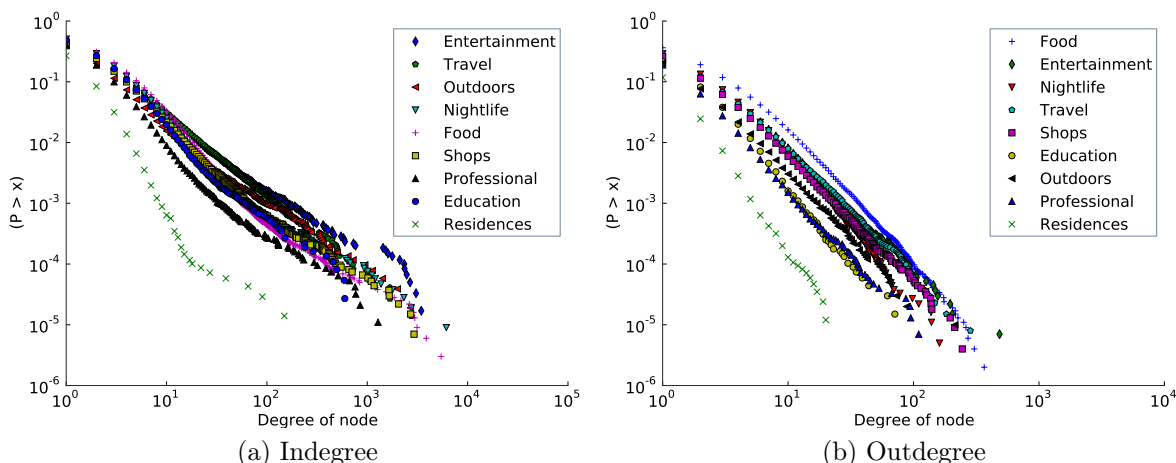


Figure 4.12: Degree Distribution of the User Network (log scale).

### 4.3.3 Measuring User Influence

The most simple and intuitive approach to estimate user influence on Foursquare based on tipping activity is to compute the number of likes received by each user. By using this strategy, we would consider a user who received 100 likes on a single tip as influential as a user who posted 100 tips and received one like in each. One could argue that the former should be considered more influential than the latter. An alternative ap-

proach is PageRank, a method used in other contexts to identify experts of influential users, that will be explained in Section 4.3.3.1.

We also propose a new method based on PageRank that addresses the shortcoming of the number of likes approach (baseline) and traditional Page Rank, by considering both the structure of the network and the number of tips posted by each user. We refer to this method as Normalized PageRank (Section 4.3.3.2). We compare our proposed approach against these two strategies (number of likes and traditional PageRank) using the graphs for each venue category in Section 4.3.3.3.

#### 4.3.3.1 Measuring User Influence using PageRank

PageRank is a link analysis algorithm proposed by Page et al. [1998] for ranking web pages. The ranking system, based on a random walk algorithm, evaluates the probability of finding a random surfer on any given page. The algorithm assumes that the presence of a hyperlink from page  $i$  to page  $j$  is an evidence of the importance of page  $j$ . In addition, this importance is determined by the importance of  $i$  itself, and is inversely proportional to the number of pages  $i$  points to.

Intuitively, we can interpret the distribution of PageRank values in terms of a random walk [Henzinger et al., 1999]. Let’s consider the case of a network representing web pages connected by hyperlinks. The PageRank value of a page can be interpreted as the fraction of time a random surfer would spend visiting the page by interactively following links from page to page [Zhang et al., 2007]. In other words, if a surfer visits page (node)  $i$ , the random walk is in state  $i$ . At each step, the Web surfer either follows a link chosen uniformly at random from those on the current page with probability  $\alpha$  (damping factor), or the web surfer jumps (teleport) to any other page on the Web chosen uniformly at random (with probability  $1 - \alpha$ ).

By regarding pages as nodes and hyperlinks as arcs between nodes, Equation 4.2 computes the PageRank  $P(j)$  of a node  $j$ , which belongs to a network of size  $N$ :

$$P(j) = \frac{(1 - \alpha)}{N} + \alpha \sum_{i \in M(j)} \frac{P(i)}{N_i} \quad (4.2)$$

where,  $M(j)$  is the set of nodes that have an direct arc to  $j$ ,  $P(i)$  and  $P(j)$  are the PageRank values of nodes  $i$  and  $j$  respectively, and  $N_i$  is the number of edges coming out of  $i$ . The PageRank formula consists of two components weighted by the damping factor  $\alpha$  ( $0 \leq \alpha \leq 1$ ), usually set to 0.85<sup>1</sup>. The first component  $\frac{1}{N}$  represents the probability that a surfer will jump to  $j$  from any other random node from the network,

<sup>1</sup>We tried various values around 0.85, but it did not make significant difference.

while the second component  $\sum_{i \in M(j)} \frac{P(i)}{N_i}$  models the contribution of incoming arcs to node  $j$  normalized by the number of incoming arcs ( $\frac{1}{N_i}$ ).

The key idea of PageRank is to allow propagation of influence along the network of web pages, instead of simply counting the number of web pages pointing at the web page. This algorithm has proven to be a useful tool for ranking nodes in a graph in many contexts, such as the identification of potential experts in specialized forums [Zhang et al., 2007], influential users on Twitter [Haveliwala, 2002; Weng et al., 2010; Kwak et al., 2010] and, spam detection [Gyöngyi et al., 2004].

In our context, PageRank can be applied to identify influential users in terms of their tipping activity, using the same user network defined in Section 4.3.2. As already stated, PageRank computes the influence of a user using the direct and indirect influence of all other users by propagating the individual influences over the network. Thus, a user has a high probability to be influential if an influential user has liked his tip. A drawback of this approach, in the specific case of Foursquare, is that the number of tips posted by the user is not considered. For instance, if a user got 100 likes in 10 tips while another user got the same number of likes in 200 tips, the method would favor the latter. Once again, one might argue that an influential user should consistently produce popular content, i.e., their influence should hold across all their tips.

Therefore, we propose an alternative strategy based on the original PageRank algorithm, which we call Normalized PageRank, described in the next section. Our algorithm takes into account: the number of likes received, the number of tips posted by the user, and the PageRank of the users who liked the tips of this user. We assume that a user should be considered influential if she has a large average number of likes in her tips and receives likes from other influential users.

#### 4.3.3.2 Normalized Page Rank

The Normalized PageRank proposed here differs from the traditional PageRank by taking the number of tips posted by the user into account. We weigh the arcs by how frequently one user likes tips from another. Thus, for the Normalized PageRank, the weight  $I$  of each arc  $(i, j)$  is weighted by the number of tips posted by user  $j$  that received a like from user  $i$ :

$$I(i, j) = \frac{w(i, j)}{t(j)} \quad (4.3)$$

where  $w(i, j)$  is the number of likes given by user  $i$  in tips posted by user  $j$ , and  $t(j)$  is the total number of tips posted by  $j$ . This definition captures and quantifies the success



of the tips posted by the user  $j$  in terms of the amount of feedback (likes) received from others. Therefore, the larger the number of tips posted by  $j$  are marked as liked by  $i$ , the greater is the influence of  $j$  on  $i$ . The normalized PageRank is computed as follows:

$$P(j) = \frac{(1 - \alpha)}{2} \left[ \frac{1}{N} + \frac{t(j)}{\sum_{k \in V} t(k)} \right] + \alpha \sum_{i \in M(j)} I(i, j) * P(i) \quad (4.4)$$

where  $\alpha$  is the damping factor;  $N$  and  $V$  are, respectively, the total number and the set of users represented in the graph, and  $M(j)$  is the set of arcs pointing to  $j$ , i.e. the set of users who liked any tip posted by  $j$ .

As in Equation 4.2, the rank score  $P(j)$  of a user is composed of a part corresponding to the rank contribution from the other users linked to  $j$  and a part corresponding to the probability of a random jump to  $j$  from any other user represented in the network. The probability of selecting any user during a random jump is also divided into two components: the first one ( $\frac{1}{N}$ ) has the same goal of the traditional PageRank, and the second part makes the probability of random jump to  $j$  proportional to the number of tips posted by  $j$ , i.e.,  $t(j)$ . The factor  $\frac{1}{\sum_{k \in V} t(k)}$  normalizes this probability by the total number of tips used to built the entire graph. Thus, this part of the formula makes that users who publish many tips have a higher probability to get visited by the random surfer. Finally, the factor  $I(i, j)$  defined in Equation 4.3 weighs the PageRank  $P(u)$  of each user  $u$  that likes at least one tip posted by  $v$ .

### 4.3.3.3 Experimental Results

In the previous section, we proposed a novel ranking model, Normalized PageRank, describing the modifications performed on the traditional PageRank to capture the number of tips posted by each user and the frequency that a user likes a tip posted by another user. Any measure of influence is necessarily subjective [Romero et al., 2011]. Thus, there is no clear ground truth of what should be the ranking of users by influence. Thus, as in previous work focused on social influence [Zhang et al., 2007; Cha et al., 2010; Weng et al., 2010], we assess the performance of our proposed method by comparing it against other algorithms: a baseline method based solely on the number of likes received, and against the traditional Page Rank method (defined in Section 4.3.3.1). Our goal here is not to point out which is the best method, but show differences in the three approaches.

Moreover, as in previous studies [Zhang et al., 2007; Weng et al., 2010], our main

evaluation metric is the correlation between the rank lists generated by the different methods, measured by the Kendall  $\tau$  coefficient [Kendall and Gibbons, 1990]. This coefficient is in the range of  $-1 \leq \tau \leq 1$ :  $\tau = 1$  means that the two lists are exact the same whereas  $\tau = -1$  implies that one list is the reverse of the other. Table 4.7 lists the  $\tau$  values between the rank lists generated by each pair of methods, whereas Table 4.8 lists the top-5 most influential Foursquare users, using each of the graphs described in Table 4.6, according to each method.

Table 4.7: Kendall  $\tau$  Correlation Values Between Rankings Lists.

<b>Category</b>	<b># likes vs. PageRank</b>	<b># likes vs. Normalized PageRank</b>	<b>PageRank vs. Normalized PageRank</b>
General	-0.0048	0.0985	0.5041
Entertainment	-0.0421	-0.0819	0.5100
Education	-0.0086	-0.0617	0.5412
Food	-0.0371	0.0220	0.4859
Outdoors	-0.0861	-0.1503	0.4758
Nightlife	-0.0253	-0.0349	0.5134
Professional	-0.1236	-0.1793	0.4654
Residences	-0.3187	-0.3101	0.4296
Shops	-0.0702	-0.0829	0.4637
Travel	-0.0677	-0.0967	0.4711

Table 4.8: Top-5 Most Influential Users Overall and per Venue Category According to Each Method.

Category	Number of likes	PageRank	Normalized PageRank
<b>General</b>	History Channel, Bravo, MTV, Wall Street J., Visit PA	History Channel, Bravo, MTV, Wall Street J., Zagat	History Channel, Bravo, MTV, Wall Street J., Zagat
<b>Entertainment</b>	History Channel, Explore Chicago, Visit PA, Wall Street J., NHL	History Channel, Wall Street J., NHL, Visit PA, Explore Chicago	History Channel, Wall Street J., La Vida, <sup>u</sup> TLC,Explore Chicago
<b>Education</b>	Arizona State U, Mizzou, UW-Madison,Cal, Stanford U	History Channel, Sports Authority, Mizzou, Stanford U., Graphic Master <sup>c</sup>	bookrenter.com, Northeastern U., Sports Authority, Mizzou, old main <sup>u</sup>
<b>Food</b>	Bravo, Zagat, Eater.com, Wall Street J., Thrillist	Bravo, Zagat, Thrillist, Visit PA, Wall Street J	Bravo, Foodspotting <sup>c</sup> ,Zagat, Britt r, <sup>u</sup> Wall Street J.
<b>Outdoors</b>	History Channel, Wall Street J., Explore Chicago, Windows Live P.G., Visit PA	History Channel, Wall Street J., Explore Chicago, Bravo, Windows Live P.G.	History Channel, Wall Street J., Explore Chicago, Healthy Pools <sup>u</sup>
<b>Nightlife</b>	MTV, Bravo, LogoTV, Thrillist, History Channel	MTV, Bravo, Thrillist, History Channel, LogoTV	Bravo, MTV, Thrillist, History Channel, Logo TV
<b>Professional</b>	History Channel, Wall Street J., Explore Chicago, National Post, Huffpost	History Channel, Wall Street J., Explore Chicago, Visit PA, Graphic Master <sup>c</sup>	History Channel, oregonvotes.org, <sup>u</sup> USF Athletics <sup>u</sup> , Wall Street J., Liam B. <sup>u</sup>
<b>Residences</b>	History Channel, Greensboro NC <sup>u</sup> , Huffpost, Lee <sup>c</sup> , Mizzou	History Channel, Huffpost, Lee <sup>c</sup> , Mizzou, Supermodelme	David K, <sup>u</sup> Beth D, <sup>u</sup> Lee <sup>c</sup> , sarah w, <sup>u</sup> capture the market <sup>u</sup>
<b>Shops</b>	Bravo, Visit PA, Graphic Master <sup>c</sup> , Mazda, (red)	Bravo, Visit PA, Graphic Master <sup>c</sup> , History Channel, Mazda	Bravo, A bag's life, <sup>u</sup> AT&T, cew, <sup>u</sup> Visit PA
<b>Travel</b>	History Channel, National Post, Wall Street J., National Post, KLM	History Channel, Wall Street J., Bravo, KLM, Explore Chicago	History Channel, Wall Street J. KLM, AT&T, Tubus <sup>u</sup>

<sup>c</sup> celebrity user<sup>u</sup> ordinary user

We note that the baseline method generates a ranked list different from those generated by the other two methods since the values of  $\tau$  are very far from 1 for all graphs. It is observed that the Normalized PageRank has a higher agreement with the traditional PageRank than with the baseline method. Considering the rankings built for users across all categories (general), four of the five most influential users appear in the three rankings (see Table 4.8). However, the two strategies based on PageRank are able to identify one user (Zagat) that was not listed as one of the top-5 users by the baseline.

Moreover, some of the top-5 most influential users across all categories do not appear in the top-5 of some venue categories which suggests that some of them are more influential in certain categories. For instance, the *MTV* user was identified by the three methods as top influential in the *Nightlife* category. We also observe that different methods identify different users as top-5 in some categories. For example, the *Foodspotting* user is among the top-5 most influential users in the *Food* category listed only by the Normalized PageRank method, while *Greensboro NC (Residences)* and *History Channel (Shops)* are listed only by the baseline and traditional PageRank approaches lists, respectively. We also note that the differences between the top-5 most influential users identified for each method are larger for categories with smaller number of users who post tips (e.g., *Education*, *Residences*, and *Professional*)

We also observe that most of the top-5 influential users in any list are composed by brand users. This is expected since these users are engaged to promote their businesses, providing many tips on a large variety of venues. Their presence on top positions of the ranks suggests that they are succeeding in attracting attention to the their posted content. In addition, we note that some methods are able to identify some ordinary users as influential in some of the Foursquare categories. For instance, the *Greensboro NC* user is listed by the baseline method among the top-5 most influential users in the *Residences* category.

Our results show that there are differences between the general ranking lists (all categories) and the lists generated for each venue category. Thus, users that are influential in general, are not as influential in certain categories. This observation is important for the task of automatically predicting the tip's popularity, since a tip posted by a user in a venue category which he is an influential/expert has more potential to become popular. Moreover, while most of the top-5 listed users are brand users, the PageRank based approaches are able to identify ordinary users and celebrities as top influentials in some categories.

Finally, the proposed method offers other few contributions that are not entirely related to our scope of popularity prediction. The ranking generated by the baseline is

more susceptible to the attack of malicious users than the PageRank based approaches, since malicious users can easily create many identities to inflate the number of likes received by a user. Even though PageRank is more robust to attacks than the baseline, there are studies [Du et al., 2007; Adal et al., 2012] that describe methods for artificially boosting pageranks scores (e.g., link farms).

Moreover, the presented techniques can be exploited to improve search and content recommendation services (e.g., by prioritizing content posted by influential users), as well as detection of the malicious users. Foursquare has a limited number of moderators who are responsible for filtering malicious activities. The Normalized PageRank method minimizes the recommendation of spammers since our method takes into account the user relative influence and the amount of feedback received from other users. We believe that this methodology can be used on other social networks where influence is measured based on the interaction between users (e.g., recommendation of posts on Facebook, photos on Instagram, etc.). As future work, a interesting direction is the comparison with another link-based ranking algorithm, for example the HITS method [Kleinberg, 1999].

## 4.4 Dynamics of Tip Popularity Evolution

In this section, we analyze the dynamics of tip popularity in Foursquare. We start by discussing how the number of likes of a tip evolves over time (Section 4.4.1), and how it is affected by the social network of the tip’s author (Section 4.4.2). We then analyze tip popularity at and around the peak (Section 4.4.3), and assess to which extent the rich-get-richer phenomenon is present in the popularity evolution of tips (Section 4.4.4). For these analyzes, we used the dataset 2.

For the sake of analyzing tip popularity dynamics, we group tips with at least one like by breaking their popularity distribution (Figure 4.6a) into 10 slices, each one containing tips whose popularity fall into a certain range of the distribution<sup>1</sup>. For example, slice 0-10% contains the top-10% most popular tips, while slice 10%-20% contains the tips whose popularity fall between the 10<sup>th</sup> and 20<sup>th</sup> percentile of the popularity distribution. This partitioning is the same used in van Zwol [2007] for analyzing Flickr photos, since it is more balanced and less biased towards the more popular tips. Table 4.9 shows the number tips as well as total number of likes per slice.

We also examine the fraction of likes coming from the social network (friends and followers) of the user who posted the tip (i.e., the tip’s author). Table 4.9 shows

---

<sup>1</sup>Note that we exclude tips with no likes from these slices.

Table 4.9: Distribution of Likes for Groups of Tips

Slice	# of Tips	Total # of Likes	% Social Likes	Group
0-10%	23,746	202,804	30.8%	$G_1$
10-20%	23,746	72,824	48.4%	$G_2$
20-30%	23,746	47,492	49.0%	$G_3$
30-40%	23,746	47,492	49.0%	$G_3$
40-50%	23,746	24,163	48.2%	$G_4$
50-60%	23,746	23,746	49.1%	$G_4$
60-70%	23,746	23,746	48.5%	$G_4$
70-80%	23,746	23,746	48.2%	$G_4$
80-90%	23,746	23,746	48.5%	$G_4$
90-100%	23,750	23,750	48.4%	$G_4$

the percentages of likes coming from the social network, referred to as *social likes*, for tips in each slice. We note that for all slices but the first one, almost half of the likes received by tips come from the user’s social network, highlighting the importance of friends and followers to the popularity of those tips. In contrast, for the most popular tips, the fraction of social likes is smaller (31%), suggesting that most likes probably come from venue visitors. We further analyze the importance of the social network to tip popularity in Section 4.4.2.

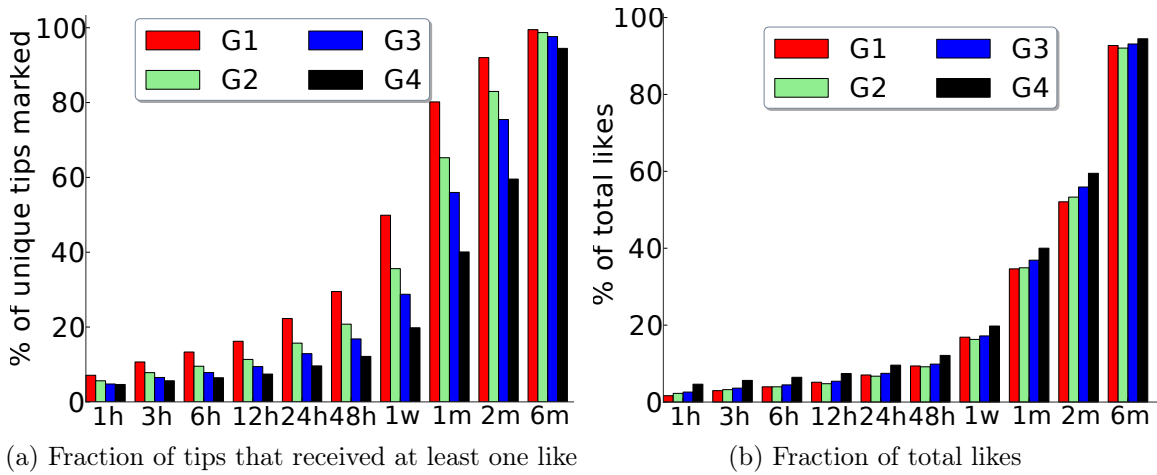


Figure 4.13: Distribution of Tip Popularity over Time.

We aggregate the slices into 4 major groups, as shown in Table 4.9. Groups 3 and 4 contain tips that received, on average, 2 and 1 likes, respectively. We analyze tip popularity separately for each slice. However, as the same conclusions hold for tips in different slices of the same group, we present results for each group only.

### 4.4.1 Popularity Evolution

We start by analyzing how the popularity of tips in each group of slices defined in Table 4.9 evolves over time. We focus on the first six months after the tip is posted. Figure 4.13a plots the fraction of unique tips in each group that received at least one like within the first  $x$  hours (h), weeks (w) or months (m) after posting time. We observe that within the first 48 hours, 29% of the tips in the most popular group (G1) had already received at least one like, while in one and two months this fraction grows up to 80% and 92%, respectively. That is, 20% of the top-10% most popular tips take more than one month to attract their first likes. This slow popularity evolution is even more clear for tips in the other (less popular) groups. Figure 4.13b shows the cumulative fraction of the total number of likes (as observed in our dataset) received by tips in each group over time. Note that, for all four groups, between 41% and 48% of the likes are received *after* 2 months since posting time.

Thus, in general, tips tend to live long in the system, presenting a gradual increase of interest. Indeed, tip popularity evolves much more slowly compared to other types of content, even for tips that end up becoming very popular. For example, news articles have a very short lifespan [Tatar et al., 2014] acquiring all comments within the first day of publication, while a large fraction of *views* of Flickr photos are generated within the first two days after upload [van Zwol, 2007]. In contrast, we here find a significant fraction of tips that can take quite a few months to attract likes and become popular. This longer lifecycle was also observed in the acquisition of *fans* by Flickr photos [Cha et al., 2009].

We further analyze the popularity evolution of tips in each group by showing in Figure 4.14 the curves of the 10<sup>th</sup> and 90<sup>th</sup> percentiles as well as the median of number of likes over time during the first one month since the tip was posted. For all groups, the 10<sup>th</sup> percentile curve is equal to zero through the whole period, implying that 10% of the tips in each group did not receive any like within the first month in the system. Around half of the most popular tips (G1) starts receiving likes after 7 days since posting time, achieving only 20% of the total likes after a month. For the second most popular group (G2), we note that half of the tips start receiving likes after 15 days. In contrast, tips in group G3 and G4 take more than 20 and 30 days, respectively, to start attracting likes.

We also analyze the amount of time it takes for a tip to receive at least  $X\%$  of their total likes, for  $X$  equal to 10, 50, 70, 90 and 100%. Figure 4.15 shows those distributions for the most popular tips (G1). Note that 57% of the tips in this group take at least 2 (3) months to reach 50% (70%) of its total observed popularity. In sum,

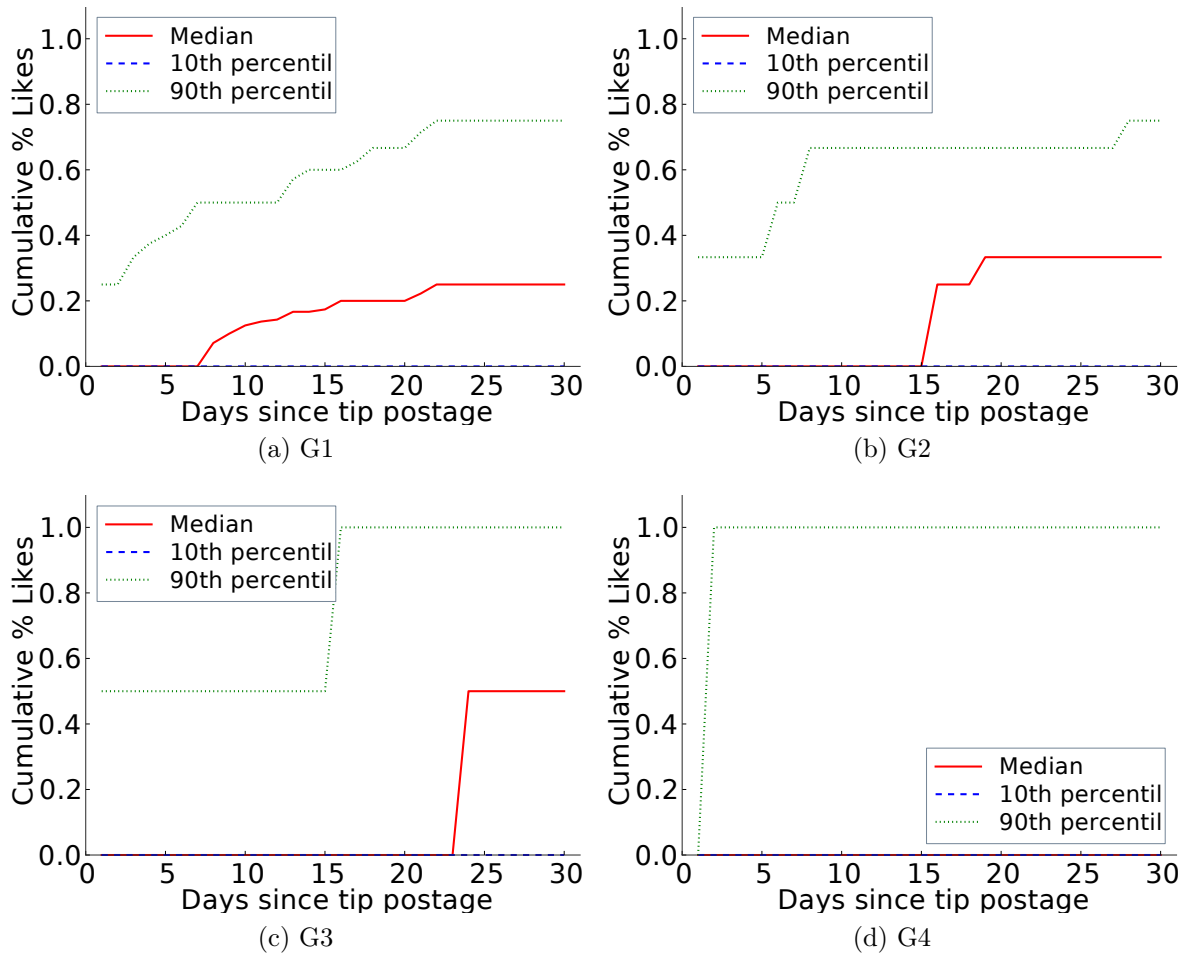


Figure 4.14: Distribution of Percentage of Likes Received During the First Month after Posting Time.

many tips do take a few months to attract likes, even those that end up being the most popular ones.

The great variability in popularity evolution across tips, allied to the somewhat slow popularity dynamics observed in the figures above, motivate the use of prediction methods that exploit early popularity measurements. Yet, they also raise a question as to whether the joint use of other features along with such measurements can improve prediction accuracy over exploiting only the latter (as in Szabo and Huberman [2010]; Pinto et al. [2013]). Similarly, the slow popularity evolution also raises a question as to how robust the prediction models are to long-term predictions. We address these questions in Chapter 6.4.



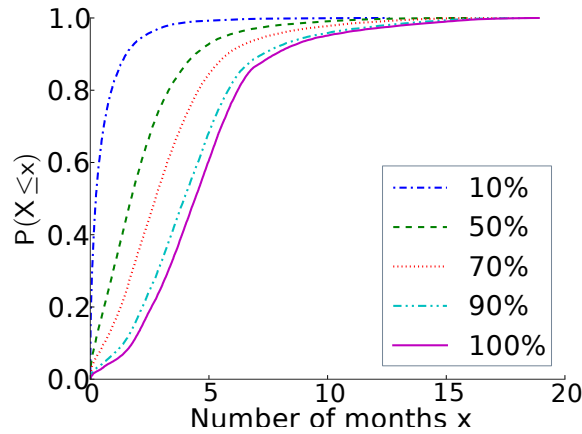


Figure 4.15: Distribution of Time Until  $x\%$  of Total Likes are Received for the Most Popular Tips (G1)

#### 4.4.2 The Role of the Social Network

The popularity evolution of a tip is directly related to how users find the tip: either by visiting the venue page or through activity notifications from their friends and followees. Thus, the number of likes received by a tip depends on a combination of its visibility and interest by the social network of the tip's author and by others.

Recall that in Section 4.1.1, we briefly analyzed the role of social network, concluding that the network influences the tip popularity, but there we focused on aggregate popularity. Now, we assess the role of social network of the tip's author on how the tip popularity *evolves over time* for tip with different popularity groups. To that end, we revisit Figure 4.13b by separating likes coming from the author's social network (social likes) and likes coming from other users (non-social likes). Figure 4.16 shows the cumulative fraction of likes, in both categories, for tips in each group. Note that the author's social network has an important influence on the popularity of a tip throughout its lifetime: at least half of all likes received in any period of time (up to 6 months since posting) come from the author's social network, for tips in *all* four groups. This fraction is higher in the earlier periods after posting time, and tends to decrease with time as the tip becomes visible to other users (e.g., venue visitors). For example, the social likes correspond to 62% of all likes received by the most popular tips (G1) in the first hour since posting time, decreasing to 54% after 6 hours. Interestingly, the social network seems to have an even more important role for the least popular tips. For example, for tips in G2, G3 and G4, the social likes correspond to more than 70% of all likes received by a tip in the first week in the system.

These results indicate that the social network of a tip's author may be responsible for boosting its popularity, particularly during early periods after posting. As

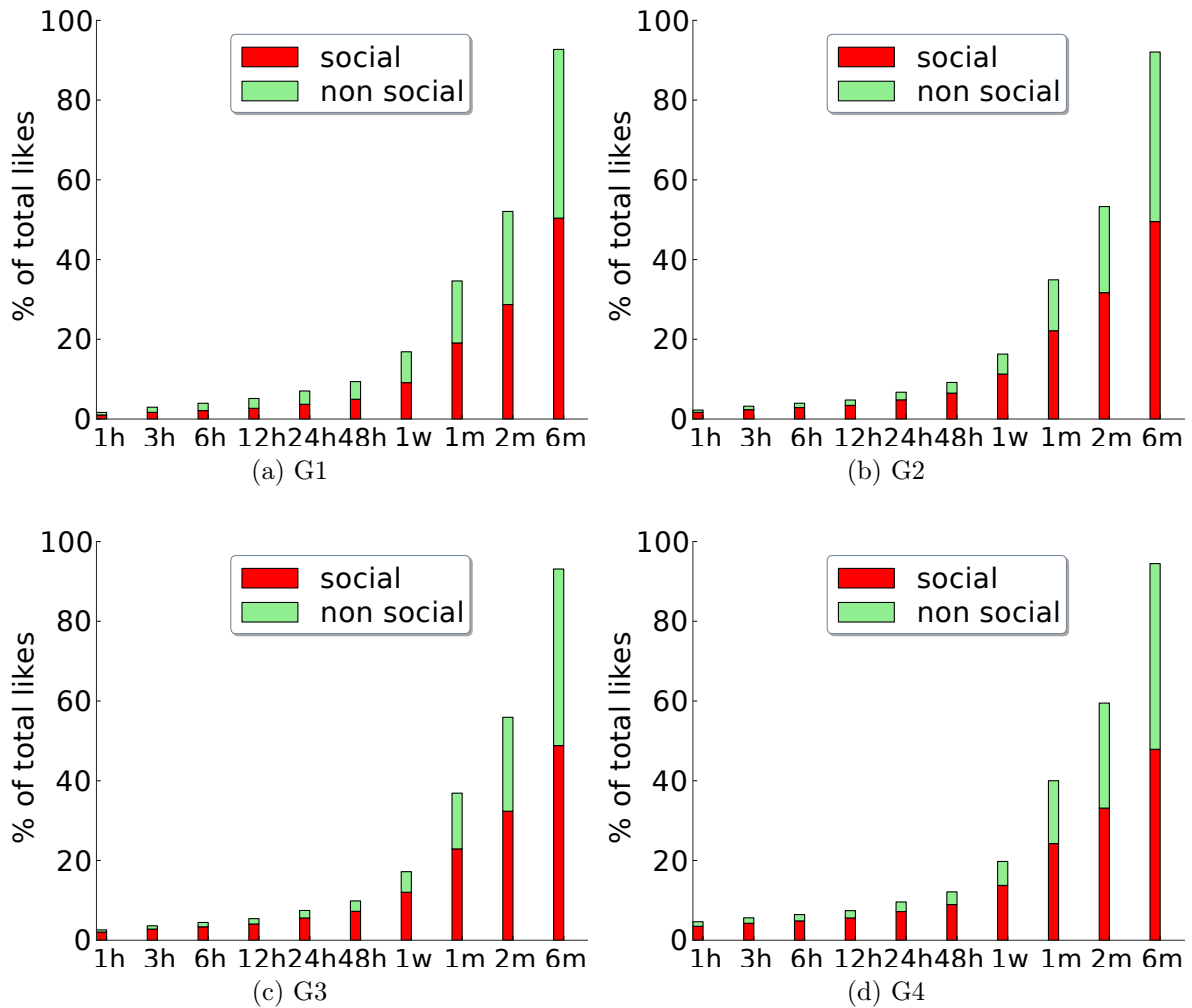


Figure 4.16: Social vs. Non Social Likes: Distribution of Percentage of Likes Received over Time.

consequence, they also suggest that it might be possible for a recently posted tip to become more popular than other tips that had already attracted many likes and thus gained visibility in the system.

### 4.4.3 Popularity Peak

We further analyze tip popularity evolution by focusing on the popularity peak. Considering the daily popularity time series of each tip, we define the peak  $k_{p_i}$  of tip  $p_i$  as the largest number of likes received by  $p_i$  on any single day. We then compute the time (in number of days) it takes for  $p_i$  to reach its popularity peak<sup>1</sup>. We also measure the

<sup>1</sup>In case of ties, we pick the first day with  $k_{p_i}$  likes.

fraction of the total likes  $p_i$  received *at*, *before* and *after* the peak. For this analysis, we focus on the most popular tips (G1).

Figure 4.17a shows the cumulative distribution of the time until the popularity peak. Around 18% of the tips experience its popularity peak one day after posting time, and around 72% of the tips reach their popularity peak within a month since posting. This implies that most tips do not take too long (less than a month) to reach its daily popularity peak. Yet, we observe that, for many tips, this peak represents only a small fraction of the total observed popularity. This is illustrated in Figure 4.17b, which presents the cumulative distributions of the median, 10<sup>th</sup> and 90<sup>th</sup> percentiles of the fraction of likes received at and after the peak day. As a complement, Figure 4.17c shows the cumulative distribution of the fraction of likes received *before* the peak day. Like observed for other types of online content (e.g., videos and news [Crane and Sornette, 2008; Pinto et al., 2013; Tatar et al., 2014]), some tips do experience heavy bursts of popularity on the peak day: for 10% of the tips, the daily peak corresponds to at least 67% of their total popularity (see 90<sup>th</sup> percentile curve in Figure 4.17b).

However, for half of the tips (median curve), the peak corresponds to only 25% of all likes. Moreover, Figure 4.17c shows that most tips (82%) receives their first like in the peak day, and only a very small fraction of the tips (3.3%) receive more than 50% of the likes before the peak day. Thus, a large fraction of tips receive most of their likes *after* the peak day, suggesting, once again, that tips experience a slow popularity evolution.

Contrasting our findings to the acquisition of fans by Flickr photos [Cha et al., 2009], we observe that both photo fans and tip likes tend to be acquired after a longer period of time after posting/upload, compared to, for example, tweets. Also, as in Cha et al. [2009], we do not observe an exponential growth on popularity as suggested by existing models of information diffusion [Valente, 1995; Figueiredo et al., 2014a]. However, comparing our results (particularly Figure 4.14), with similar ones presented in Cha et al. [2009], we find that tip popularity seems to increase even more slowly than photo fans on Flickr. For example, we do not observe a period of steady linear popularity growth during the first month, as observed for photos.

#### 4.4.4 The Rich-Get-Richer Phenomenon

Most online systems offer their users the option to see different pieces of content (or *objects*) sorted by their posting dates or by some estimate of their popularity. The adopted strategy may have a direct impact on the visibility of different objects. For example, by displaying objects sorted in decreasing order of popularity, a website may

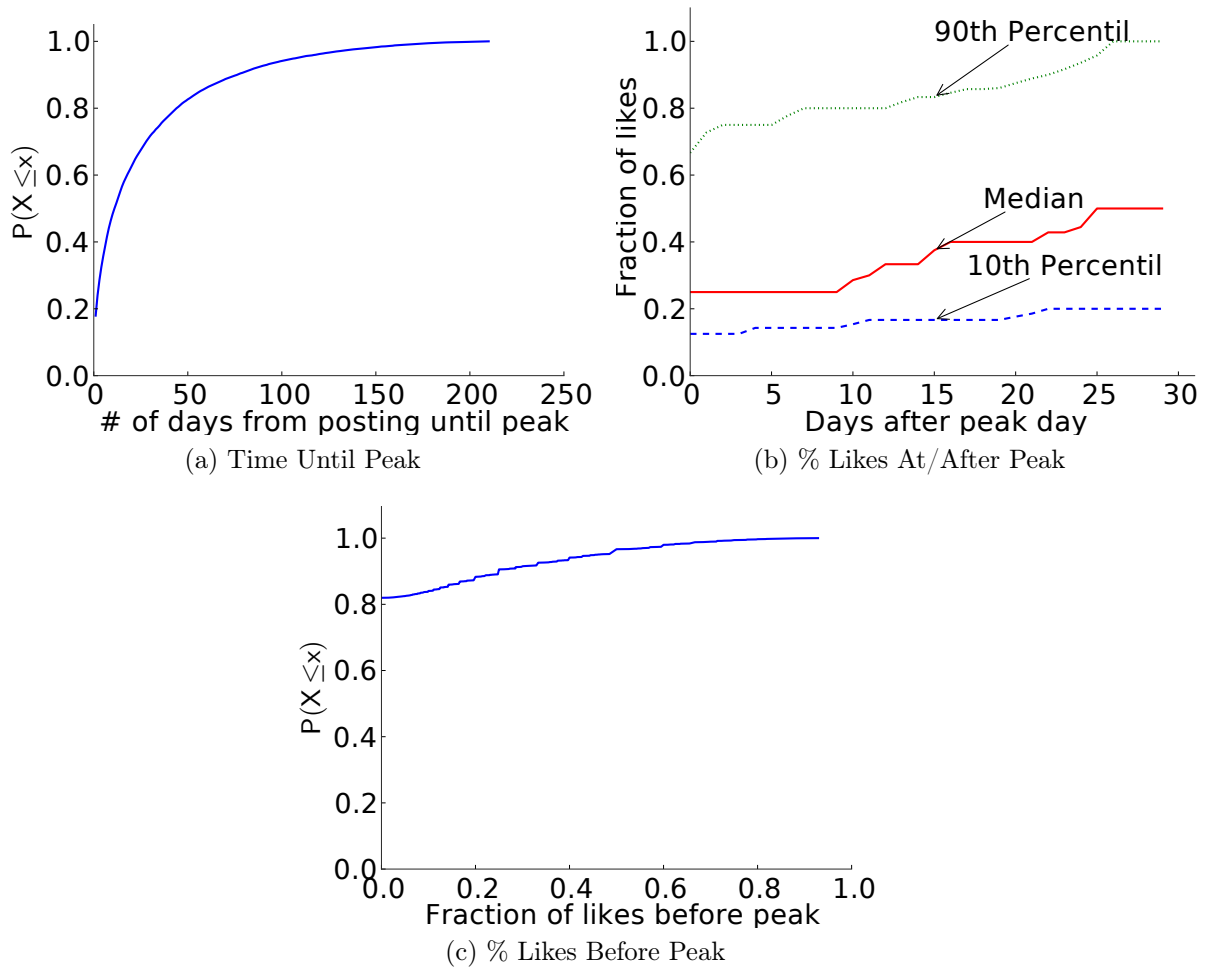


Figure 4.17: Cumulative Distributions of Popularity Peak for Most Popular Tips (G1).

contribute to further increasing the popularity of an object that is already very popular, a phenomenon that is known as *rich-get-richer* [Barabasi and Albert, 1999]. Indeed, prior work has already suggested that popularity of some types of online content (e.g., YouTube videos) evolves according to this phenomenon [Borghol et al., 2012; Szabo and Huberman, 2010].

Foursquare tips may be sorted by the number of likes (in increasing/decreasing order) or by posting time, but only the former is available in the mobile application. Thus, we here assess to which extent the rich-get-richer phenomenon can explain tip popularity evolution.

The rich-get-richer, or preferential attachment, models define that the probability of a tip  $p_i$  experiencing an increase in popularity is directly proportional to  $p_i$ 's current popularity [Barabasi and Albert, 1999]. As in Borghol et al. [2012], we consider a model where the probability that a tip  $p_i$  with  $l_{p_i}$  likes receives a new like is a power

Table 4.10: Rich-get-Richer Analysis: Coefficients  $\alpha$  (and 95% Confidence Intervals) and  $R^2$  of Linear Regressions from (log) Popularity in  $t_r$  to (log) Popularity  $t_r + \delta$ .

$t_r + \delta$	$t_r$	Tips in G1		All Tips	
		$\alpha$	$R^2$	$\alpha$	$R^2$
1 month	1 day	$0.763 \pm 0.016$	0.26	$0.822 \pm 0.006$	0.21
1 month	1 wk	$0.838 \pm 0.009$	0.57	$0.887 \pm 0.004$	0.49
2 months	1 day	$0.594 \pm 0.017$	0.17	$0.673 \pm 0.007$	0.13
2 months	1 wk	$0.681 \pm 0.011$	0.40	$0.753 \pm 0.004$	0.31
2 months	1 mo	$0.834 \pm 0.006$	0.74	$0.856 \pm 0.003$	0.65
6 months	1 day	$0.309 \pm 0.015$	0.07	$0.397 \pm 0.007$	0.05
6 months	1 wk	$0.394 \pm 0.010$	0.20	$0.489 \pm 0.005$	0.16
6 months	1 mo	$0.504 \pm 0.008$	0.40	$0.562 \pm 0.003$	0.33

law, i.e.,  $Prob(p_i) \propto l_{p_i}^\alpha$ .

We analyze the rich-get-richer effect using a univariate linear regression to observe the impact of the number of likes of a tip after a monitoring time  $t_r$  (predictor variable) in the total number of likes of the tip at target time  $t_r + \delta$  (response variable), using log-transformed data. The case of  $\alpha=1$  corresponds to a linear preferential selection [Barabasi and Albert, 1999], and  $\alpha > 1$  implies in a case where the rich gets much richer with time. The sublinear case ( $\alpha < 1$ ) results in a (stretched) exponential popularity distribution, which reflects a much weaker presence of the rich-get-richer effect [Krapivsky et al., 2000]. We perform this analysis separately for tips in each popularity group as well as for all tips.

Table 4.10 shows the coefficients  $\alpha$  (along with corresponding 95% confidence intervals) and the coefficients of determination  $R^2$  of the univariate regressions performed using various predictor and response variables (defined by different values of  $t_r$  and  $\delta$ ), for tips in G1 as well as for all tips. For all considered cases, we find  $\alpha < 1$ , which indicates an exponential popularity evolution that could result in a much less skewed popularity distribution than suggested by the pure (linear) rich-get-richer dynamics. This has also been observed for a set of different YouTube videos [Borghol et al., 2012], although the values of  $\alpha$  found in that case (0.93 on average) are much larger than those we observed in all considered scenarios. This suggests that the rich-get-richer effect might be weaker in Foursquare tips than in YouTube videos, even considering all tips jointly. This also implies that other factors might strongly impact tip popularity. Indeed, as discussed in Section 4.4.2, the social network of the tip’s author is responsible for a significant fraction of the likes received by the tip, and thus might contribute to reduce the impact of the rich-get-richer effect.

The univariate regression model has also been proposed as a means to predict the future popularity of YouTube videos and Digg stories [Szabo and Huberman, 2010].

This prediction strategy was motivated by a strong linear correlation observed between the (log-transformed) popularity of objects and earlier measures of user accesses (also log-transformed). For example, the authors observed Pearson linear correlations above 0.90 between the popularity of Digg stories measured at 1 hour and at 30 days after upload as well as between the popularity of YouTube videos measured at 7 and 30 days after upload. These correlations are stronger than those observed for tips. For example, the  $R^2$  value of the regression from popularity in 1 week to popularity in 1 month is only 0.57 (for tips in G1) and 0.49 (for all tips), which correspond to linear correlations of 0.75 and 0.7, respectively<sup>1</sup>. For shorter monitoring periods  $t_r$  or longer values of  $\delta$ , the  $R^2$  values are much lower, indicating that popularity at time  $t_r$  can only explain a small fraction of the total popularity acquired by the tip at  $t_r + \delta$ .

This result motivates the development of more sophisticated prediction models, such as those proposed in Chapter 6.4, which exploit other factors (e.g., characteristics of the user who posted the tip and the venue where it was posted) to estimate the future popularity of a given tip.

## 4.5 Summary

In this chapter, we analyzed how Foursquare users behave when they interact with each other using tips and likes. This comprehension is important to derive useful insights about the design of prediction models as well as to guide us in interpreting prediction results. Our analyses of some selected features, related to the three main entities that are key to the tip popularity prediction problem (user, venue and tip's content), revealed that most of them exhibit very large variability, with great concentration on few users, venues and tips. In particular, the very skewed distribution of number of likes per tip brings extra challenges to the prediction task, as it leads to great imbalance in the training data. In the next chapter, we try to minimize the detrimental impact of such imbalance on prediction accuracy by employing undersampling in the training data. Furthermore, we found that there are low correlations between the number of tips posted by a user and the number of likes the user gives and/or receives, implying that users who tip more do not necessarily receive or give more feedback on previous tips.

Moreover, in Section 4.2, we identified four user profiles that differ in terms of their levels of tipping activity in the system. Two profiles correspond to regular users who differ in terms of their levels of tipping activity in the system. A third profile consists of

---

<sup>1</sup>The  $R^2$  is the square of the linear correlation between predictor and response variables.

users who seem engaged in posting tips at a large variety of venues. These users, some of which are famous businesses and brands, typically receive a large amount of feedback from others regarding their tips. Finally, we identified a group of users characterized by posting tips containing links at many different venues. A manual inspection of a sample of these users confirmed them as potential spammers, since they posted tips that are unrelated to the venue. However, we also showed that some spammers do succeed in attracting the attention of many users.

We also provided the first pieces of evidence of spamming activity in Foursquare. Spam has been observed in many other online social systems, including Facebook [Gao et al., 2010], YouTube [Benevenuto et al., 2009], and Twitter [Grier et al., 2010]. As a result, a number of efforts towards designing effective strategies to detect and remove spam from these systems are available [Costa et al., 2013; Aggarwal et al., 2013; Thomas et al., 2011]. Although it is debatable whether the kind of spamming activity we uncovered here corresponds to malicious/opportunistic acts that deserves punishment, we hope that our analyses serve as motivation for future discussions on the matter.

In Section 4.3 we modeled user interactions through tips using a graph to identify the most influential users. We proposed a variation of the traditional PageRank algorithm, previously applied to this purpose [Zhang et al., 2007; Weng et al., 2010], which is more adequate to the present context as it weighs the arcs by the number of tips posted by each node (user). Our method performed very similar to PageRank: it was able to identify some expert/influential users not indicated by the traditional method in some of the venue categories. These findings suggest that the category of the venue where a tip is posted must be taken into account in the tip popularity prediction task.

Finally, in Section 4.4, we analyzed the tip popularity dynamics. Although prior work has tackled the popularity dynamics of various types of user generated content, we are not aware of prior temporal analysis of online reviews. We found that most tips have a slow popularity evolution, acquiring most of their likes after a few months, and that the social network of the tip's author plays an important role to draw attention to the tip, particularly soon after posting time. We also found that most tips reach their daily popularity peak within a month in the system, although most of their likes are received after the peak. Moreover, compared to other types of content, we observed a weaker presence of the rich-get-richer phenomenon, indicating a lower correlation between the early and long-term popularities of the tip. This suggests that the tip popularity prediction may require more sophisticated models, exploring other factors related to the tip, besides their current popularity.

In the next chapter, we explore several features analyzed here to design models to predict a tip's future popularity.





## Chapter 5

# Predicting the Popularity Ranking of a Set of Tips

In Chapter 4, we observed that tips have longer lifespans than other types of online content (e.g, tweets, photos), and that tip popularity dynamics may be more strongly influenced by factors other than simply their current popularity (e.g., social network). We now further analyze this issue by assessing to which extent the *relative* popularity of a set of Foursquare tips can be predicted using only their popularity at prediction time, and to which extent the use of other attributes may improve prediction accuracy. The popularity of a tip is estimated as the number of likes received by the tip, which reflects the number of users who agreed with the tip’s author<sup>1</sup>. There are several different prediction tasks depending mainly on the application domains and the user interest.

In this chapter, we address the prediction as a ranking task, which aims at ranking a group of tips based on their predicted popularity at a future time. Thus, we model the prediction task as a *ranking* problem. The ranking of the most popular tips helps to summarize a large set of tips focusing on the most popular ones for a scenario of interest (e.g., a city, a venue), instead of looking at the tips individually. Another way of tackling the popularity prediction problem is to predict the popularity *level* of a single tip, formulating it as a classification task. We leave the discussion of the classification task to Chapter 6. In both chapters, we use dataset 2 to evaluate our proposed solutions.

We first formally define our prediction task (Section 5.1), present the ranking strategies (Section 5.2), and the features used as input to them (Section 5.3). We

---

<sup>1</sup>The number of likes reflects only approximately the level of agreement with the tip’s content, as some users may have chosen not to click on “Like” regardless of their opinions.

then discuss our experimental setup (Section 5.4) followed by our experimental results (Section 5.5).

## 5.1 Popularity Prediction Task

Our goal in this chapter is to develop models to predict the ranking of a set  $P_d$  of tips posted in the previous  $d$  time units ( $d \in (0, \infty)$ ) that meet a certain criterion  $c$ , rank those tips according to their expected popularity, measured in terms of the total number of likes they will receive up to time  $t_r + \delta$ , where  $t_r$  is the time when the ranking (i.e., prediction) is performed. Thus,  $\delta$  defines the prediction window. Criterion  $c$  may be, for example, tips posted at venues of a given city and/or category (e.g., “Food”), or even at a given venue. An empty criterion implies in no further constraint on set  $P$ . Figure 5.1 illustrates the prediction scenarios considered in our study.

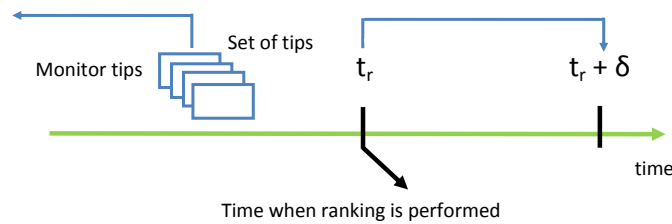


Figure 5.1: Monitoring Time Scheme.

As in other prediction tasks, the prediction model is learned using a training set, which consists of a subset of tips along with associated information about the users who posted them, the venues where they were posted and textual characteristics extracted from their content. The learned prediction model is then evaluated using a different set of tips (test set). Both training and test sets are built considering three different types of entities: a set  $P = \{p_1, \dots, p_K\}$  of  $K$  tips, a set  $U = \{u_1, \dots, u_N\}$  of  $N$  users (tip authors), and a set  $V = \{v_1, \dots, v_O\}$  of  $O$  venues.

Each tip is represented by a tuple  $(p, u, v)$ , and each entity ( $p$ ,  $u$ , and  $v$ ) has a set of attributes (or features)  $F$  associated with it. The features represent the inputs (predictors) associated with a given instance. There are also relationships between these sets of entities: a function  $L : P \rightarrow V$  maps each tip  $p_i$  to a unique venue  $v_i$ , and an authorship function  $A : P \rightarrow U$  maps each tip  $p_i$  to a unique user  $u_i$ .

Thus, given the input data  $\langle p, u, v \rangle$ , we want to learn a prediction model  $M$  that, for a set  $P_d$  of tips posted in the previous  $d$  time units, ranks those tips according to their expected popularity at  $t_r + \delta$ . A tip is represented as an  $f$ -dimensional real vector  $\mathbf{p}$  over a space feature  $F$  built from the information in  $P$ ,  $U$  and  $V$ . The model is

thus a function  $M : \mathbb{R}^f \rightarrow \mathbb{R}$  that maps a tip's feature vector to a numerical popularity (number of likes). Our proposed solution to this problem consists in: (1) determining the set of features used to represent the tips, and (2) applying a learning algorithm to predict the ranking of a set of tips ( $P_d$ ), given  $d$ ,  $t_r$ , and  $\delta$ .

Note that different tips in  $P_d$  may have been posted at different times within the time window  $[t_r - d, t_r]$ . Thus, we associate a posting time  $t_{p_i}$  with each tip  $p_i$  in  $P_d$ . For evaluation purposes, we consider that each tip  $p_i \in P_d$  is *labeled* with a numeric value that represents the number of likes received by  $p_i$  in the time interval  $[t_{p_i}, t_r + \delta]$  (i.e., the true popularity acquired by  $p_i$  up to  $t_r + \delta$ ), as further discussed in Section 5.4. The values of these features for a tip  $p_i$  are computed considering all the information available up to the time when the ranking is performed ( $t_r$ ).

The choice of criterion  $c$  allows for different scenarios where the tip ranking problem becomes relevant. One scenario is that of a user who is interested in quickly finding tips with greater potential of becoming popular, and thus of containing valuable information, posted in any venue in her home city. A different scenario is that of a user who is particularly interested in retrieving tips regarding restaurants in her home city (or neighborhood). A business owner can also benefit from a ranking restricted to tips posted at venues of a specific category to get feedback about her business and about her competitors. Also, changes in the current and future tip popularity rankings can help with indirect analysis such as the influence of certain users whose tips got promoted in the future and the potential market share gains or losses for certain venues or venue categories.

## 5.2 Ranking Strategies

Recall that our goal is to assess to which extent using only the tips' current popularity ranking is enough to accurately predict their ranking at a future time. Thus, we here consider two ranking strategies. The first approach simply uses the ranking of the tips at prediction time ( $t_r$ ) as an estimate of their ranking at the future time  $t_r + \delta$ . If the popularity ranking is stable, this approach should lead to perfect predictions. Thus, by analyzing the effectiveness of this approach we are indirectly assessing the stability of the tip popularity ranking. We use this approach as our *baseline*.

In order to assess the potential benefit of exploiting other factors to this prediction task, we consider a second approach that combines multiple features. To that end, we rely on an ordinary least square (OLS) multivariable regression model to predict the popularity of each tip  $p_i$  in  $P_d$  at time  $t_r + \delta$  and then rank the tips by their predictions.

In this approach, the logarithm of the number of likes of a tip  $p_i$ ,  $\mathcal{R}_t$ , is estimated as a linear function of  $k$  predictor variables or features (presented in the next section), i.e.:

$$\mathcal{R}_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

We note that various other algorithms could be used to exploit multiple features to predict the popularity ranking of a set of tips. Indeed, we also experimented with Support Vector Regression (SVR) [Drucker et al., 1997] with radial basis function kernel as well as with a state-of-the-art learning-to-rank algorithm called Random Forests [Breiman, 2001]. However, when applied with the same set of features, their results are similar (or even worse in some cases) than those obtained with the simpler OLS regression<sup>1</sup>. Thus, in order to avoid hurt readability and to focus our discussion on the benefits of adding the other features, we present only OLS results.

### 5.3 Tip Features

Having presented our prediction models, we now turn to the predictor variables  $x_1$ ,  $x_2$ ,  $\dots$ ,  $x_k$ . These variables are features of the tip  $p_i$  whose popularity we intend to predict.

One of the primary goals of this dissertations is to understand which factors impact the popularity achieved by a tip. This involves defining a set of features and then assessing their relative importance to the tip’s popularity. We explore several features related to the three central entities – tip, user, venue – which, intuitively, should be related to the tip’s popularity. Some of these features have been also explored to discuss the helpfulness of online reviews [Kim et al., 2006; Zhang and Varadarajan, 2006; O’Mahony and Smyth, 2009] and predict the ratings of (long) reviews [Hsu et al., 2009; Lu et al., 2010; Siersdorfer et al., 2010].

As a type of online content, tips also can be evaluated for their credibility as source of information. Fogg et al. [2001] described credibility as a perceived quality composed by multiple dimensions. The authors investigate how seven web site design elements, namely *Real-World Feel*, *Ease of Use*, *Expertise*, *Trustworthiness*, *Tailoring*, *Commercial Implications*, and *Amateurism*, impact its credibility. The study revealed that four of these elements – *Real-World Feel*, *Ease of Use*, *Expertise*, and *Trustworthiness* – have relative impact on increasing credibility. Some of our proposed features are inspired from these elements.

For example, the *Real-World Feel* element refers to aspects indicating that a web site has a physical location and can be contacted. In our specific context, it indicates

---

<sup>1</sup>As we will see in Chapter 6, we also found OLS to be as good as (if not better than) SVR when applied to the (different) task of predicting the popularity level of a given tip (Chapter 6).

to the users that real people wrote the tip and can be reached for questions. *Expertise* is related to how respected the user is in the system, which can be reflected by the number of likes received by that user. The *Ease of Use* element did not fit in any of our features since this element reflects usability aspects, which are standard for all Foursquare users. Finally, *Trustworthiness* was mapped into features that characterize users and venues that were reputable.

For the ranking task, we exploit  $k = 53$  features related to the user  $u_i$  who posted tip  $p_i$ , the venue  $v_i$  where  $p_i$  was posted, and the textual content of  $p_i$ . The values of these features are computed at the time when the ranking is performed ( $tr$ ). Our selected features, which are grouped into user, venue and tip’s content features, are introduced in the next sections.

### 5.3.1 User Features

User features describe the tip’s author past behavior, aiming to identify key behavioral attributes that may impact the tip’s future popularity. In other words, our goal is to identify factors that are useful to access the user credibility.

The full set of user features we consider is the following:

- **Number of tips (`user_tip_num`):** total number of tips posted by the user up to  $tp_i + \varepsilon$ .
- **Number of distinct venues (`user_tipped_venues`):** number of venues where the author posted tips in the past up to time  $tp_i + \varepsilon$ .
- **Number of received likes (`user_total_likes`)<sup>1</sup>:** total number of likes received by tips previously posted by the author in the past up to time  $tp_i + \varepsilon$ . Variations of this feature also included in the feature set are: median (`user_median_likes`), average (`user_avg_likes`) and standard deviation (`user_std_likes`) of the same metric.
- **Number of given likes (`user_given_likes`):** total number of likes given by the tip’s author in the past up to time  $tp_i + \varepsilon$ .
- **Number of friends/followers (`user_sn_size`):** number of Foursquare users who follow or are friends of the tip’s author.

---

<sup>1</sup>Based on the Fogg’s design element *Expertise*.

- **Social Likes (user\_sn\_likes)**: fraction of all likes received by the author coming from his social network in the past up to time  $tp_i + \varepsilon$  (i.e., friends and followers).
- **Tips by the social network (user\_tip\_by\_sn)**: total number of tips posted by the author's social network in the past up to time  $tp_i + \varepsilon$ . Other variations included are: median (**user\_median\_tip\_by\_sn**) and average (**user\_avg\_tip\_by\_sn**) of the same metric.
- **Likes by the social network (user\_like\_by\_sn)**: total number of likes given by the author's social network (in any tip) in the past up to time  $tp_i + \varepsilon$ . Other variations included are: median (**user\_median\_like\_by\_sn**), average (**user\_avg\_like\_by\_sn**) and standard deviation (**user\_std\_like\_by\_sn**) of the same metric.
- **User visibility at venue (user\_venue\_visibility\_total)**: fraction of all likes received by the tip's author that are associated with tips posted at the same venue of the current tip  $p_i$ , but after  $p_i$  was posted. This feature tries to capture an estimation of the user visibility at the venue where the tip was posted. Other variations included are: median (**user\_venue\_visibility\_median**), average (**user\_venue\_visibility\_avg**) and standard deviation (**user\_venue\_visibility\_std**) of the metric.
- **User type (user\_type)**<sup>1</sup>: user category defined by Foursquare.
- **Number of mayorships (user\_mayorships\_num)**: total number of mayorships won by the author.
- **Mayor (user\_mayor)**<sup>2</sup>: binary feature indicating whether the tip's author was a mayor of the venue where the tip was posted.

### 5.3.2 Venue Features

The second set of features considered are related to the venue where the tip was posted. The selected features capture the activity at the venue or its visibility by other users. For example, a tip may have a higher chance of becoming popular if it is posted at a venue that has more visibility. Another piece of information that we try to capture is

---

<sup>1</sup>Celebrity and brand accounts may represent respectable people or organizations. This item was also listed by Fogg in his study [Fogg et al., 2001]

<sup>2</sup>Based on the Fogg's design element *Expertise*.

the strategy adopted by Foursquare to display the tips posted at the same venue, which also may impact the visibility of a tip. Foursquare tips may be sorted by their number of likes or by posting time, but only the former is available in the mobile application. Thus, because of this placement, the top tips may accumulate more and more likes, while recent reviews may be rarely read and thus not rated, a manifestation of the rich-get-richer effect. Indeed, as we discussed in Chapter 4, the rich-get-richer effect on Foursquare is present, but it is weaker compared to other types of content (e.g., video and photos).

The following venue features are considered:

- **Number of tips (`venue_tip_num`)**<sup>1</sup>: total number of tips posted at the venue up to  $tp_i + \varepsilon$ .
- **Number of received likes (`venue_total_likes`)**<sup>2</sup>: total number of likes received by all tips previously posted at the venue until  $tp_i + \varepsilon$ . Other variations of this metric included are: median (`venue_median_likes`), average (`venue_avg_likes`) and standard deviation (`venue_std_likes`).
- **Number of check ins (`venue_cks_num`)**<sup>2</sup>: total number of checkins at the venue.
- **Number of visitors (`venue_visitors_num`)**<sup>2</sup>: total number of unique visitors.
- **Verified venue (`venue_verified`)**<sup>3</sup>: binary feature indicating whether the tipped venue was verified by Foursquare.
- **Category (`venue_category`)**: venue category defined by Foursquare.
- **Position in the ranking by likes**: position of the tip in the ranking of tips of the venue sorted by number of received likes in both ascending (`venue_like_rk_pos_asc`) and descending (`venue_like_rk_pos_dsc`) orders.
- **Position in the ranking by date**: position of the tip in the ranking of tips of the venue sorted by posting date in both ascending (`venue_date_rk_pos_asc`) and descending (`venue_date_rk_pos_dsc`) orders.

---

<sup>1</sup>Based on the Fogg's design element *Trustworthiness*. We can interpret the number of tips, the number of received likes, the number of check ins or unique visitors as estimates of the venue popularity. Popularity can reflect recommendation from other users.

<sup>2</sup>Based on the Fogg's design element *Trustworthiness*.

<sup>3</sup>Based on the Fogg's design element *Real-world Feel*. A venue is verified, when it was claimed by a real business owner. Thus, the venue exists outside the Internet, in the real world.

### 5.3.3 Tip’s Content Features

Finally, we consider various features related to the textual content of the tip. Some of these features are the number of characters (**tip\_char\_num**), number of words (**tip\_wrd\_num**), and number of URLs or e-mail addresses (**tip\_url\_num**) in the content of the tip, which were analyzed in Section 4.1.3. Note that, by considering **tip\_char\_num** and **tip\_wrd\_num**, we aim at analyzing to what extent the size of the tip may impact its popularity. However, we also consider various other features that capture linguistic and semantic characteristics of the textual content, as discussed next.

A number of studies [Liu et al., 2008; Lu et al., 2010; O’Mahony and Smyth, 2010; Wagner et al., 2012] have shown that the linguistic style can be a good indicator of the utility/helpfulness of the review or quality of other user generated contents [Agichtein et al., 2008; Chen et al., 2011; Dalip et al., 2011; Momeni et al., 2013]. As in [Liu et al., 2008; Lu et al., 2010; Momeni et al., 2013], we have chosen to model syntactic features using the Part-Of-Speech (POS) tagging of the words in the tip’s text. In this step, each word is assigned a label, which represents its position/role in the grammatical context (e.g., noun, adjective, verbs, etc.). For our experiments, we used the Stanford Part of Speech tagger [Stanford NLP Group, 2012], which uses probabilistic methods to build parse trees for sentences aiming at representing their grammatical structure. Thus, for each tip, we parse it and count the number of words with each tag. We then divide each number by the total number of words in the tip, for normalization purposes. The features created based on the POS tags are shown in Table 5.1.

Table 5.1: Tip’s Syntactic Content Features.

<i>Feature Name</i>	<i>Feature Description</i>
tip_pos_nn	ratio of nouns
tip_pos_adj	ratio of adjectives
tip_pos_adv	ratio of adverbs
tip_pos_comp	ratio of comparatives
tip_pos_ver	ratio of verbs
tip_pos_fw	ratio of foreign words
tip_pos_num	ratio of numbers
tip_pos_sup	ratio of superlatives
tip_pos_sym	ratio of symbols
tip_pos_pp	ratio of punctuation symbols

Furthermore, we also included three scores – positive, negative and neutral – that captures the tip’s sentiment. These scores are computed using SentiWordNet



[Esuli and Sebastiani, 2006], an established sentiment lexical for supporting opinion mining in English texts. SentiWordNet is a lexical resource built on top of WordNet [Fellbaum, 1998], which is a lexical database of English that groups nouns, verbs, adjectives and adverbs into sets of synonyms, each expressing a distinct concept. In SentiWordNet, each term is associated with a numerical score in the  $[0, 1]$  range for positive, negative and objectivity (neutral) sentiment information. We compute the scores (**tip\_pos\_score**, **tip\_neg\_score** and **tip\_neu\_score**) of a tip by taking the averages of the corresponding scores over all words in the tip that appear in SentiWordNet. We adopted SentiWordNet because it showed to be the best and the cheapest method among several state-of-the-art supervised strategies to detect the polarity of tips [Moraes et al., 2013b]. To handle negation, we adapted a technique proposed in Pang et al. [2002], that reverse the polarity of words between a negation word (“no”, “didn’t”, etc.) and the next punctuation mark<sup>1</sup>.

Since different tips in the set  $P_d$  may have been posted at different times, we also add the age of the tip (in number of hours since posting time  $tp_i$ ) and the number of likes it has already received (**tip\_age\_hours** and **tip\_likes\_current**, respectively). All features are summarized in Table 5.2.

## 5.4 Experimental Setup

We build two scenarios to evaluate the prediction strategies: ranking all tips recently posted at venues located in New York, the city for which we have the largest number of tips, and ranking tips posted at venues of a specific category (Food) (also the largest category) located in New York<sup>2</sup>. In both scenarios, we consider only tips posted in the previous month (i.e.,  $d = 30$  days), and produce rankings based on their predicted popularity  $\delta$  days later. We compare the effectiveness of both prediction strategies for various values of  $\delta$ . Table 5.3 summarizes these two datasets, presenting the total numbers of tips, venues and users in each of them (the two rightmost columns are discussed below).

As we did for the classification task, we split the tips chronologically into training and test sets. Figure 5.2 illustrates this chronological splitting used. For comparison purposes, we also evaluate the baseline only in the test sets.

<sup>1</sup>Another option is to use the negation detection tool of the Stanford POS tagger.

<sup>2</sup>Other scenarios, such as ranking tips posted at a venue, are also possible. However, the highly skewed distribution of tips per venue leads to severe data sparsity, which, in turn, poses a challenge to the training of the regression model.

Table 5.2: Complete Set of Features for Tip Popularity Prediction

Type	Feature Name	Feature Description
User	user_tip_num	Total number of tips
	user_tipped_venues	Number of distinct venues
	user_total_likes	Total number of received likes <sup>1</sup>
	user_given_likes	Total number of given
	user_sn_size	Number of friends/followers
	user_sn_likes	Social likes
	user_tip_by_sn	Total number of tips by the social network <sup>1</sup>
	user_like_by_sn	Total number of likes given by social network <sup>1</sup>
	user_venue_visibility_total	User visibility at venue <sup>1</sup>
	user_type	Foursquare user category
	user_mayorships_num	Total number of mayorships
	user_mayor	If the author was mayor of the venue
Venue	venue_tip_num	Total number of tips
	venue_total_likes	Total number of received likes <sup>1</sup>
	venue_cks_num	Total number of check ins
	venue_visitors_num	Total number of visitors
	venue_verified	If the venue was verified
	venue_category	Foursquare venue category
	venue_like_rk_pos_asc	Like ranking position in ascending order
	venue_like_rk_pos_dsc	Like ranking position in descending order
venue_date_rk_pos_asc	Date ranking position in ascending order	
Content	tip_likes_current	Total number of likes received until time $t_r$
	tip_age_hours	Hours since posting time until time $t_r$
	tip_char_num	Total number of characters
	tip_wrd_num	Total number of words
	tip_url_num	Total number of URLs or email addresses
	tip_pos_nn	Fraction of nouns
	tip_pos_adj	Fraction of adjectives
	tip_pos_adv	Fraction of adverbs
	tip_pos_comp	Fraction of comparatives
	tip_pos_ver	Fraction of verbs
	tip_pos_fw	Fraction of non-English words
	tip_pos_num	Fraction of numbers
	tip_pos_sup	Fraction of superlatives
	tip_pos_sym	Fraction of symbols
	tip_pos_pp	Fraction of punctuation
	tip_pos_score	Positive tip score
	tip_neg_score	Negative tip score
	tip_neu_score	Neutral tip score

<sup>1</sup> Median, average and standard deviation are also included.

Table 5.3: Overview of Datasets and Scenarios of Evaluation

Scenarios	# of tips	# of users	# of venues	# of tips in training sets	Avg # of tips in test sets
NY	169,393	55,149	31,737	516	4,697.87
NY Food	81,742	32,961	8,927	244	2,365.0

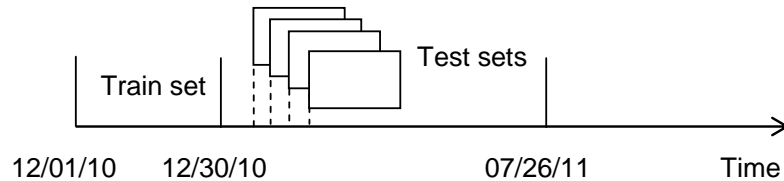


Figure 5.2: Temporal Data Split into Train and Test Sets.

The training set is composed of all tips posted from December 1<sup>st</sup> to 30<sup>th</sup>, 2010. These tips are used to learn the (regression-based) ranking model. We assume the ranking of the training instances is done on December 30<sup>th</sup>, and thus use the total number of likes received by these tips at the target date (i.e.,  $\delta$  days later) as the ground truth to build the regression model.

Recall that the distribution of number of likes per tip is highly skewed towards very few number of likes (Chapter 4), which might bias the regression model and ultimately hurt its accuracy<sup>1</sup>. Thus, we adopt the following approach to reduce this skew<sup>2</sup>. We group tips in the *training set* according to a threshold  $\omega$  for the number of likes received by the tip at the target date. Two classes are defined: all tips with at least  $\omega$  likes are grouped into the high popularity class and the others are grouped into the low popularity class. We then build balanced training sets according to the two popularity classes by performing under-sampling<sup>3</sup>. We repeat this process 5 times, thus building multiple (balanced) training sets and allowing us to assess the variability of our results. We note that this under-sampling approach (and threshold  $\omega$ ) is applied *only to the training set*. The test sets (described next) remains unchanged (imbalanced). Table 5.3 (5<sup>th</sup> column) presents the total number of tips in the training sets for each scenario.

We then use tips posted from December 31<sup>st</sup> until February 27<sup>th</sup> 2011 to build 30 different test sets, as follows. Since tips can be continually liked, the predicted ranking may become stale. Thus, we evaluate the effectiveness of the ranking methods by using them to build a new ranking by the end of each day (starting on January 29<sup>th</sup>), always

<sup>1</sup>Great imbalance in the training set, as observed in our datasets, is known to have a detrimental impact on the effectiveness of classification and regression algorithms.

<sup>2</sup>We also adopt the same under-sampling strategy used for the classification task.

<sup>3</sup>For illustration purposes, we note that the original training set for the NY scenario had 5,225 tips in the low popularity class and only 258 tips in the other (smaller) class.

considering the tips posted in the previous  $d = 30$  days. Thus, 30 test sets are built by considering a window of 30 days and sliding it 1 day at a time, 30 times. Table 5.3 (6<sup>th</sup> column) shows the average number of tips in each test set. For each test set, we report average results produced by all 5 training sets, and corresponding 95% confidence intervals.

For both training and test sets, the features of each tip are computed using all data collected up to the time when ranking is performed ( $t_r$ ), including (for the regression model) information associated with tips posted before the beginning of each training set. Moreover, feature values are computed by first applying a logarithm transformation on the raw numbers to reduce their large variability, and then scaling these results between 0 and 1. We note that, in order to have enough historical data about users who posted tips, for both training and test sets, we consider only tips posted by users with at least five tips. We determine the best parameters of the regression models by minimizing the least squared errors of predictions for the candidate tips in the training set.

We evaluate each ranking method by computing the Kendall  $\tau$  rank distance of the top- $k$  tips in the rankings produced by the method and the top- $k$  tips in the ideal ranking defined by the number of likes accumulated by each tip until  $t_r + \delta$  (i.e., the tip's label). We refer to it as  $K\tau@k$ . Since we are comparing two top- $k$  lists ( $\tau_1$  and  $\tau_2$ ), we use a modified Kendall  $\tau$  metric [Konagurthu and Collier, 2013], that uses a penalty parameter  $p$ , with  $0 \leq p \leq 1$ , to account for the distances between non-overlapping tips in  $\tau_1$  and  $\tau_2$ <sup>1</sup>. The modified Kendall  $\tau$  is defined as follows:

$$K\tau(\tau_1, \tau_2)@k = (k - |\tau_1 \cap \tau_2|)((2+p)k - p|\tau_1 \cap \tau_2| + 1 - p) + \sum_{i \in \tau_1 \cap \tau_2} \kappa_{i,j}(\tau_1, \tau_1) - \sum_{i \in \tau_1 - \tau_2} \tau_1(i) - \sum_{i \in \tau_2 - \tau_1} \tau_2(i) \quad (5.1)$$

where  $\tau_1(i)$  (or  $\tau_2(i)$ ) is the position in the ranking of the item  $i$  and  $\kappa_{i,j}(\tau_1, \tau_2) = 0$  if  $\tau_1(i) < \tau_1(j)$  and  $\tau_2(i) < \tau_2(j)$ , or  $\kappa_{i,j}(\tau_1, \tau_2) = 1$ , otherwise.  $K\tau@k$  ranges from 0 to 1, with values close to 1 indicating greater disagreement between the predicted ranking and an ideal ranking.

---

<sup>1</sup>We use  $p = 0.5$  which was recommended by Fagin et al. [2003].

## 5.5 Experimental Results

We discuss our results by first assessing how the popularity ranking of tips varies over time (Section 5.5.1), and then comparing the prediction based only on the current ranking (baseline), the regression-based prediction that uses a richer set of features (Section 5.5.2). Finally, in Section 5.5.3, we analyze the impact in the model accuracy of removing features one at time according to Information Gain metric.

### 5.5.1 Ranking Stability

Using the experimental setup described in Section 5.4, we investigate the differences between the *true* popularity rankings of tips at times  $t_r$  and  $t_r + \delta$ , for various values of  $\delta$ . To that end, we quantify the correlation between these two rankings using Kendall’s  $\tau$  coefficient. Recall that the closer to 1 the value of  $K\tau$  is, the larger the disagreements between both rankings.

Figure 5.3 shows the  $K\tau@k$  for each day in the test fold of both NY and NY Food scenarios, for values of  $\delta$  varying from 1 to 5 months. We focus on the top-10 most popular tips ( $k=10$ ). Focusing first on the NY scenario, Figure 5.3a shows that the disagreements between both rankings increase as we increase  $\delta$ . Indeed, for a fixed test day (fixed set of tips), the  $K\tau@10$  varies from 0.26 to 0.72 as we increase  $\delta$  from 1 to 5 months. Moreover, we can still observe some discrepancies even if we predict for only one month ahead in the future ( $\delta=1$  month). Indeed, as discussed in Section 4.4, over 40% of the likes of most tips arrive after two months since posting time. Since the tips in each test fold are at most 1 month old, most of them are still at very early stages of their popularity curves, and the popularity ranking, even considering only the top-10 tips, will change. Very similar results were also observed for the NY Food scenario, as shown in Figure 5.3b, although the values of  $K\tau@10$  (and thus the disagreements between current and future rankings) seem somewhat smaller on some days, particularly for larger values of  $\delta$ .

Examining the top most popular tips in each test fold for the NY scenario, we found that some of them referred to special events occurring in the city. Examples are a venue created to be a “promoted” venue during the *2010 Super Bowl* game event and a “meme” venue, *Snowpocalypse 2010*, created to celebrate the major New York City snow storm [Seward, 2011; Barbierri, 2011]. These tips exhibit a somewhat different pattern: all of their likes are received until the event occurs. Thus, once they reach the top of the ranking, they tend to remain there for a while, which contributes to lower the discrepancies between predicted and future rankings.

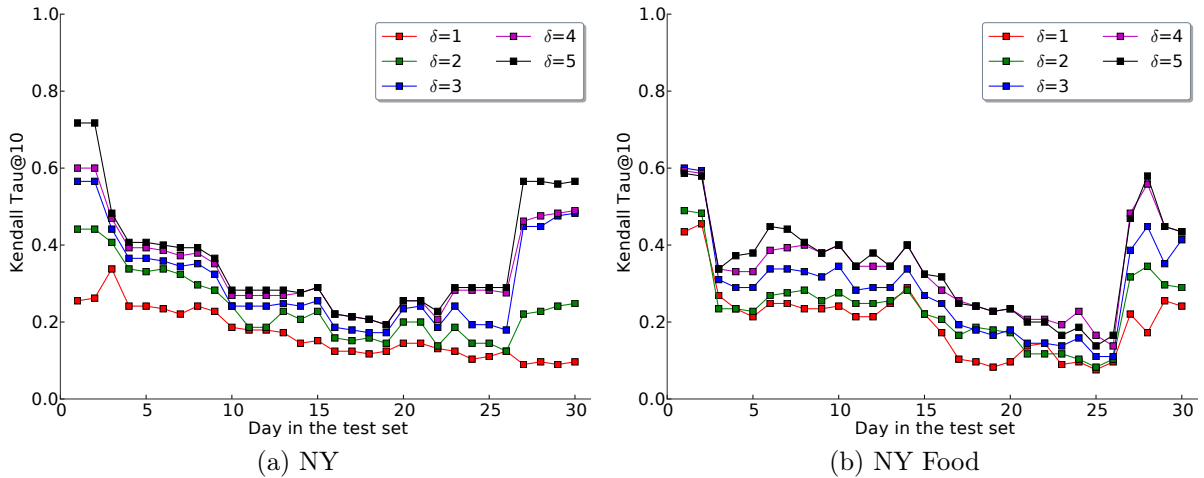


Figure 5.3: Correlations between the top-10 most popular tips at time  $t_r$  and at time  $t_r + \delta$  ( $\delta$  in months).

Overall, these results corroborate our discussion in Section 4.4.4, and suggest that there are some noticeable discrepancies between the current and the long-term popularity of tips (even within the top-10 most popular tips). Thus, models that use only early measurements, such as those proposed by Szabo and Huberman [2010] as well as Pinto et al. [2013], may lead to inaccurate predictions not only of popularity measures (as will be discussed in Chapter 6) but also of popularity ranking. Next, we assess to which extent such ranking predictions can be improved by exploiting a multidimensional set of predictors.

## 5.5.2 Prediction Results

We now compare the prediction results using only the popularity ranking at time  $t_r$  (baseline) against the prediction produced by using the OLS regression model jointly with the features defined in Section 5.3. Figure 5.4 shows the average daily  $K\tau@10$  along with 95% confidence intervals for the two ranking methods and each value of  $\delta$ , for the NY scenario. For  $\delta$  equal to 1 month, both methods produce  $\tau@k$  results below 0.4, showing a high correlation between the predicted ranking and the true popularity ranking at  $t_r + \delta$ . However, the OLS regression model produces results that are significantly better (lower  $K\tau@10$ ) than those produced by the baseline in 67% of the days (reductions in up to 69%).

Moreover, as we predict further into the future, increasing  $\delta$  to 2, 5 and 6 months, we observe increasing values of  $K\tau@10$  for both methods. This implies that the discrepancies with the true ranking tend to increase as both methods start using outdated

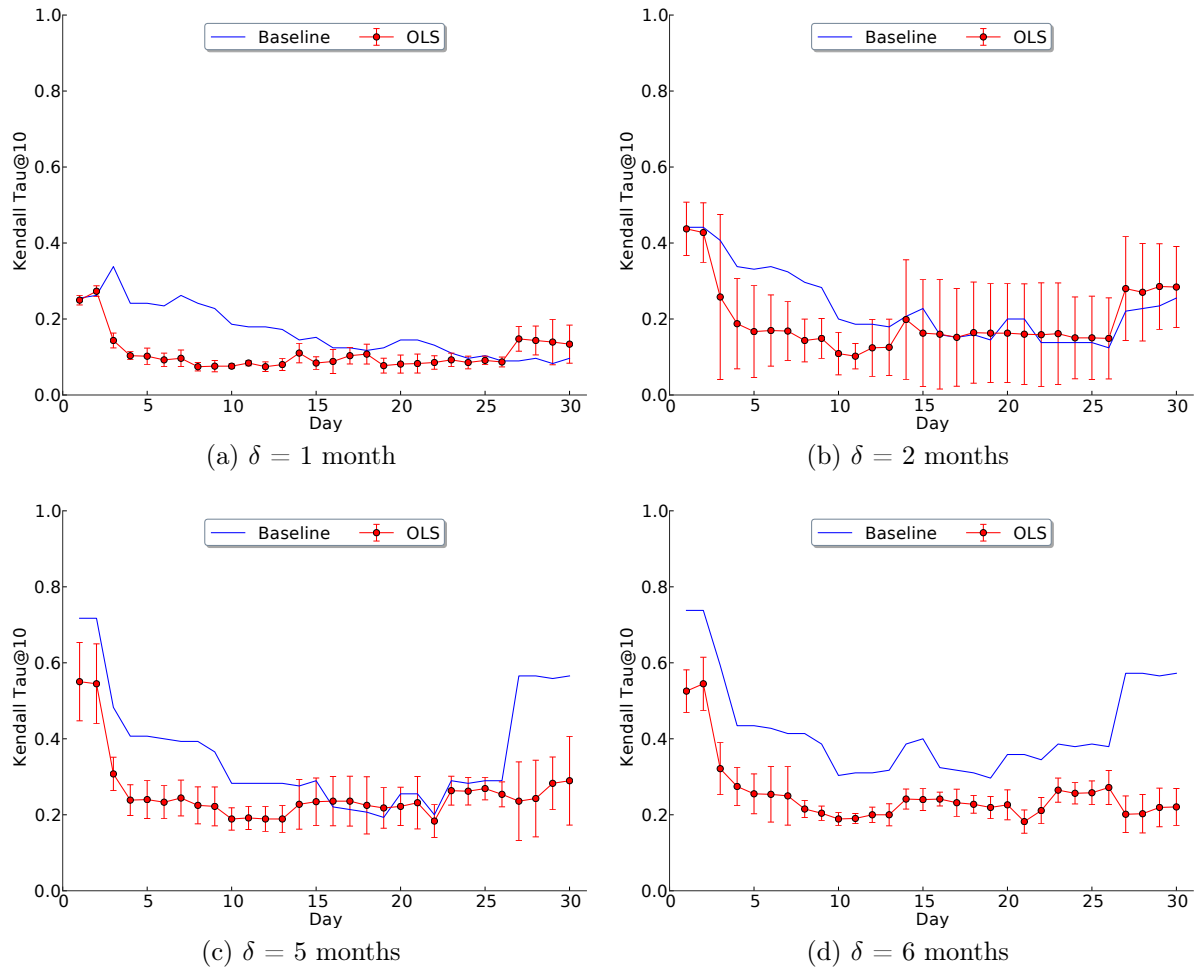


Figure 5.4: Effectiveness of Ranking for Varying Target Time  $t_r + \delta$ : NY Scenario (Avg and 95% Confidence Intervals).

and possibly inaccurate data. Yet, the gap between the baseline and the OLS regression model tends to increase (reaching up to 65% for  $\delta$  equal to 6 months). This result shows that taking factors other than simply the current popularity of the tips into account is important and can improve prediction accuracy of the long-term popularity ranking.

We note, however, that there are some cases where the baseline performs as good as the more sophisticated OLS model. These specific cases are explained by the following: some of the most popular tips (which referred to real events), acquired most of their likes very early on before time  $t_r$  (before the event). Figure 5.5 shows a top-5 ranking of tips taken from our experimental datasets. We observed that the first two tips in the ranking  $T_1$  and  $T_2$  achieved 95% and 90% of their likes at time  $t_r$ ). These two tips referred to the *Snowpocalypse 2010* event. Thus, they quickly reached top

Figure 5.5: Tips Ranking Example.

Tip	Tip age (days)	# likes at $t_r$	# likes at $t_r + \delta$
$T_1$	4	128	135
$T_2$	4	85	94
$T_3$	28	16	57
$T_4$	23	25	54
$T_5$	28	5	47

positions of the ranking, remaining there until  $t_r + \delta$ . For such cases, the use of other features produces only marginal improvements in prediction.

Figure 5.6 shows similar results for the NY Food scenario. In this case, we see smaller differences between both methods. In most cases, the baseline is just as good as the more sophisticated OLS method, although the use of the extra features does provide improvements (up to 30%) in some of the days for large values of  $\delta$ . These results reflect the higher stability of the tip popularity ranking in the NY Food scenario, discussed in Section 5.5.1. Moreover, as shown in Table 5.3, the number of tips in the training set of this scenario is almost half of that used in the NY scenario, which also impacts the accuracy of the regression model. That is, the benefits from using more features are constrained by the limited amount of data to train an accurate model<sup>1</sup>.

These results highlight that the accurate prediction of the tips popularity ranking of a set of tips is a challenging task. Although tip popularity ranking remains roughly stable over short periods of time (e.g., 1 month), there are still significant discrepancies that occur in the top of the ranking. Moreover, the use of other features related to the tip’s author, venue and tip’s content can improve prediction accuracy to some extent, provided that enough information about the features is available to train the model.

### 5.5.3 Experiments Removing Features

Focusing on the OLS prediction strategy, we now assess the relative importance of each feature using the well-known Information Gain feature selection technique [Yang and Pedersen, 1997].

Information Gain, originally used to compute splitting criteria for decision trees, is often used as a measure of how well a given feature separates the given dataset

<sup>1</sup>Recall that we did experiment with other prediction strategies based on SVR and Random Forests, but OLS provided the best results across all scenarios.



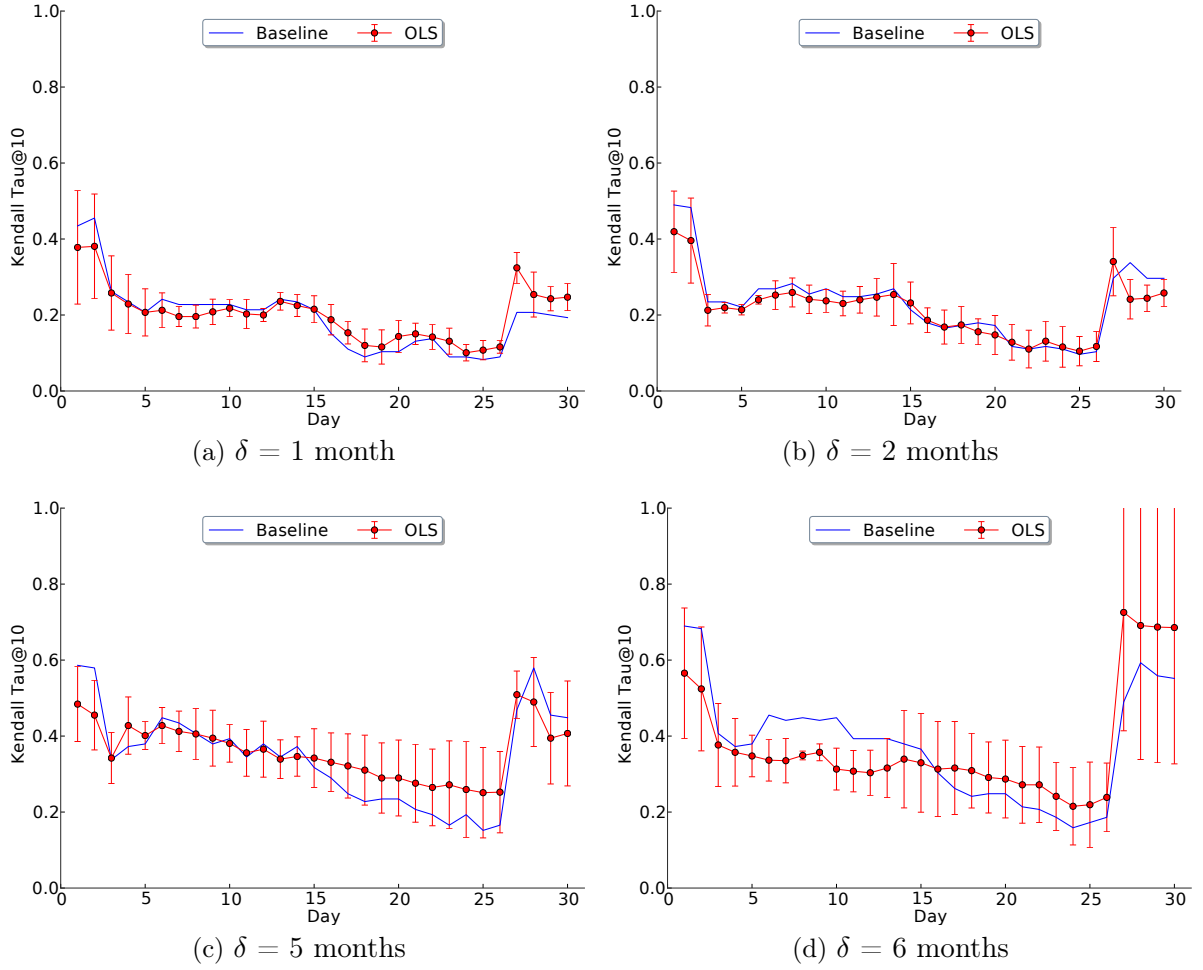


Figure 5.6: Effectiveness of Ranking for Varying Target Time  $t_r + \delta$ : NY Food Scenario (Avg and 95% Confidence Intervals)

[Janecek et al., 2008]. Before computing the Information Gain, we must compute the overall entropy  $I$  for the dataset  $S$ , defined as:

$$I(S) = - \sum_{i=1}^C p_i \log_2 p_i \quad (5.2)$$

where  $C$  is the number of classes and  $p_i$  is the fraction of instances from class  $i$  in  $S$ . It takes the maximum value when the instances are equally distributed amongst the  $C$  classes. Information Gain is the expected reduction in entropy caused by partitioning the instances according to a given feature. Thus, the Information Gain  $IG(S, F)$  of a given feature  $F$  is defined as:

$$IG(S, F) = I(S) - \sum_{v \in \text{values}(F)} \frac{|S_{F,v}|}{|S|} I(S_{F,v}) \quad (5.3)$$

where  $v$  is the set of all possible values for feature  $F$ , and  $S_{F,v}$  is the set of instances where  $F$  has value  $v$ . Note that the first term in the equation is just the entropy of the whole dataset  $S$ , and the second term is the expected value of the entropy after  $S$  is partitioned using feature  $F$ .

Table 5.4: Features Ranked by Information Gain

Pos.	Feature	Pos.	Feature
1	tip_likes_current	28	venue_std_likes
2	user_total_likes	29	tip_pos_pp
3	user_avg_likes	30	tip_pos_nn
4	user_std_likes	31	venue_avg_likes
5	user_median_likes	32	tip_pos_adj
6	user_tipped_venues	33	tip_pos_adv
7	user_tip_num	34	venue_tip_num
8	user_avg_tip_by_sn	35	tip_neg_score
9	user_venue_visibility_total	36	venue_date_rk_pos_asc
10	user_sn_size	37	venue_total_likes
11	user_tip_by_sn	38	tip_pos_ver
12	user_like_by_sn	39	user_venue_visibility_median
13	user_type	40	tip_neu_score
14	user_std_like_by_sn	41	tip_pos_score
15	user_avg_like_by_sn	42	tip_pos_num
16	user_sn_likes	43	user_median_tip_by_sn
17	user_venue_visibility_std	44	tip_pos_fw
18	user_venue_visibility_avg	45	tip_pos_sup
19	user_given_likes	46	venue_category
20	user_mayorships_num	47	venue_median_likes
21	tip_char_num	48	tip_pos_comp
22	tip_wrd_num	49	user_median_like_by_sn
23	tip_age_hours	50	venue_verified
24	venue_cks_num	51	user_mayor
25	venue_like_rk_pos_asc	52	tip_url_num
26	venue_like_rk_pos_dsc	53	tip_pos_sym
27	venue_visitors_num		

Table 5.4 shows features ranked by Information Gain. We found that the most important feature is, unsurprisingly, the current popularity of the tip. It is followed by features related to the user’s popularity, such as the total number of likes in previous tips (*user\_total\_likes*). Features related to the social network of the tip’s author (average number of tips posted by his social network, and the number of followers and friends) are also in the top-10 most important features (8th and 10th positions, respectively).

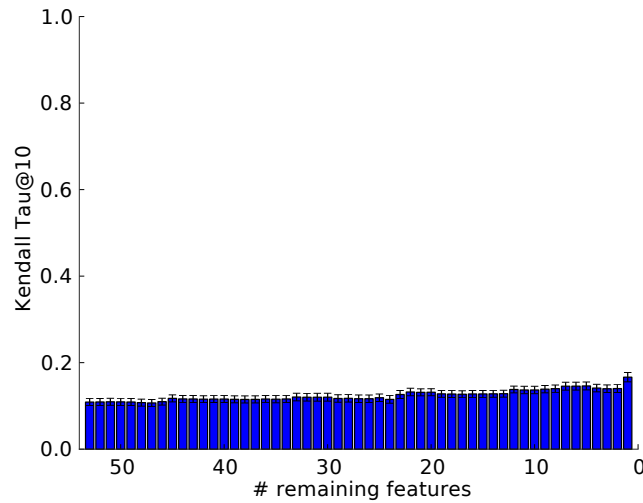


Figure 5.7: Effectiveness of Ranking when Removing One Feature at a Time: NY Scenario (Avg and 95% Confidence Intervals for All Considered Days)

The most important venue feature is the total number of check-ins, followed by the current position of the tip in the ranking of tips of the venue sorted by increasing number of likes. However, these features, like the other venue related features, are much less important than the user features, occupying only the 24<sup>th</sup> and 25<sup>th</sup> positions of the ranking. Similarly, the most important content feature is the number of characters in the tip, but it occupies only the 21<sup>st</sup> position of the ranking. Thus, unlike in other efforts to assess the helpfulness of online reviews [Kim et al., 2006; Zhang and Varadarajan, 2006], textual features play a much less important role in the tip popularity ranking prediction task, possibly due to the inherent different nature of these pieces of content.

Finally, we extend our evaluation of the relative importance of each feature by assessing the accuracy of the OLS strategy as we remove one feature at a time, in increasing order of importance given by the Information Gain. Figure 5.7 shows the impact on the average  $K\tau@10$  as each feature is removed, starting with the complete set of user, venue and content features. For example, the second bar shows results after removing the least discriminative feature (fraction of symbols in the tip text). Note that the removal of many of the least discriminative features has no significant impact on  $K\tau@10$ , indicating that these features are redundant.

However, we observe statistical significant losses (up to 18.8%) when we remove the features related to the number of check ins at the venue (24<sup>th</sup>), the Foursquare user type (13<sup>th</sup>) and the user’s total number of likes (2<sup>nd</sup>). After each one of these losses, we also observe periods of stability (no losses) between the removal of some features. For example, the period between the removal of the 24<sup>th</sup> and 13<sup>th</sup> features. We thus

built a model that focus only on these four features (number of check ins at the venue, user type, user’s total number of likes and current popularity of the tip) in which we observed the largest losses. Figure 5.8 shows the same curves of Figure 5.4a, including the new OLS model with only four features. This result shows that the model with four features also produces results that are also statistically better than those produced by the baseline (in 76.7% of the days) and, on average, is statistically tied with the model with all features. In sum, using only four features, we are able to produce predictions that are as accurate as those using the complete set of features.

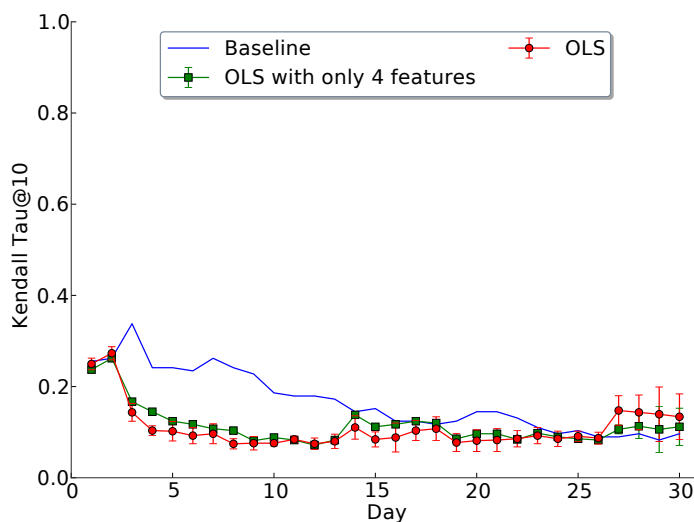


Figure 5.8: Effectiveness of Ranking When Using Only 4 Features for  $\delta = 1$  month (Avg and 95% Confidence Intervals)

## 5.6 Summary

In this chapter, we addressed the popularity prediction problem as a ranking task, which consists in estimating the ranking by popularity of a given set of tips at a future time. We evaluated the stability of tip popularity ranking over time, observing that there are some noticeable disagreements between the current and future popularity rankings, even when considering only the top-10 most popular tips and a time window of only 1 month. This suggests that predicting the future ranking based only on the current ranking may not be accurate.

We thus investigated to which extent we can improve such predictions by using a regression model and exploiting a multidimensional set of features related to the tip’s author, the venue where it was posted and its content. Our results showed that the

use of these features can improve to some extent the prediction accuracy, given that enough training data is available.

Finally, we assessed the relative importance of each features of the regression model using the Information Gain feature selection technique. We observed a model that used only four features (number of check ins at the venue, user type, user's total number of likes and current popularity of the tip) is as accurate as those using a complete set of features.



## Chapter 6

# Predicting the Popularity Level of a Tip

In the previous chapter, we tackled the popularity prediction problem as a ranking task, which aims at ranking a group of tips based on their predicted popularity at future time. In this chapter, we investigate a different prediction task, modeling it as a *classification task*. Using some of the features defined in the previous chapter, such task aims to predict the popularity *level* of a single tip. Ranking as classification tasks support different applications. For example, tip ranking supports filtering and recommendation at finer granularity (as opposed to the number of popularity levels) which is useful to users and venue owners. The classification task focus on the popularity of a given tip that was just posted. So, users and venue owners who would like to be able to predict if a tip will have enough visibility ahead of time can react quickly to promote it if it is the case.

In this chapter, we also focus on the hardest prediction scenario, i.e., prediction at posting time, when the only information available about the tip consists of its content and historical patterns related to the user and the venue associated with it. However, many prior efforts to predict the popularity of online content exploit information about how the user population reacts to the content during an initial monitoring time (e.g., its early popularity measures). Another challenge is that we estimate the popularity of a single tip, instead of a relative popularity outputted by the ranking task, which is also a harder task.

As in previous related efforts, we here tackle the problem of predicting the popularity level of a tip will reach using both regression [Chen et al., 2011; Hsu et al., 2009; Kim et al., 2006] and classification [Liu et al., 2008, 2007] techniques. Specifically, like in Liu et al. [2008], when applying regression, we use the predicted value as a threshold

to classify a tip into different popularity levels, defined based on ranges of number of likes received. We do so by determining into which range the predicted number of likes fall.

We start this Chapter discussing the popularity levels we consider in Section 6.1. Next, we formally define the problem of predicting the popularity *level* of a tip on Foursquare (Section 6.2). We then discuss the techniques adopted to design our prediction models (Section 6.3) and the features used as input to them (Section 6.4). We present our evaluation methodology (Section 6.5) followed by our main experimental results (Section 6.6). We then investigate the accuracy of our models when varying monitoring time and prediction target time. Finally, in Section 6.8, we discuss the impact of model specialization.

## 6.1 Popularity Levels

In this chapter, we investigate models to predict the level of popularity of a tip at a certain point in time in the future. For a given tip, its popularity is defined as the expected number of times the tip will be marked as liked. Moreover, as in previous work [Hong et al., 2011; Bandari et al., 2012; Anderson et al., 2012], instead of predicting the exact number of likes a tip will receive, we categorize tips into various level of popularity, defined by ranges of the number of likes received, and predict this level instead. We choose to predict the level of popularity instead of the exact number of likes of the tip because the latter is harder, particularly given the very skewed distribution of number of likes per tip (shown in Figure 4.6a). Moreover the former should be good enough for various purposes (e.g., early identification and highlight of future popular tips as soon as they are posted, so they can be more visible on the site or definition of revenue schemes for ads embedded on tips). In the previous chapter, we discussed solutions to an alternative popularity prediction task, which consists of ranking a group of tips based on their predicted popularity at a future time.

We here consider two scenarios of popularity levels: two levels (*low* or *high* popularity), and three levels (*no*, *low* or *high* popularity). Table 6.1 presents, for each scenario, the ranges of number of likes per tip in each level as well as a numerical representation of each category (used for computation purposes). The rightmost column will be discussed in Section 6.5.1. In the first scenario, tips with at most 4 likes are in the low popularity level, whereas the others are grouped into the high popularity level. In the second scenario, tips with low popularity receive from 1 to 4 likes whereas tips with high popularity receive at least 5 likes (like in scenario 1).



Table 6.1: Distribution of Candidates for Prediction Across Different Popularity Levels

Scenario	Popularity Level	Tips’s Category	Number of Likes per Tip	Number of Candidate Tips
1	Low	0	$< 5$	703,827
	High	1	$\geq 5$	3,427
2	No	0	0	589,044
	Low	1	1-4	114,783
	High	2	$\geq 5$	3,427

We note that we also experimented with other number of categories (e.g., 4 different categories) as well as different range definitions. Fundamentally, the limiting factor here is the imbalance between the tips. So, even if we consider two levels, one composed by tips that did not received any like and another level with tips that received at least one like, the imbalance would be still severe and influential in the results. Another point to consider when defining these categories is the availability of enough examples from each category for training, i.e., learning the model parameters. We here show results for the two scenarios presented in Table 6.1, although, in general, the main conclusions hold with all other categorizations we experimented with. We also note that we tried to cluster tips using a set of attributes (the attributes discussed in Section 6.4) and various types of clustering techniques (e.g., k-means [Hartigan and Wong, 1979], x-means [Pelleg and Moore, 2000], spectral clustering [Shi and Malik, 2000] and density-based clustering [Ester et al., 1996]). However, the resulting clusters were not stable in the sense that clustering results varied with the seeds provided to the algorithm, and no single number of clusters could be determined as the best one. The same clustering behavior was previously observed in the prediction of quality of questions in question answering services [Li et al., 2012]. This may be due to the lack of discriminative attributes *specifically for the clustering process*, which does not imply that the attributes cannot be useful for the prediction task, our main goal here.

## 6.2 Tip Popularity Prediction: Formal Definition

The problem we tackle in this chapter can be formally defined as follows. Given a tip  $p_i$ , posted at time  $tp_i$ , predict the *popularity level* of  $p_i$  at a given future target time  $tp_i + \delta$ , where the popularity of the tip is estimated by the total number of likes received until  $tp_i + \delta$  (prediction window). Moreover, we consider various scenarios

where the predictions are performed at time  $tp_i + \varepsilon$ , where  $\varepsilon < \delta$  defines a period after posting time, during which the tip is monitored up until prediction. Note that  $\varepsilon$  may be zero, corresponding to predictions at posting time. Figure 6.1 illustrates the prediction scenarios considered in our study.

We also consider sets  $V$  and  $U$  of venues and users, respectively, where  $u_i \in U$  is the user who posted  $p_i$ , and  $v_i \in V$  is the venue where it was posted. Each tip  $p_i$ ,  $U$  and  $V$  has a set of features  $F$  associated with it. Collectively, the features associated with  $p_i$ ,  $u_i$ , and  $v_i$ , are used as inputs to a prediction model (see below) representing the given tip instance. For evaluation purposes, each tip  $p_i \in P$  is labeled with a numeric value that represents the popularity level of  $p_i$  at time  $tp_i + \delta$ . The values of the features associated with each entity are computed using information available up to  $tp_i + \varepsilon$ , which defines when the prediction is done.

Next, we define the learning methods used in our experiments to predict the popularity level of a tip.

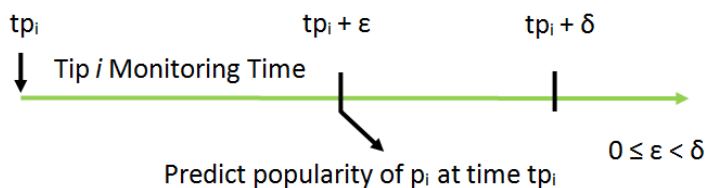


Figure 6.1: Monitoring Time Scheme.

### 6.3 Prediction Methods

We explore the use of two techniques – regression and classification – in the design of our models to predict the popularity level of a tip. We experiment with one classification algorithm – Support Vector Machine (SVM) [Joachims, 1998] – and two types of regression algorithms – ordinary least squares (OLS) multivariable linear regression and Support Vector Regression (SVR) [Drucker et al., 1997] algorithm. These methods are described in the following sections.

As already mentioned, we focused on predicting the tip’s popularity level instead of the total number of likes received by it. Previous work, mentioned in Chapter 2, has predicted the helpfulness of a review defined as a fraction  $\frac{x}{y}$ ,  $x$  out of  $y$  people found the following review helpful [Liu et al., 2008; Hong et al., 2012], the review average rating (real number between 0 and 5) [Lu et al., 2010; Moghaddam et al., 2012], and the number of tweets about an article [Bandari et al., 2012]. All these previous efforts were pursued using regression approaches. We here experiment with two types of

regression algorithms. Both produce as output a real value, which is rounded to the nearest integer that represents a popularity level, as done for predicting the helpfulness of movie reviews in Liu et al. [2008]. We chose to do so after performing some initial experiments comparing the use of regression models to predict the number of likes of a tip and its popularity level. These experiments indicated using regression to predict the popularity level leads to better results.

### 6.3.1 Support Vector Machines (SVM)

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (class label) and several observed variables or features. The goal of the classification algorithm is to learn a model based on the training data which predicts the class labels of the test data given only the test data features.

The support vector machine (SVM) learning algorithm [Joachims, 1998] is among the most popular supervised classification methods, and has proven successful across many domains such as document classification [Manevitz and Yousef, 2002], detection of malicious users [Benevenuto et al., 2012], as well as prediction of the localization of proteins [Hua and Sun, 2001].

Let a training set of instance-label pairs  $(\mathbf{x}_i, y_i)$ ,  $i=1, \dots, k$ , where  $\mathbf{x}_i \in \mathbb{R}^k$  is a vector representing the observed features and  $y_i$  is the class to which the sample instance belongs, which, in our case, corresponds to the tip's popularity level. A function  $\phi$  maps the training vectors  $\mathbf{x}_i$  into a higher dimensional space, such that the data under consideration has become separable by a hyperplane with maximum margin. SVM selects the plane that lies furthestmost from all defined classes (maximal margin hyperplane) from a set of infinite number of existing separating hyperplanes. Since the  $\mathbf{x}_i$  vector can have many dimensions or even be infinite, the  $\phi$  function computation becomes non trivial and computationally costly. To simplify this computation, SVM uses a class of functions called kernel functions. In particular, in our experiments, we used both linear and radial basis function (RBF) kernels, available in the LIBSVM open source package [Chang and Lin, 2001]. We note that the latter (RBF) does handle non-linear relationships between the target value and the predictor variables (features).

### 6.3.2 Ordinary Least Square Regression (OLS)

We also considered an ordinary least square (OLS) multivariable linear regression model to estimate the popularity level of a tip  $p_i$ , at a given point in time  $t$ ,  $\mathcal{R}(p_i, t)$ <sup>1</sup>, as a linear function of  $k$  predictor variables,  $x_1, x_2 \dots x_k$ , i.e.:

$$\mathcal{R}(p_i, t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (6.1)$$

Model parameters  $\beta_0, \beta_1 \dots \beta_k$  are determined by the minimization of the least squared errors [Jain, 1991] in the training data, as will be discussed in Section 6.5.1.

### 6.3.3 Support Vector Regression (SVR)

We also consider the more sophisticated Support Vector Regression (SVR) algorithm [Drucker et al., 1997], which is a state-of-the-art method for regression learning. SVR has been applied to several problems including the estimate of the quality of articles in collaborative digital libraries [Dalip et al., 2011] and regional logistic demand forecasting [Yang et al., 2010]. Unlike our OLS model, SVR does not consider errors that are within a certain distance of the true value (i.e., within the *margin*). Moreover, as SVM, SVR allows the use of different kernel functions, which helps solving a larger set of problems, compared to linear regression. Like for SVM, we explore both linear and RBF kernels, also available in the same package.

In addition to the SVM, OLS and SVR models, we also consider a very simple strategy that predicts the level of popularity of a tip  $p_i$  posted by a user  $u_i$  using the median number of likes received by tips previously posted by  $u_i$  as an estimate of the number of likes the new tip will receive. Our interest is to assess the improvements in prediction accuracy of using either the SVM, OLS or SVR models over this simpler (baseline) approach which takes into account only the popularity of previous posted tips of the user. We refer to it as the *median* strategy.

## 6.4 Tip Features

For the classification task, we exploit the same set of features proposed for the ranking task (Chapter ) with exception of two features: the number of hours since posting time (tip\_age\_hours) and the number of likes it has already received (tip\_likes\_current)<sup>2</sup>

<sup>1</sup>As already mentioned, the tip's popularity level is obtained rounding  $\mathcal{R}_t$  to the nearest integer.

<sup>2</sup>This feature is used when we evaluate scenarios were  $\varepsilon > 0$ .

that are more specific to scenarios in which we consider a set of tips posted at different times. Recall, that here we are predicting the popularity of a single tip at posting time.

We considered two scenarios when computing the user features, namely: (1) all tips posted by the user, and (2) only tips posted by the user at venues of the same category of the venue where  $p_i$  was posted. To distinguish between these two sets of user features, we refer to the latter as **user/category features**.

In addition to the features listed in Section 5.3, we also considered other linguistic features extracted from the tip’s textual content. These features are based on aggregate statistics that capture the readability<sup>1</sup>, informativeness, and structural features<sup>1</sup> of the tip’s text. Readability features or tests are used to estimate the difficulty readers may have in reading and comprehending a text. These tests produce scores that combine several factors affecting the clarity of the text, such as the number of words, syllables and sentences [Dalip et al., 2013]. We consider six of such tests:

- *Automated Readability Index (ARI)* (**tip\_read\_ari**) [Senter and Smith, 1967]: is derived from the ratios representing word difficulty (number of characters per word) and sentence difficulty (number of words per sentence).
- *Flesch Reading Ease* (**tip\_read\_flesch**) [Flesch, 1948]: measures the difficulty to read a text, in a 100-point scale, based on the average number of syllables per word and words per sentence. The higher the Flesch Reading Ease score, the easier it is to understand the document.
- *Flesch Kincaid Grade Level* (**tip\_read\_kincaid**) [Ressler, 1993]: translates the Flesch Reading Ease score into U.S. grade level of education required to understand the text<sup>2</sup>.
- *Coleman-Liau* (**tip\_read\_coliau**) [Coleman and Liau, 1975]: consists of the average number of characters per word and the number of sentences in a fragment of 100 words<sup>2</sup>.
- *Gunning Fog* (**tip\_read\_fog**) [Gunning, 1952]: indicates the number of years of education required for a reader to understand the text<sup>2</sup>. It uses the average number of words per sentences and the average number of complex words (words with more than 3 syllables).
- *SMOG* (Simple Measure of Gobbledygook) (**tip\_read\_smog**) [McLaughlin, 1969]: estimates the years of education a person needs to understand a piece

---

<sup>1</sup>Extracted from style readability tool (<http://www.gnu.org/software/diction/>)

<sup>2</sup>It outputs approximately the U.S. grade level necessary to comprehend the text.

of writing using the average of polysyllabic words (number of words of more than two syllables) taken from a sample of 30 sentences<sup>2</sup>.

Informativeness (**tip\_informat**) of a tip measures the novelty of the tip’s terms with respect to other tips posted at the same venue. This metric was used in [Hsu et al., 2009; Wagner et al., 2012; Momeni et al., 2013] to predict quality or helpfulness of comments or reviews. We derive informativeness using the Term Frequency Inverse Document Frequency (TF-IDF) measure, where we sum over the TF-IDF values for all terms in a single tip  $p_j$  (Equation 6.2):

$$inform(p_j) = \sum_{w_i \in p_j} TF_{i,j} IDF_i \quad (6.2)$$

The Term Frequency (TF) term captures the concept of popularity, being defined as (Equation 6.3):

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (6.3)$$

where  $n_{i,j}$  is the number of occurrences of the term  $w_i$  in the tip  $p_j$  and the denominator is the sum of the number of occurrences of all terms in the tip  $t_j$ . The Inverse Document Frequency (IDF) term measures the concept of specificity, valuing terms that are infrequently across all tips in the dataset. It is defined as (Equation 6.4):

$$IDF_i = \frac{\log|K(w_i)|}{|K|} \quad (6.4)$$

where  $|K(w_i)|$  is the number of tips where the term  $w_i$ , is found and  $|K|$  is the number of tips in the dataset.

We consider a range of structural features used in previous studies of on-line reviews [O’Mahony and Smyth, 2010; Castillo et al., 2011; Dalip et al., 2013; Momeni et al., 2013] In particular the following set of structural features are extracted from the tip’s text: fraction of capitalized words in the tip (**tip\_cap\_wrd**), fraction of unique words (**tip\_uniq\_wrd**), fraction of capitalized characters (**tip\_cap\_char**), entropy of tip word sizes (**tip\_ent\_wrd**), fraction of the characters that are spaces in the tip (**tip\_frac\_space**), number of sentences in the tip text (**tip\_sent\_num**), number of syllables (**tip\_syl\_num**), average number of syllables per word (**tip\_avg\_syl\_wrd**), average number of characters per word (**tip\_avg\_char\_wrd**), number of words in the longest sentence (**tip\_long\_sent\_wrd**), number of words with 3 or more syllables (**tip\_comp\_wrd\_num**), number of sentences in the tip’s text begin-

ning with a conjunction, article, interrogative pronoun, preposition, pronoun and subordinating conjunction (**tip\_beg\_conj**, **tip\_beg\_art**, **tip\_beg\_int\_pron**, **tip\_beg\_prep**, **tip\_beg\_pron**, **tip\_beg\_subor\_conj**), total number of conjunctions (**tip\_conj\_num**), total number of sentences that are questions (**tip\_quest\_sent**), total number of prepositions (**tip\_prep\_num**), number of passive voice sentences (**tip\_pass\_sent**), number of uses of verb “to be” (**tip\_tobe\_num**), and total number of pronouns (**tip\_pron\_num**) in the tip’s text.

Complementary to the above features, we also use semantic and topical features. This set includes the total number of named entities (**tip\_num\_entities**)<sup>1</sup>, number of distinct types of named entities (**tip\_num\_dist\_entities**)<sup>1</sup> as well as, psychological characteristics and sentiment polarities, further discussed next.

As in Momeni et al. [2013], we also use the psychological dimensions defined by LIWC. LIWC [Tausczik and Pennebaker, 2010] is a dictionary that includes more than 2,300 English words classified in psychologically categories that reflects people’s emotional and cognitive perceptions. It has been widely used in many contexts, specially for sentiment analysis in social networks [Wu et al., 2011b; Quercia et al., 2011; Park et al., 2012; Momeni et al., 2013]. In this work, we focused on the following categories [Inquiry and Count, 2007]: first, second and third persons (**tip\_liwc\_first**, **tip\_liwc\_second**, **tip\_liwc\_third**), positive and negative emotions (**tip\_liwc\_pos**, **tip\_liwc\_neg**), cognitive processes (**tip\_liwc\_cog**, **tip\_liwc\_inhib**, **tip\_liwc\_cause**), perceptual processes (**tip\_liwc\_perc**), time or temporal context (**tip\_liwc\_time**), verb tenses (**tip\_liwc\_past**, **tip\_liwc\_pres**, **tip\_liwc\_fut**), quantifiers (**tip\_liwc\_quant**), numbers (**tip\_liwc\_num**), personal concerns (**tip\_liwc\_work**, **tip\_liwc\_leisure**, **tip\_liwc\_money**, **tip\_liwc\_home**, **tip\_liwc\_relig**), swear words (**tip\_liwc\_swear**), social processes (**tip\_liwc\_social**, **tip\_liwc\_family**, **tip\_liwc\_friend**, **tip\_liwc\_humans**), negative feelings (**tip\_liwc\_anxiety**, **tip\_liwc\_anger**, **tip\_liwc\_sad**), biological processes (**tip\_liwc\_body**, **tip\_liwc\_health**, **tip\_liwc\_sex**, **tip\_liwc\_ingst**), relativity (**tip\_liwc\_relativ**, **tip\_liwc\_space**, **tip\_liwc\_motion**), achievement (**tip\_liwc\_achieve**), affective processes (**tip\_liwc\_affect**), and spoken category (**tip\_liwc\_nonfl**). We also used a feature to count the number of words that match any word in LIWC (**tip\_liwc\_match**).

Finally, we extract three more content features from the tip, also inspired in

---

<sup>1</sup>For the extraction of the Named Entity features, we used the Stanford Named Entity Recognizer (NER)[Stanford NLP Group, 2013].

studies about content quality [Agichtein et al., 2008; Chen et al., 2011; Dalip et al., 2013]: number of bad words in the tip text (**tip\_bad\_wrd**)<sup>1</sup>, number of words in the tip that are not in the English lexical WordNet (**tip\_not\_wordnet**), and number of words present in a list of common misspellings (**tip\_miss\_wrd**)<sup>2</sup>.

Thus, we represent each tip  $p_i$  by  $k = 125$  features related to the user  $u_i$  who posted tip  $p_i$ , the venue  $v_i$  where  $p_i$  was posted, and the textual content of  $p_i$ . The values of these features are computed up to  $(tp_i + \varepsilon)$ . All features are summarized in Table 6.2.

## 6.5 Methodology Evaluation

Having defined the features used by our strategies for predicting the popularity level of a tip, we now discuss our evaluation methodology. We start by presenting our experimental setup in Section 6.5.1, and then discuss our evaluation metrics in Section 6.5.2.

### 6.5.1 Experimental Setup

Our experimental setup consists, in general terms, of dividing the available data into training and test sets, learning model parameters using the training set, and evaluating the accuracy of the learned model in the test set. We split the tips chronologically into training and test sets, rather than performing random split, to avoid that the dataset used to build a model to estimate a given tip’s popularity level includes data related to tips that were posted *after* the time the given tip was created (or after the monitoring period  $tp_i + \epsilon$  expired). Next, to generate multiple runs, we slide both training and test windows along the time axis<sup>3</sup>. We do so 5 times, thus producing 5 results, as illustrated in Figure 6.2.

For the training sets, we considered the most recent tip posted by each user as *candidate* for popularity prediction, using the other tips to compute the features used as predictor variables. For the test set, we considered all tips posted during 1 month after the end of the training set.

For each candidate for prediction in both training and test sets, we computed the feature values by first applying a logarithm transformation on the raw numbers to reduce their large variability (as discussed in Section 4.1), and then scaling the

<sup>1</sup>This list was extracted from <https://gist.github.com/jamiew/1112488>

<sup>2</sup>[http://en.wikipedia.org/wiki/Wikipedia:Lists\\_of\\_common\\_misspellings](http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings)

<sup>3</sup>Another approach would be using rolling forecast validation scheme, increasing the training set in each run.



Table 6.2: Complete Set of Features for Tip Popularity Level Prediction

Type	Feature Name	Feature Description
User	user_tip_num	Total number of tips
	user_tipped_venues	Number of distinct venues
	user_total_likes	Total number of received likes <sup>1</sup>
	user_given_likes	Total number of given
	user_sn_size	Number of friends/followers
	user_sn_likes	Social likes
	user_tip_by_sn	Total number of tips by the social network <sup>1</sup>
	user_like_by_sn	Total number of likes given by social network <sup>1</sup>
	user_venue_visibility_total	User visibility at venue <sup>1</sup>
	user_type	Foursquare user category
	user_mayorships_num	Total number of mayorships
user_mayor	If the author was mayor of the venue	
Venue	venue_tip_num	Total number of tips
	venue_total_likes	Total number of received likes <sup>1</sup>
	venue_cks_num	Total number of check ins
	venue_visitors_num	Total number of visitors
	venue_verified	If the venue was verified
	venue_category	Foursquare venue category
	venue_like_rk_pos_*	Like ranking position *(ascending/descending order)
venue_date_rk_pos_asc	Date ranking position in ascending order	
Content	tip_char_num	Total number of characters
	tip_wrd_num	Total number of words
	tip_url_num	Total number of URLs or email addresses
	tip_pos_nn	Fraction of nouns
	tip_pos_adj	Fraction of adjectives
	tip_pos_adv	Fraction of adverbs
	tip_pos_comp	Fraction of comparatives
	tip_pos_ver	Fraction of verbs
	tip_pos_fw	Fraction of non-English words
	tip_pos_num	Fraction of numbers
	tip_pos_sup	Fraction of superlatives
	tip_pos_sym	Fraction of symbols
	tip_pos_pp	Fraction of punctuation
	tip_read_*	*(ARI, Flesch, Kincaid, Coleman-Liau, Gunning Fog and SMOG) Index
	tip_informat	Informativeness
	tip_cap_wrd	Percent of capital words
	tip_cap_char	Percent of capital characters
	tip_ent_wrd	Entropy of word sizes
	tip_frac_space	Space density
	tip_sent_num	Number of sentences
	tip_syl_num	Number of syllables
	tip_avg_syl_wrd	Average syllables per word
	tip_avg_char_wrd	Average characters per word
	tip_long_sent_wrd	Longest sentence size
	tip_comp_wrd_num	Number of complex words
	tip_beg_*	Number of sentences beginning with a * (conjunction, article, interrogative, pronoun, preposition, pronoun and subordinating conjunction)
	tip_conj_num	Total number of conjunctions
	tip_quest_sent	Total number of questions
	tip_prep_num	Total number of prepositions
	tip_pass_sent	Total number of passive sentences
	tip_tobe_num	Total number of to be verbs
	tip_pron_num	Total number of pronouns
	tip_num_entities	Total number of entities
	tip_num_dist_entities	Total number of distinct types of entities
	tip_liwc_*	Psychological characteristics *(LIWC classes)
	tip_*_score	*(Positive, negative, neutral) tip score
	tip_bad_wrd	Number of bad words
	tip_not_wordnet	Number of words not in the WordNet
	tip_miss_wrd	Number of misspellings

<sup>1</sup> Median, average and standard deviation are also included.

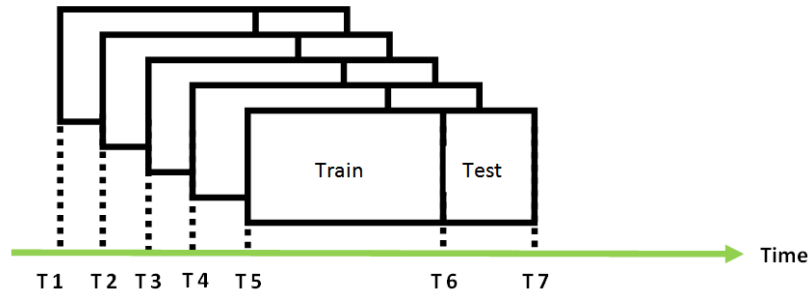


Figure 6.2: Chronological Split of Training and Test Sets: Sliding Windows Over Time.

results between 0 and 1. Moreover, in order to have enough historical data about users who posted tips, we considered only users who posted at least 5 tips. We also only considered tips that were posted at least one month before the end of training/test sets (otherwise the tip certainly has not received a lot of attention by the time of prediction). We also focused on tips written in English, since the textual features are computed by using tools which are available only for that language. After applying these filters, we ended up with roughly 700 thousand tips that are *candidates* for prediction. The distribution of these tips into the different popularity levels is shown in Table 6.1 (rightmost column).

OLS model parameters were defined by minimizing the least squared errors for predictions for the candidate tips in the training data. This can be done as their popularity levels (i.e., number of likes) is known. SVR model parameters were defined in a similar manner, as performed by the LIBSVM package. Moreover, SVR and SVM have some parameters that impact the error minimization process (e.g., parameters  $C$  and  $\epsilon$ ) [Drucker et al., 1997]. As in previous work [Dalip et al., 2011; Hsu et al., 2009], the best values of these parameters were selected using cross-validation within the training set, using the parameter selection tool provided by the LIBSVM package.

As shown in Table 6.1, the distribution of tips across the popularity levels (i.e., categories) is quite unbalanced. Such imbalance, particularly in the training data, poses great challenges to the regression and classification accuracy [He and Garcia, 2009]. Indeed, results from initial experiments using all available data were very poor and unsatisfactory.

A widely used strategy to cope with the effects of class imbalance in the training data is under-sampling [Liu et al., 2009]. Suppose there are  $n$  tips in the smallest category in the training set. We produce a balanced *training* set by randomly selecting equal sized samples from each category, each with  $n$  tips. Note that under-sampling is

performed only in the training set. The test set remains unchanged.<sup>1</sup> Because of the random selection of tips for under-sampling, we performed this operation 5 times for each sliding window, thus producing 25 different results, in total. The results reported in this chapter are thus averages of 25 results, along with corresponding 95% confidence intervals.

## 6.5.2 Evaluation Metrics

Many metrics have been used for evaluating the performance of classification tasks. A common technique is to evaluate the confusion matrix produced as a result of the testing phase of the classifier. A confusion matrix  $M$  summarizes all information about the actual and predicted values done by the classifier. Each matrix entry  $M[i, j]$  is computed as the number of testing samples belonging to class  $i$  which were assigned by the classifier to class  $j$ . The diagonal elements  $M[i, i]$  shows the numbers of correct classifications made for each class, whereas the off-diagonal elements show the errors.

Table 6.3 illustrates a three-by-three confusion matrix. From these entries, simple measures can be directly obtained. The true positive for each class  $i$ ,  $tp_i$ , is the number of samples correctly assigned to class  $i$ . The false positive for class  $i$ ,  $fp_i$ , is the number of samples that do not belong to class  $i$  but were incorrectly assigned to class  $i$  by the classifier. For example,  $fp_A = e_{BA} + e_{CA}$  in Table 6.3. The false negatives of class  $i$ ,  $fn_i$ , is the number of samples that were not assigned to class  $i$  by the classifier but actually belong to it. In Table 6.3,  $fn_A = e_{AB} + e_{AC}$ .

Table 6.3: Confusion Matrix for a Three-Class Classification Task

		Predicted class		
		A	B	C
Known Class	A	$tp_A$	$e_{AB}$	$e_{AC}$
	B	$e_{BA}$	$tp_B$	$e_{BC}$
	C	$e_{CA}$	$e_{CB}$	$tp_C$

The most commonly used metric derived from a confusion matrix is the overall *accuracy*. It is computed as the sum of correct classifications divided by the total number of classifications. However, with highly skewed data distribution, as found in our dataset, the accuracy metric does not always provide the full picture. For instance, a model that predicts all samples as negative has high accuracy, but it is inefficient to

<sup>1</sup>Although under-sampling changes the distribution of tip categories between the training and test sets, this technique has demonstrated to be effective in other applications with skewed data [Dalip et al., 2011; Benevenuto et al., 2012].

detect rare positive samples [Tang et al., 2009]. Thus, the performance of a classifier in applications with class imbalance should not be expressed in terms of average accuracy.

The *precision* for each class  $i$ , defined as  $\frac{tp_i}{tp_i + fp_i}$ , is also sensitive to changes in data distribution as will be discussed in Section 6.6. In contrast, the *recall* metric, defined as  $\frac{tp_i}{tp_i + fn_i}$ , provides a less sensitive measure of the classifier performance [He and Garcia, 2009]. We also use macro-averaged precision and recall over all the experiments. Macro-average scores are calculated by first computing precision and recall for each class  $i \in C$ , where  $C$  is the set of class labels, and then taking the average of these values (Equation 6.5 and 6.6). These macro-averaged measures also prevent the results from being biased towards the larger class.

$$\text{Precision}_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{tp_i}{tp_i + fp_i} \quad (6.5)$$

$$\text{Recall}_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{tp_i}{tp_i + fn_i} \quad (6.6)$$

## 6.6 Experimental Results: Predictions at Posting Time

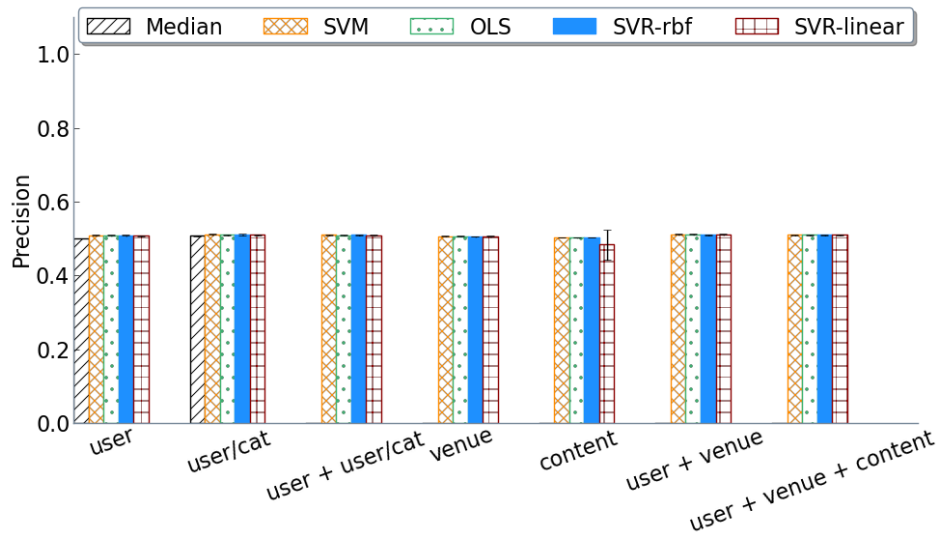
We now discuss representative results for predictions performed at the time the tip is posted. That is, we fix  $\varepsilon = 0$  and evaluate our proposed models to predict at time  $tp_i$  the popularity level that tip  $p_i$  will achieve at time  $tp_i + \delta$ . We also set  $\delta$  equal to 1 month and leave the evaluation for other values of  $\delta$  and  $\varepsilon$  to next sections.

We start by first investigating how the sets of features related to the three central entities – user, venue and tip – affect the prediction of tip popularity (Section 6.6.1). We then analyze the importance of each feature individually (Section 6.6.2).

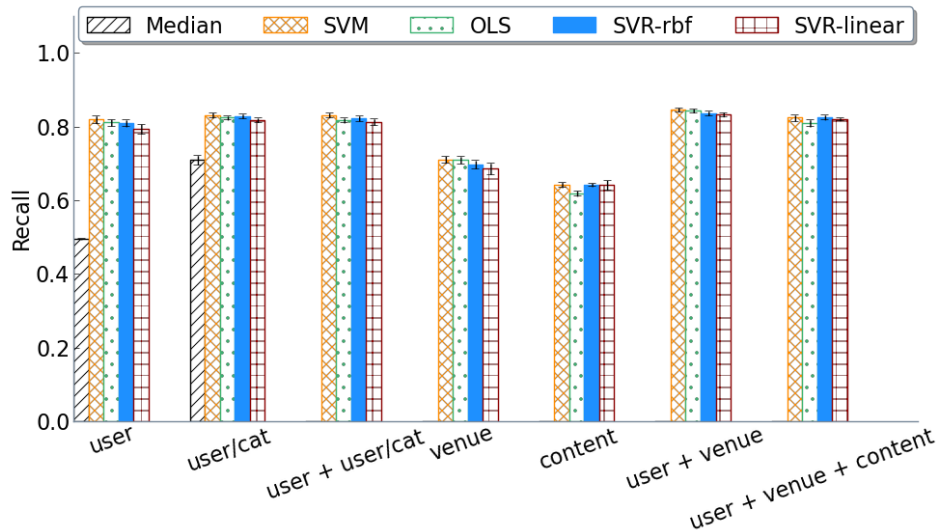
### 6.6.1 Analysis of the Groups of Features

Figures 6.3 and 6.4 show the macro-average precision and recall results along with corresponding 95% confidence intervals, for 30 different strategies for predicting a tip’s popularity level, considering two and three popularity levels respectively. These strategies emerge from the combination of five prediction algorithms with alternative sets of predictor variables. In particular, for OLS, SVR (linear and RBF kernels) and SVM<sup>1</sup> algorithms, we consider the following sets of predictors: only user features, only venue

<sup>1</sup>We show only results for the RBF kernel, as the linear kernel produced similar results.



(a) Macro-Average Precision

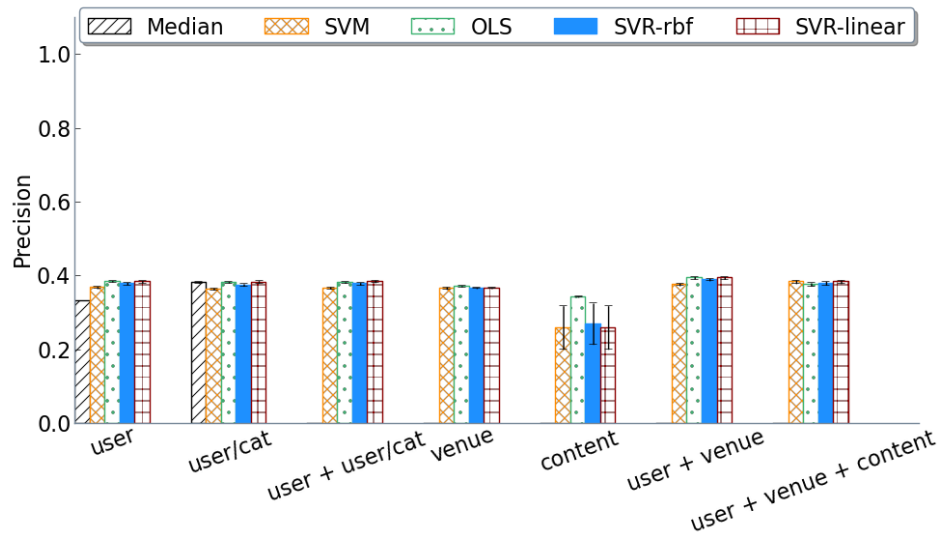


(b) Macro-Average Recall

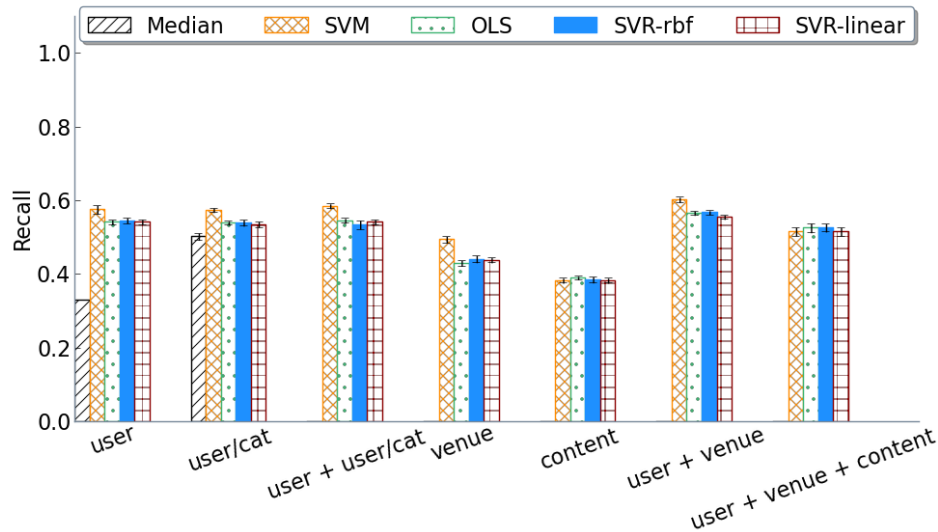
Figure 6.3: Macro-Average Results for Two Popularity Levels.

features, only content features, all venue and user features, all user, venue and content features. We also consider only user features restricted to the category of the venue where the tip was posted (user/cat features). For the predictions using the median number of likes of the user, here referred to as median strategy, we compute this number over all tips of the user and only over the tips posted at venues of the same (target) category<sup>1</sup>. The significance of the results was assessed using both one-way ANOVA and Kruskal-Wallis ([Allen, 1990]) tests with 95% confidence.

<sup>1</sup>The Figures 6.3, 6.4, 6.5 and 6.6 show results of median prediction model only for the user and user/cat sets of predictor variables since they are the same for the other sets.



(a) Macro-Average Precision



(b) Macro-Average Recall

Figure 6.4: Macro-Average Results for Three Popularity Levels.

We start by analyzing the two different scenarios of popularity levels. Comparing precision and recall results for both scenarios, we note that the two-level scenario produces the best results (gains of up to 93% for average precision and up to 68% for average recall). This result is not surprising, since it may be simpler to build a classifier to distinguish only between two classes than to consider more than two classes as the decision boundaries in the former case can be simpler [Galar et al., 2011]. Analyzing the confusion matrices for the scenario with three levels, we note that most of the misclassifications occur between the no popularity and the low popularity classes, which suggests that the proposed features cannot distinguish between them very well.

Thus, we only focus on the scenario with 2 popularity levels for the next analyses.

We observe that there is little difference in macro-average precision across the prediction algorithms, except for SVR with linear kernel using only content features, which performs worse (Figure 6.3a). The same degradation of SVR is observed in terms of macro-average recall (Figure 6.3b). Moreover, the superiority of SVM, SVR and OLS over the simpler median strategy, in terms of recall, is clear. For any of those strategies, the best macro-average recall was obtained using user and venue features, with gains from 1% (SVR-RBF) up to 36% (OLS) over the other sets of predictors. Comparing the different techniques using user and venue features, we see small but statistically significant gains in macro-average recall for SVM and OLS over SVR (up to 1.15% and 0.84%, respectively), while OLS and SVM produce statistically tied results. However, OLS is much simpler than both SVM and SVR, as will be discussed later in this section.

We also note that recall results are higher than precision results in Figures 6.3-a,b). This is because the precision for the smaller class (high popularity) is very low even with a high recognition rate (large number of true positives). This can be better analyzed when we look at the confusion matrices. Table 6.4 shows two confusion matrices taken from our experimental results with two popularity levels. Recall that under-sampling is applied only to the training set, and thus class distribution in the test set is still very unbalanced and dominated by the larger class (low popularity). Thus, even small false negative rates for the larger class results in very low precision for the other (smaller) class (see Table 6.4a). Thus, macro-average precision scores by themselves do not provide a completely fair picture about the number of correct predictions in all strategies. For that reason, towards a deeper analysis of the prediction strategies, we also discuss precision and recall results for each class (low and high popularity category).

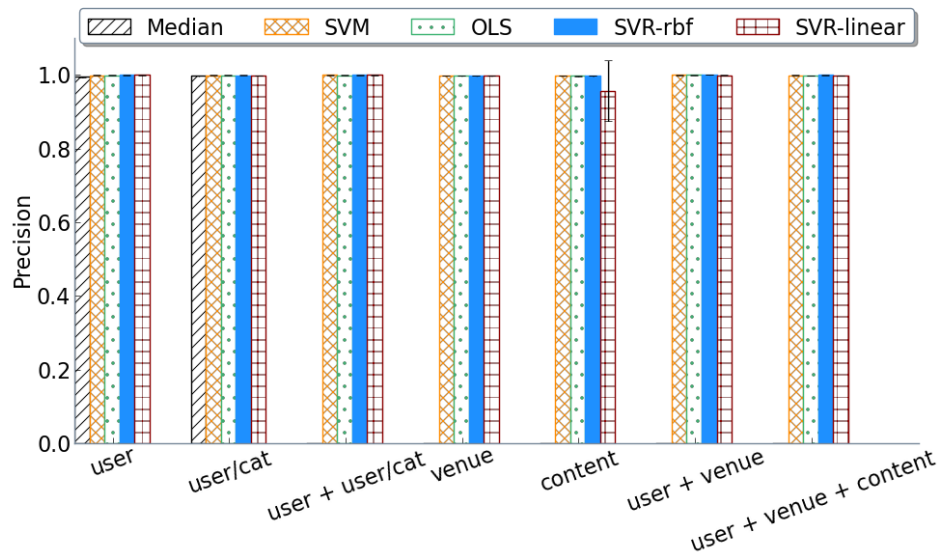
Predicted class				Predicted Class			
		Low	High			Low	High
Known Class	Low	74,054	30,733	Known Class	Low	89,237	15,550
	High	115	529		High	558	86

(a) OLS Method with User + Venue + Textual

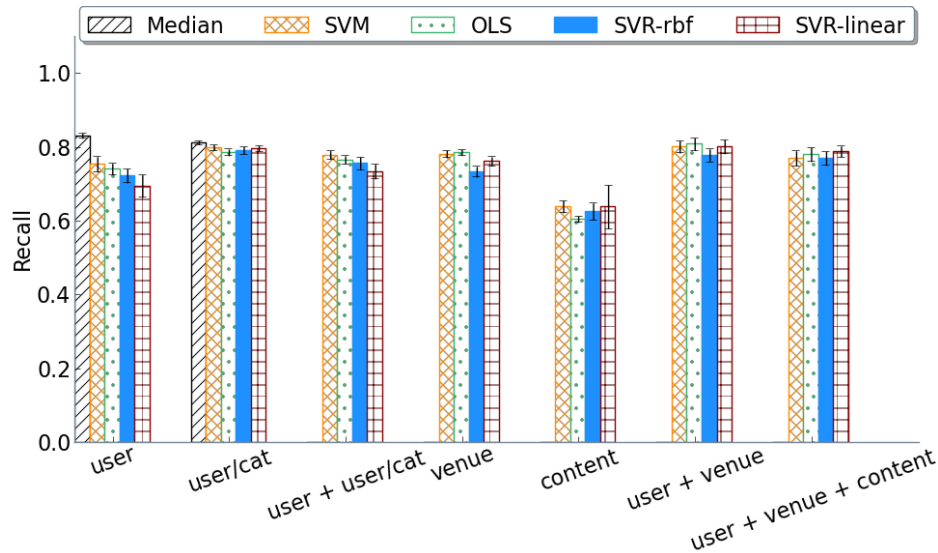
(b) Median Method

Table 6.4: Examples of Confusion Matrices for a Two-Class Classification Task

Figure 6.5 and Figure 6.6 show the average precision and recall along with corresponding 95% confidence intervals for tips in the low and high popularity categories, respectively. As discussed above, the precision values for tips in the low popularity class is much higher than the precision for tips in the high popularity category (shown



(a) Average Precision



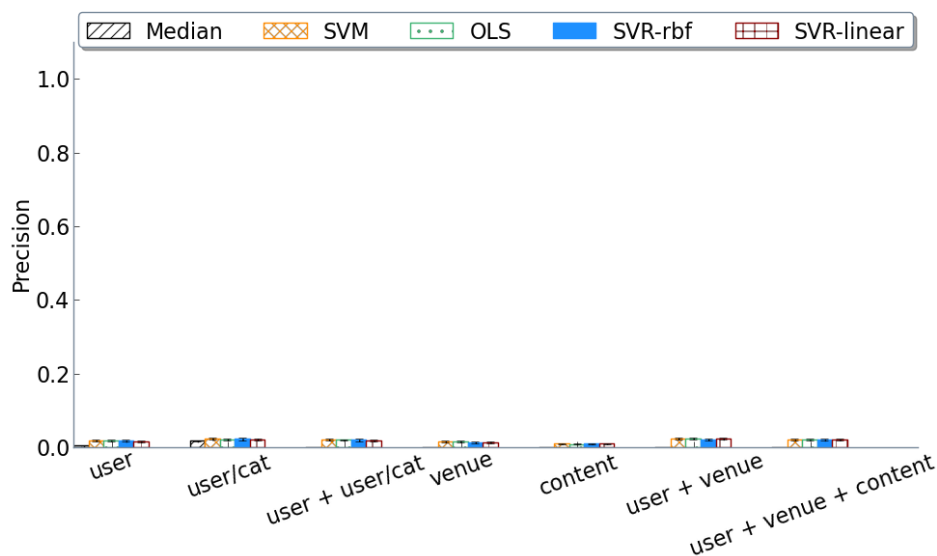
(b) Average Recall

Figure 6.5: Results for Tips in the Low Popularity Category.

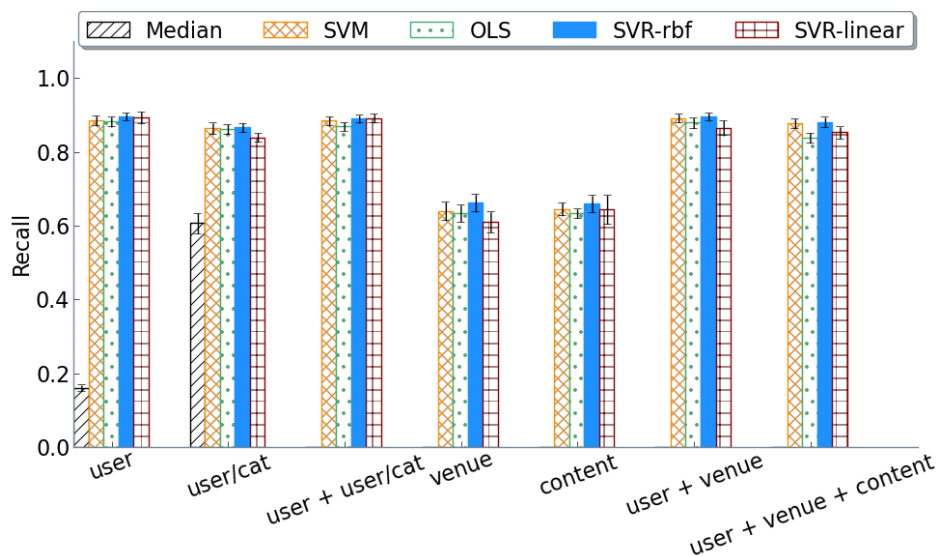
in Figure 6.6). For tips with low popularity, the best average recall results are produced by the median method. However, this result is a side effect of the large number of users with median number of likes equal to zero (as discussed in Chapter 4) which favors the median strategies. That is, the median strategies predict that most tips will have few likes as illustrated in Table 6.4b. This leads to large precision and recall for the larger class (low popularity) but very poor results for the smaller class.

Comparing the considered sets of predictor variables, we find a reasonable gain in recall (up to 15% for SVR with linear kernel), for tips in the low popularity class





(a) Average Precision



(b) Average Recall

Figure 6.6: Results for Tips in the High Popularity Category.

using both user and venue features over using either set of features separately. For tips with high popularity, there are no significant differences in the recall across the various sets of predictors, provided that user features are included. Moreover, the use of venue features jointly with user features improves the precision of the high popularity class (Figure 6.6a) in up to 46% (35% for the OLS algorithm). That is, adding venue features reduces the amount of noise when trying to retrieve potentially popular tips, with no significant impact on the recall of that class, and is thus preferable over exploring only user features. Note also that including content features as predictors lead to no further

(statistically significant) gains.

Comparing OLS, SVR and SVM with user and venue features as predictors, we find only limited gains (if any) of the more sophisticated SVR and SVM algorithms over the simpler OLS strategy. In particular, SVR (with RBF kernel) leads to a small improvement (2% on average) in the recall of the high popularity class, being statistically tied with SVM<sup>1</sup>. Yet, in terms of precision of that class, OLS outperforms SVR by 13.5% on average, being statistically tied with SVM. We note that SVR tends to overestimate the popularity of a tip, thus slightly improving the true positives of the high popularity class (and thus its recall), but also increasing the false negatives of the low popularity class, which ultimately introduces a lot of noise into the predictions for the high popularity class (hurting its precision). Note that the limited gains in recall of SVR over OLS come at the expense of a 30 times longer model learning process, for a fixed training set.

Thus, the simpler OLS model produces results that, from a practical perspective, are very competitive to (if not better than) those obtained with SVM and SVR.

## 6.6.2 Feature Importance

In the previous section, we concluded that the simpler linear OLS model using both user and venue features produces the best trade-off between prediction accuracy and complexity time of the model training phase. In this section, we evaluate the relative importance of all features, including textual features. Focusing on the OLS prediction strategy, we use a very popular feature selection technique, called Information Gain [Yang and Pedersen, 1997], to analyze the importance of each feature to prediction in our model (Section 6.6.2.1). Next, we examine the relationship between the response variable (popularity level) and each feature used as predictor (Section 6.6.2.2). We also investigate whether multicollinearity occurs in the regression and how it affects the prediction accuracy (Section 6.6.2.3). Finally, we analyze the impact on the model accuracy of removing features one at time according to the Information Gain metric (Section 6.6.2.4).

### 6.6.2.1 Ranking of Features

As we did for evaluating the features of the ranking task (Section 5.5.3), we have also sorted the features used by the OLS method using the Information Gain feature selection technique [Yang and Pedersen, 1997].

---

<sup>1</sup>For the low popularity level, OLS outperforms SVR (with RBF kernel) by 4% on average being statistically tied with SVM and SVR with linear kernel.

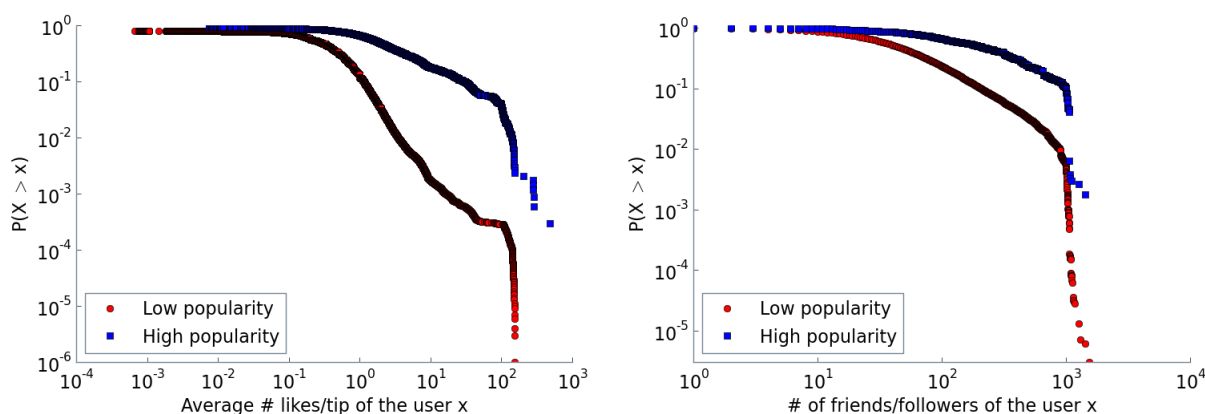
Table 6.5: Features Ranked by Information Gain.

Pos.	Feature	Pos.	Feature	Pos.	Feature
1	user_avg_likes	44	tip_liwc_match	87	tip_conj_num
2	user_total_likes	45	tip_num_entities	88	tip_sent_num
3	user_std_likes	46	tip_ent_wrd	89	tip_liwc_affect
4	user_sn_size	47	tip_pos_ver	90	tip_url_num
5	venue_median_likes	48	venue_verified	91	tip_liwc_pos
6	venue_visitors_num	49	tip_num_dist_entities	92	tip_liwc_pres
7	venue_cks_num	50	tip_pos_adv	93	tip_tobe_num
8	venue_category	51	tip_pos_adj	94	tip_liwc_quant
9	user_like_by_sn	52	tip_comp_wrd_num	95	tip_liwc_perc
10	user_type	53	user_median_tip_by_sn	96	tip_liwc_first
11	venue_total_likes	54	tip_cap_char	97	tip_liwc_past
12	venue_std_likes	55	tip_liwc_cog	98	tip_liwc_neg
13	venue_like_rk_pos_dsc	56	tip_pos_nn	99	tip_liwc_money
14	venue_avg_likes	57	tip_pos_comp	100	tip_liwc_body
15	venue_tip_num	58	tip_liwc_social	101	tip_liwc_cause
16	venue_date_rk_pos_asc	59	tip_pos_sup	102	tip_beg_prep
17	user_venue_visibility_total	60	tip_read_ari	103	tip_liwc_humans
18	user_sn_likes	61	tip_liwc_leisure	104	tip_pass_sent
19	user_tip_num	62	tip_pos_pp	105	tip_beg_pron
20	user_tipped_venues	63	user_mayor	106	tip_liwc_home
21	user_tip_by_sn	64	tip_liwc_time	107	tip_beg_art
22	venue_like_rk_pos_asc	65	tip_read_smog	108	tip_bad_wrd
23	user_std_like_by_sn	66	tip_uniq_wrd	109	tip_liwc_inhib
24	user_venue_visibility_std	67	tip_read_kincaid	110	tip_liwc_swear
25	user_venue_visibility_avg	68	tip_cap_wrd	111	tip_liwc_anger
26	tip_informat	69	tip_pos_num	112	tip_liwc_relig
27	venue_median_likes	70	tip_frac_space	113	tip_liwc_fut
28	tip_prep_num	71	tip_pron_num	114	tip_liwc_health
29	user_venue_visibility_median	72	tip_pos_fw	115	tip_liwc_anxiety
30	tip_long_sent_wrd	73	user_median_like_by_sn	116	tip_quest_sent
31	tip_syl_num	74	tip_neg_score	117	tip_liwc_sex
32	tip_pos_score	75	tip_read_fog	118	tip_beg_subor_conj
33	user_given_likes	76	tip_read_coliau	119	tip_liwc_sad
34	tip_liwc_relativ	77	tip_liwc_num	120	tip_liwc_family
35	tip_char_num	78	tip_avg_syl_wrd	121	tip_beg_int_pron
36	tip_wrd_num	79	tip_liwc_work	122	tip_beg_conj
37	user_mayorships_num	80	tip_liwc_secon	123	tip_liwc_friend
38	user_avg_like_by_sn	81	tip_liwc_ingst	124	tip_liwc_nonfl
39	tip_neu_score	82	tip_read_flesch	125	tip_miss_wrd
40	tip_liwc_space	83	tip_liwc_motion		
41	user_avg_tip_by_sn	84	tip_liwc_third		
42	tip_avg_char_wrd	85	tip_pos_sym		
43	tip_not_wordnet	86	tip_liwc_achieve		

Table 6.5 shows the ranking of the considered features according to their Information Gain, computed over all tips in our dataset. In consistency with the ranking task (Chapter 6.4), we find that the most important features are related to the popularity of the tip’s author: average, total and standard deviation of the number of likes received by tips posted by the user (*user\_avg\_likes*, *user\_total\_likes*, *user\_std\_likes*, respectively). Thus, the feedback received on previous tips of the user is the most important factor for predicting the popularity level of her future tips. Figure 6.7, which shows the

complementary cumulative distribution functions (CCDF) of the best of these features (average number of likes) for tips in each class, clearly indicates that it is very discriminative of tips with different (future) popularity levels. Similar gaps exist between the distributions of the other two aforementioned features.

Features related to the social network of the tip’s author are also important. The number of friends/followers of the author and the total number of likes given by them occupy the 4<sup>th</sup> and 9<sup>th</sup> positions of the ranking, respectively. Moreover, we find that authors of tips that achieve high popularity tend to have more friends/followers (Figure 6.7b). Thus, the social network does play a more important role for tips with very high popularity.



(a) Distribution of the Average Number of Likes per Tip of the User  
 (b) Distribution of the Number of Friends/Followers per User

Figure 6.7: Distribution of Most Important User Feature for Predicting a Tip’s Popularity Level.

The best venue feature, which occupies the 6<sup>th</sup> position of the ranking, is the number of unique visitors (Figure 6.8a). Moreover, the total number of check-ins (7<sup>th</sup> position), the venue category (8<sup>th</sup> position), and the total number of likes received by tips posted at the venue (11<sup>th</sup> position) (Figure 6.8b) are also very discriminative, appearing above other user features, such as number of tips (19<sup>th</sup> position), in the ranking.

Finally, note that the content features are the least important ones, which is consistent with our discussion, in previous chapter, about the low correlation between content features and number of likes per tip. For comparison purpose, Figure 6.9 shows the distributions of the informativeness scores (defined in Section 5.3.3) and number of prepositions per tip, the two most discriminative content features (26<sup>th</sup> and 28<sup>th</sup> positions, respectively). Clearly these features cannot discriminate tips with different

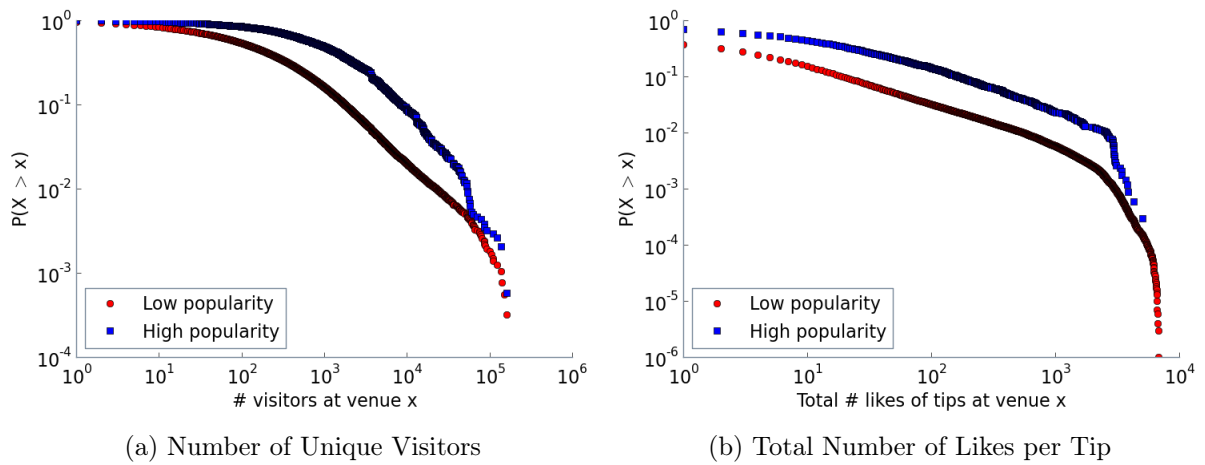


Figure 6.8: Distributions of the Most Important Venue Features for Predicting a Tip's Popularity Level.

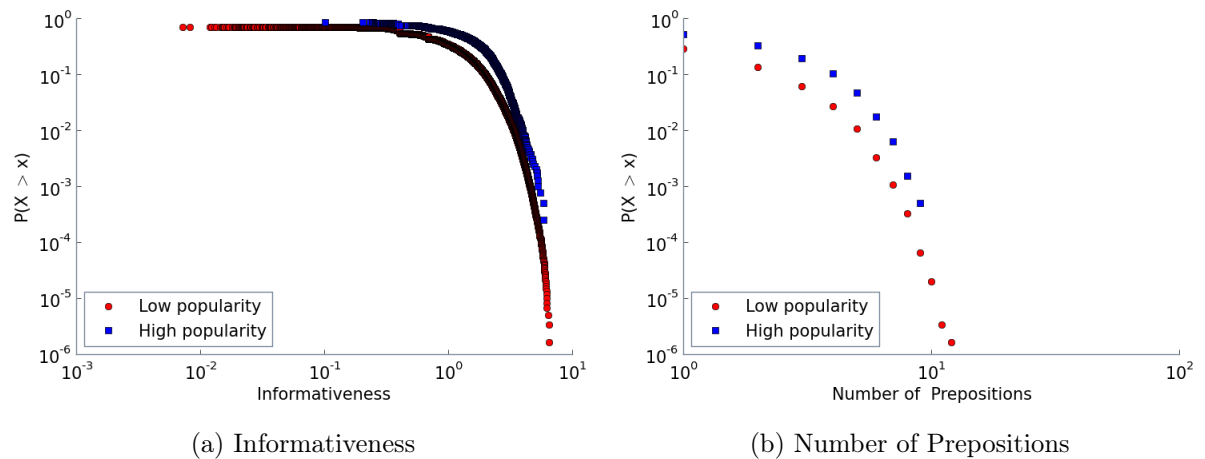


Figure 6.9: Distribution of the Most Important Content Feature for Predicting a Tip's Popularity Level.

levels of popularity very well. These results are consistent with those in Section 5.5.3 thus unlike in other efforts to predict the popularity or assess the helpfulness of online reviews [Kim et al., 2006; Zhang and Varadarajan, 2006], textual features are less important in the tip popularity prediction task. In contrast, user-related features, particularly those that capture the popularity of the tip's author in her previous tips and her social network, are much more important.

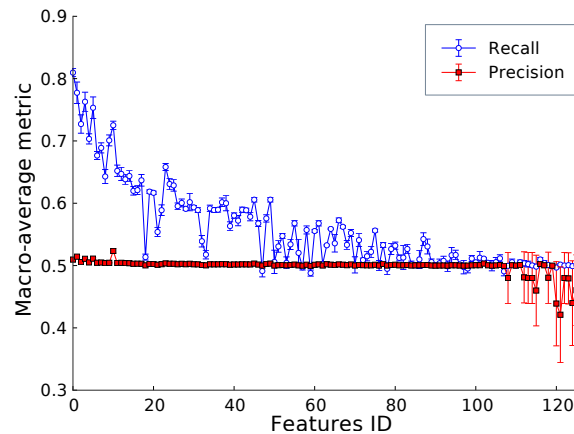


Figure 6.10: Macro-Average Precision and Recall for OLS using one feature at time

### 6.6.2.2 Effects of Individual Features

To understand the effects of each individual feature on the response variable (popularity level), we performed separate *univariate* linear OLS regressions with each feature, as in Borghol et al. [2012].

Figure 6.10 presents the macro-average precision and recall results along with corresponding 95% confidence intervals. Each point in the x-axis represents the results of the OLS regression with each given feature, sorted by the Information Gain (Table 6.5), used as single predictor. The first point ( $x = 0$ ) refers to the original model with all features. As we can see, the precision values are similar across all strategies, except at the end, in which the regressions were performed using the least discriminative features. By looking at the macro-average recall values, we note these values are larger for the most discriminative features, which, as discussed in the previous section, are related to the number of likes received by the tips previously posted by the user (*user\_avg\_likes*, *user\_std\_likes*, *user\_median\_likes*, *user\_total\_likes*). Finally, we observe that the highest recall value is obtained by the full model, which reinforces the importance of a multivariable model. Next, we present a collinearity analysis to identify correlated features, which might have a negative effect in the regression results.

### 6.6.2.3 Multicollinearity Analysis

Multicollinearity occurs in linear regression models when two or more predictor variables are linearly correlated. Although the SVM and SVR methods are capable of handling a high degree of collinearity [Howley et al., 2006], the OLS model might be severely impacted [Jain, 1991], as multicollinearity may increase the variance of the OLS coefficient estimates, degrading model predictability.

Table 6.6: Features with High Collinearity with at Least One Other Feature.

ID <sup>1</sup>	VIF	Tolerance	ID <sup>1</sup>	VIF	Tolerance	ID <sup>1</sup>	VIF	Tolerance
16	507.83 ± 44.80	0.004 ± 0.00	56	34.33 ± 1.74	0.06 ± 0.00	19	16.52 ± 2.07	0.13 ± 0.01
15	215.71 ± 15.94	0.01 ± 0.00	42	32.69 ± 2.29	0.06 ± 0.01	38	16.52 ± 2.10	0.13 ± 0.02
11	101.03 ± 4.78	0.02 ± 0.00	21	27.19 ± 3.47	0.08 ± 0.01	20	15.33 ± 1.67	0.13 ± 0.01
36	96.01 ± 9.53	0.02 ± 0.00	47	21.92 ± 0.99	0.09 ± 0.00	30	15.19 ± 0.38	0.13 ± 0.00
35	80.60 ± 7.81	0.03 ± 0.00	93	21.36 ± 1.19	0.09 ± 0.01	4	13.95 ± 1.67	0.15 ± 0.02
31	59.83 ± 4.17	0.03 ± 0.00	104	21.02 ± 1.24	0.09 ± 0.01	62	12.85 ± 0.60	0.15 ± 0.01
13	49.46 ± 3.08	0.04 ± 0.00	88	19.85 ± 1.47	0.10 ± 0.01	23	12.64 ± 0.91	0.16 ± 0.01
39	45.92 ± 2.66	0.04 ± 0.00	45	19.10 ± 1.17	0.10 ± 0.01	74	12.56 ± 0.80	0.16 ± 0.01
25	44.03 ± 1.25	0.05 ± 0.00	49	18.60 ± 1.21	0.11 ± 0.01	34	12.52 ± 0.58	0.15 ± 0.01
34	43.59 ± 2.76	0.05 ± 0.00	14	18.57 ± 1.11	0.11 ± 0.01	50	11.86 ± 0.54	0.16 ± 0.01
22	40.11 ± 1.51	0.05 ± 0.00	51	18.53 ± 0.72	0.11 ± 0.00	6	11.84 ± 0.53	0.16 ± 0.01
9	37.20 ± 6.03	0.06 ± 0.01	17	18.50 ± 0.57	0.11 ± 0.00	44	10.23 ± 0.43	0.19 ± 0.01
32	34.99 ± 2.15	0.06 ± 0.00	1	18.49 ± 0.71	0.11 ± 0.00			

<sup>1</sup> The feature IDs are defined by the position in the Information Gain ranking shown in Table 6.5.

There are several methods to test multicollinearity. As in Borghol et al. [2012], we here test whether our OLS prediction model is impacted by multicollinearity using two methods: variance inflation factors (VIF) [Stevens, 2002] and tolerance [O’Brien, 2007]. The VIF for a predictor  $k$ ,  $VIF_k$ , indicates whether there is a strong linear association between  $k$  and all the remaining predictors. If multicollinearity exists, the variance of an estimated regression coefficient  $b_i$  is inflated by  $VIF_k$  because of the correlation among the predictor variables in the model. The variance inflation factor for the  $i^{th}$  predictor is computed as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where  $R_i^2$  is the coefficient of determination obtained by a regression model built using the  $i^{th}$  predictor as response variable and the remaining predictors as inputs. Stevens [2002] recommends a heuristic VIF greater than 10 as an indication of multicollinearity requiring correction. The tolerance is measured as  $1 - R_i^2$ , where  $R_i^2$  is computed in the same way as in VIF. A tolerance of less than 0.10 indicates a multicollinearity problem [O’Brien, 2007].

We compute the VIFs and tolerances for all features of our complete OLS model. Table 6.6 lists the features with VIF values above 10, with their corresponding average VIF and tolerance values (and 95% confidence intervals). Several features are collinear with at least one other feature in the model. This is not totally unexpected as most collinear features are derived from other features (e.g., total and average numbers of likes of the user). Note also that some of these features have high Information Gain.

Next, we perform an experiment with the OLS method, eliminating one collinear feature at a time. We also compare the OLS model without the collinear feature and the

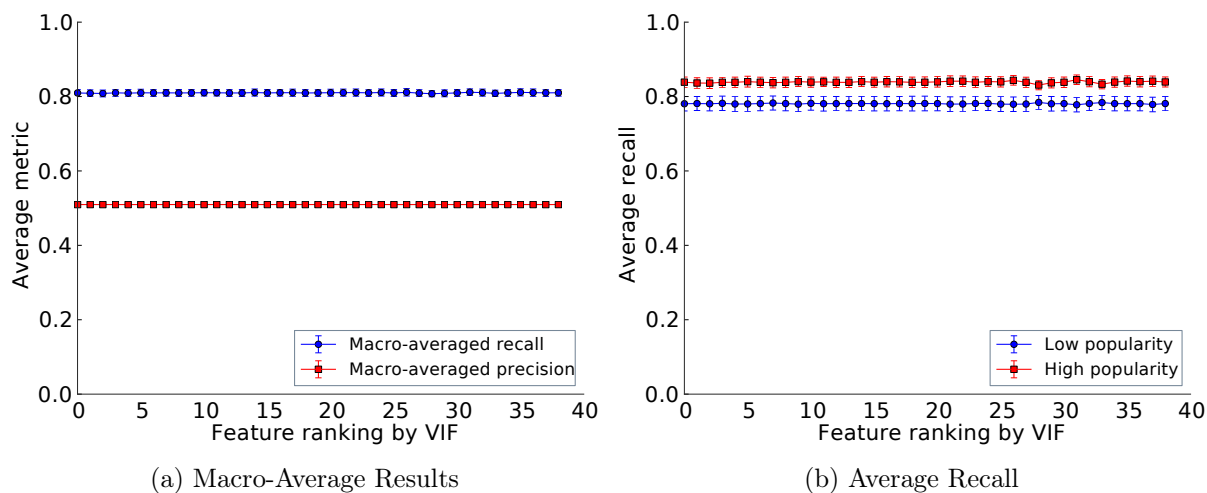


Figure 6.11: Macro-Average Results for OLS After Removing Each Collinear Feature.

original complete model. Figure 6.11a shows the macro-average results, while Figure 6.11b shows the average recall values for each popularity category. Each point in the x-axis represents a scenario when one collinear feature is eliminated: the eliminated feature is identified by its position in the ranking of VIF values (see Table 6.6), starting with the feature with largest VIF (i.e., feature ID 16). The first point ( $x = 0$ ) refers to the original method with all features. The figures show that the multicollinearity does not impact any of our metrics, i.e., the gains eliminating each multicollinear feature are not statistically significant neither for macro-average precision and recall nor for average per-class recall.

#### 6.6.2.4 Feature Removal using Information Gain

In our first experiments, we evaluated the relative importance of each group of feature, but we were not able to see which features are redundant in each group. Next, we perform an experiment removing one feature at a time cumulatively, in increasing order of importance given by the Information Gain as we did for the ranking task.

Figure 6.12 shows the impact on the macro-average recall and on the recall of the high popularity class as each feature is removed, starting with the complete set of user, venue and content features. For example, the second point in each graph shows results after removing the least discriminative feature (number of common misspellings per tip). We omit results for macro-average precision as they are not affected by the feature removal mainly because of the imbalance problem, discussed in Section 6.5. Note that as observed for the ranking task (Section 5.5.3), the classification task also has several features that are redundant, which means that the removal of many of



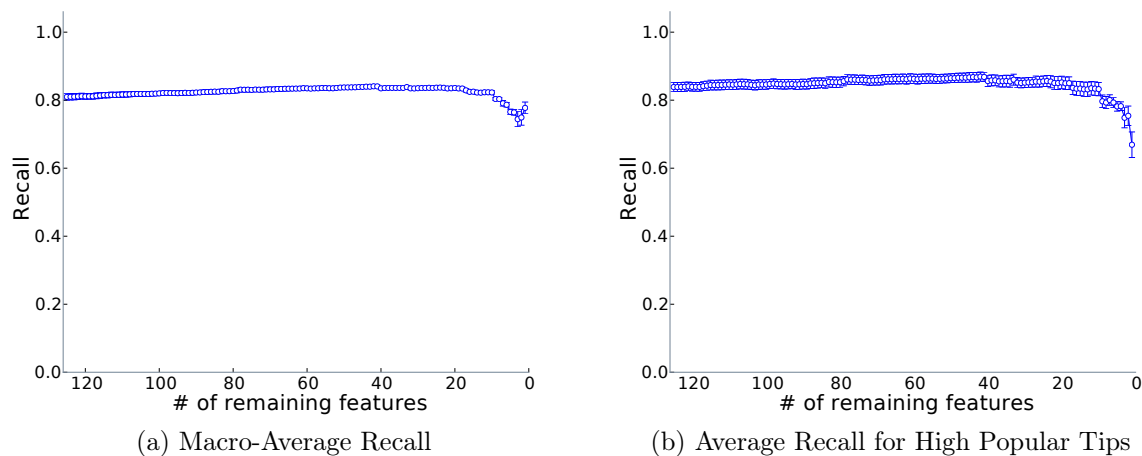


Figure 6.12: Recall for OLS When Removing One Feature at a Time.

the least discriminative features has no significant impact on recall. Significant losses in recall are observed only after we start removing features in the top-10 positions of the ranking. Among those, the largest losses are observed when the number of check ins at the venue (consistent with the ranking task), and the size of the user’s social network are removed, which reinforces the importance of including venue and social network features for the prediction task. In sum, using the top 10 most important features produces predictions that are as accurate as those of using the complete set of features.

## 6.7 Experimental Results: Other Prediction Scenarios

Ideally, one would like to predict the future popularity of a tip immediately after it is posted. This was the scenario considered in the previous section. However, tips exhibit different popularity evolution patterns, as observed in Chapter 4 as well as for other types of content (e.g., YouTube videos [Crane and Sornette, 2008; Pinto et al., 2013], Digg news [Szabo and Huberman, 2010], and tweets [Hong et al., 2011]). By monitoring the tips for a certain (short) time interval ( $\varepsilon$  units of time), we may be able to gather useful information about how its popularity is evolving, which may improve prediction accuracy.

Similarly, so far we have considered only predictions targeting one month ahead. However, depending on how the popularity of a tip evolves over time, we may be able to make accurate predictions further in the future (i.e., larger values of  $\delta$ ). One

concern in this case is that, as we predict further into the future, the information used as predictors may get outdated, hurting prediction accuracy. Thus, it is interesting to evaluate how further into the future our models are still able to produce reasonably accurate predictions.

In order to address these questions, we analyzed the impact of varying the monitoring time  $\varepsilon$  and the target prediction window  $\delta$  on prediction accuracy in Sections 6.7.1 and 6.7.2, respectively.

### 6.7.1 Prediction Results Varying the Monitoring Period $\varepsilon$

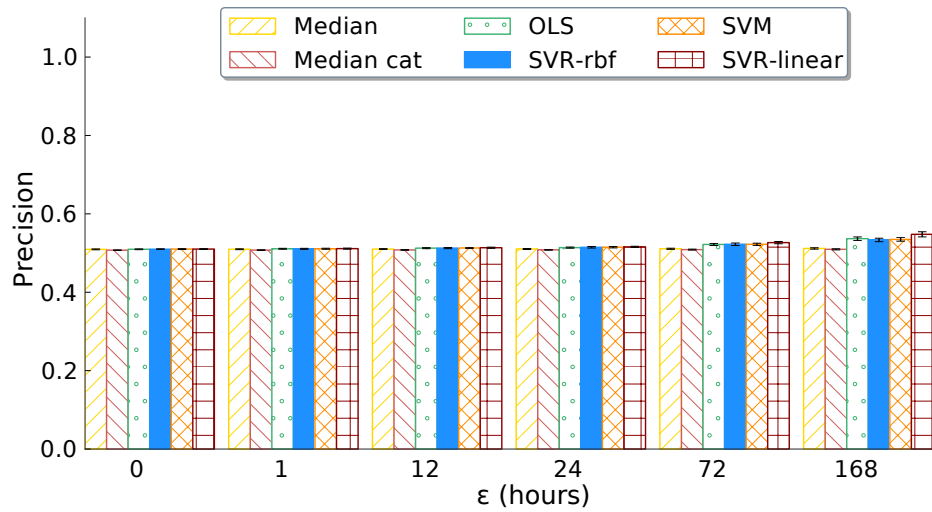
We investigated the accuracy of our prediction models for duration of the monitoring period ( $\varepsilon$ ) equal to 1, 12, 24, 72 and 168 hours (one week), fixing the target time  $\delta$  equal to 1 month. Recall that, in these scenarios, we use as predictors all the features listed in Section 6.4 as well as the number of likes already received by the tip  $p_i$ . For a given  $\varepsilon$ , the values of the predictors are computed taking all past history up to  $\varepsilon$  hours after the tip's posting time.

Figure 6.13 presents macro-average precision and recall results for each prediction method: SVM, SVR (with both linear and RBF kernels) and OLS. These results are produced using the complete set of features. The figure also shows results for the median baseline using all user features as well as only user/cat features (referred to as median-cat). Note that the results for  $\varepsilon=0$  are the same results presented in Section 6.6.1.

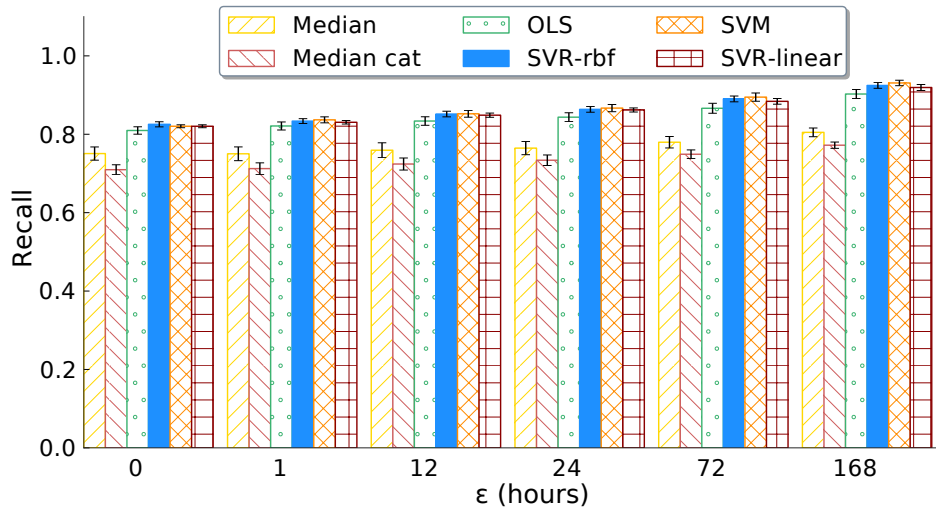
We note that extending the monitoring time to only 1 hour after the tip is posted only slightly improves the macro-average recall (up to 1.60% for SVM method), which is expected given the slow evolution of tip popularity observed (see discussion in Chapter 4). Yet, by monitoring the tip for one week ( $\varepsilon=168$  hours) we can improve the macro-average recall in up to 13% (SVM method) and the macro-average precision in up to 7% (SVR linear method) over using features computed at posting time. Moreover, such improvements are observed for all methods, although the median baselines are still much worse than our solutions. In particular, the OLS model remains as the most cost-effective prediction method as it produces results that are statistically as good as those obtained with the other (more costly) methods, for all considered values of  $\varepsilon$ .

Moreover, out of all considered features, the total number of likes received during the monitoring period ( $\varepsilon$ ) is the most discriminative feature according to the Information Gain criterion. This indicates the importance of taking the early popularity evolution as evidence for prediction.

In order to assess to which extent the other features contribute to prediction ac-



(a) Macro-Average Precision



(b) Macro-Average Recall

Figure 6.13: Macro-Average Results for Various Monitoring Times  $\varepsilon$  ( $\delta = 1$  month).

curacy, we also compare our OLS strategy against three other state-of-the-art baseline models that exploit only early popularity measurements. The first one, proposed by Szabo and Huberman [2010] and referred to as S-H model, is a univariate linear regression model on logarithmic scale that uses only the total number of likes received during the monitoring period as predictor. The other two baselines, proposed by Pinto et al. [2013], are extensions of the S-H model. One is a multivariable linear regression model that uses early popularity measures sampled at regular intervals (e.g., per day) during the monitoring period as predictors. The other builds on this model by also using Radial Basis Functions (RBFs) to capture the similarity (in the early popularity

Table 6.7: Macro-Average Results of Models that Use Early Popularity Measurements (only tips with at least 1 like,  $\varepsilon=168$  hours,  $\delta=1$  month).

Metrics	Models			
	OLS	ML	MRBF	S-H
Macro-average Recall	<b>0.8263</b> $\pm$ <b>0.0129</b>	0.7257 $\pm$ 0.0064	0.8003 $\pm$ 0.0085	<b>0.8395</b> $\pm$ <b>0.0100</b>
Recall low popularity	0.8305 $\pm$ 0.0186	<b>0.9960</b> $\pm$ <b>0.0005</b>	0.9619 $\pm$ 0.0022	0.9240 $\pm$ 0.0010
Recall high popularity	<b>0.8220</b> $\pm$ <b>0.0189</b>	0.4553 $\pm$ 0.0131	0.6386 $\pm$ 0.0176	0.7549 $\pm$ 0.0205
Macro-average Precision	0.5650 $\pm$ 0.0061	<b>0.8799</b> $\pm$ <b>0.0145</b>	0.6674 $\pm$ 0.0129	0.6158 $\pm$ 0.0112
Precision low popularity	<b>0.9930</b> $\pm$ <b>0.0012</b>	0.9827 $\pm$ 0.0021	0.9879 $\pm$ 0.0017	<b>0.9913</b> $\pm$ <b>0.0015</b>
Precision high popularity	0.1369 $\pm$ 0.0126	<b>0.7770</b> $\pm$ <b>0.0308</b>	0.3469 $\pm$ 0.0271	0.2403 $\pm$ 0.0237

measurements) between the target tip and selected examples from the training set. These variations are referred to as ML and MRBF, respectively. We use the S-H, ML and MBRF models as originally proposed, that is, to predict the total number of likes of a tip. We then use the predicted number to infer the corresponding popularity level.

We evaluate all models in the same scenario adopted by Szabo and Huberman [2010]; Pinto et al. [2013], i.e.,  $\varepsilon$  equal to 168 hours and  $\delta$  equal to 1 month. As discussed in Section 6.5.1, model parameters are defined using cross-validation in the training set. The number of RBF functions in the MRBF model was set to 100, as in Pinto et al. [2013].

We start by noting that although the S-H, ML and MRBF models can be directly applied to predict tip popularity, their use is constrained to tips with at least one like at the target time, since the models are solved by minimizing the mean relative squared error (MRSE) over the training set, which is undefined for tips with zero likes. Thus, in order to favor the baselines in our evaluation, we disregarded tips with zero likes, corresponding to 83% of all tips in our dataset<sup>1</sup>. Macro-average recall and precision results obtained with all models over the 17% remaining tips are shown in Table 6.7<sup>2</sup>.

Focusing first on recall, our primary metric of interest, we note that our OLS model produces the best average recall results for the high popularity class, with gains of 80%, 29% and 9% over the ML, MRBF and S-H models, respectively. The baselines are more biased towards the less popular tips, favoring the recall of this class. Yet, in terms of macro-average recall, the OLS model still outperforms both ML and MRBF models, being statistically tied with the S-H model. The gains of OLS in recall, particularly for the high popularity class, come at the cost of a decrease in precision. The baselines, especially ML, are able to better filter false positives out, leading to higher precision.

<sup>1</sup>In this setup, the low popularity class is defined as tips with number of likes ranging from 1 to 4.

<sup>2</sup>Note that the OLS results shown in this table are different from those in Figure 6.13, computed over all tips.

In sum, we find that the baselines, particularly the ML model, perform quite well in the considered scenario, confirming previous results, except in terms of the recall of the high popularity class. However, this metric is of particular interest if one is aiming at retrieving most of the potentially more popular tips even if this comes at the expense of some noise. In such case, our OLS model is preferable.

Moreover, we emphasize that our solution is more robust, since, unlike the baselines, it can be applied to any tip, at or after posting time. In particular, it can be applied to tips that have not received any like yet (i.e., unpopular tips or tips that have just been posted). The baselines, instead, are more suitable to types of content that exhibit a faster popularity evolution (e.g., news, videos)<sup>1</sup>. Given the results discussed in Chapter 4, it is very likely that the baselines will not be applicable to the vast majority of the tips, even for other values of  $\delta$ . Our solution is thus more general and preferable.

### 6.7.2 Prediction Results Varying Target Prediction Window $\delta$

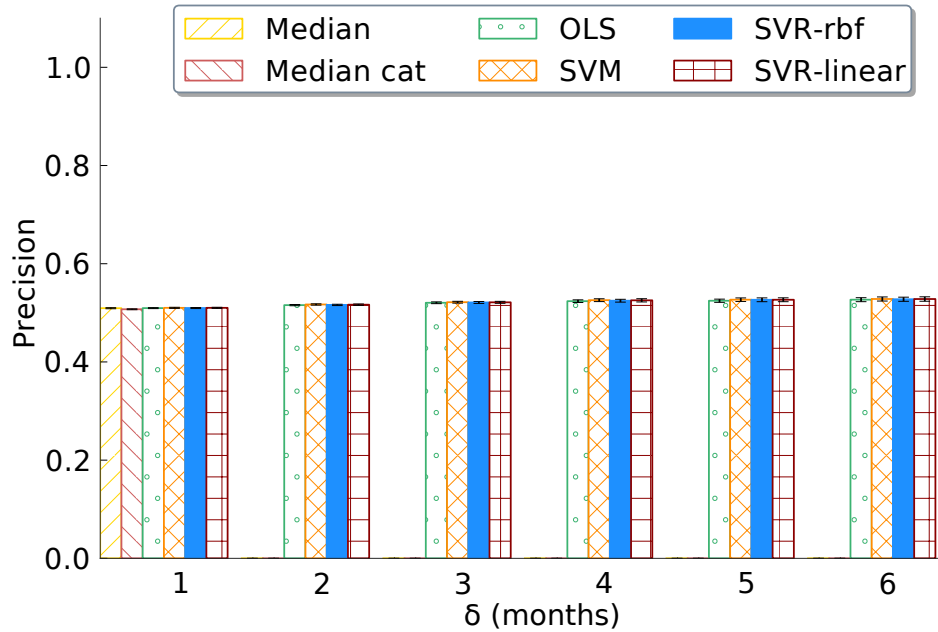
We now analyze how prediction accuracy is affected as we vary the target prediction window  $\delta$  from 1 to 6 months. In all scenarios, we set  $\varepsilon$  to 0, thus focusing on predictions at posting time. We evaluate the OLS, SVM and SVR (with both kernels) models using all features. As baselines, we consider once again the median and median-cat models, since the S-H, ML and MRBF methods are not suitable to predictions at posting time.

Figure 6.14 shows macro-average precision and recall results for all methods and values of  $\delta^2$ . Note that the results for  $\delta = 1$  are the same as those presented in Section 6.6.1. We find that, for any given value of  $\delta$ , OLS, SVR (with both kernels) and SVM produce statistically tied results in terms of macro-average precision (Figure 6.14a). Moreover, the gains (if any) in macro-average recall (Figure 6.14b) of the SVR and SVM methods over OLS are limited (up to 3.15% when  $\delta = 5$  months). Thus, once again, OLS is a cost-effective solution, from a practical perspective, for various values of  $\delta$ .

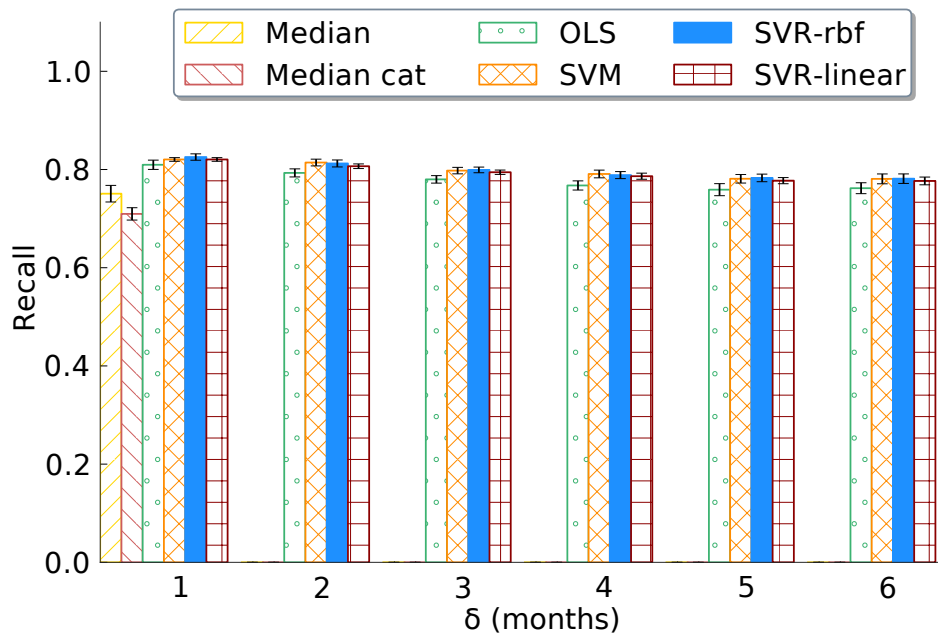
Moreover, we find the same trend for all methods: as  $\delta$  increases, precision increases slightly but at the cost of a (small) reduction on recall. For example, comparing the predictions done by OLS for  $\delta$  equal to 1 and 6 months, we observe a small improvement in macro-average precision (3.37%) for the latter, but also a decrease in

<sup>1</sup>Pinto et al. [2013] reported that less than 1.5% of the videos in their datasets had not received any view in the first month in the system. This is in sharp contrast to 83% of the tips with no likes in the same period in our dataset.

<sup>2</sup>The tip distribution across classes may not be the same for experiments with different values of  $\delta$  since tips from the low popularity class may move to the high popularity class as  $\delta$  increases.



(a) Macro-Average Precision



(b) Macro-Average Recall

Figure 6.14: Macro-Average Results for Various Target Times  $\delta$  ( $\varepsilon=0$ ).

macro-average recall (5.9%). As shown in Figure 4.15, for a large fraction of tips, most likes are gained after 3 months, which suggests that a tip may take a long time to define its popularity level. Yet, we still observe a loss in macro-average recall of 3.7% when predictions are done for  $\delta=3$ . The losses in recall occur in both classes, reaching

6.35% and 5.47% for low and high popularity classes respectively, for  $\delta$  equal 6.

The gains in macro-average precision observed as  $\delta$  increases come from a higher precision for high popularity class, which, in turn, is due to the reduction of class imbalance (which severely hurts the precision of the smaller class) as more tips migrate from the low to the high popularity class. This migration might also partially explain the losses in recall. More generally, as we predict further ahead, model inputs (feature values) become outdated and less efficient for prediction purposes. Given our results, we find that predictions for up to 2 months ahead are mostly unaffected by outdated features. For longer periods, the reduction in recall starts becoming significant.

## 6.8 Model Specialization

So far, we have built and evaluated prediction models that were trained using *all* tips in the training set. This approach produces a single general prediction model that aggregates and summarizes the relationships between the predictors (feature values) and the response (popularity level) across all tips. In this section, we analyze whether we can improve prediction accuracy by building models that are specialized to particular groups of tips such as tips posted at venues located in the same geographical region or of the same category. Model specialization might bring up patterns that are inherent to that particular group of tips, but are masked when all tips are treated jointly. For example, venues in different categories might exhibit different patterns: while “Travel & Transport” is the most popular venue category in terms of number check-ins, “Food” is the category that attracts the largest number of tips [Li et al., 2013]. Model specialization might improve accuracy as fewer instances of noise are used to train the models. On the other hand, specialization might also suffer from the lack of enough training instances, which impacts prediction accuracy as it affects the model’s capacity to generalize, or from a more severe class imbalance which, due to the need of undersampling in the training set, ends up severely restricting the amount of training examples.

We here assess the benefits from building specialized models for specific cities (Section 6.8.1) and venue categories (Section 6.8.2). To that end, we compare specialized and general models, built using the same method (OLS, SVM or SVR) on the *same* test set, when performing predictions at posting time for one month in the future (i.e.,  $\varepsilon = 0$  and  $\delta = 1$  month).

We adopt the same general experimental setup described in Section 6.5.1, learning model parameters through cross-validation in the training set. There are two key differences though. First, the set of tips used as input to the experimental procedure is

restricted to tips posted in venues of the target city or category in case of specialization. Second, when training either the general or the specialized models, we here consider multiple tips posted by the same user as candidates for prediction, since, using only the most recent tip of each user, as discussed in Section 6.5.1, severely constrains the amount of data available for training.

### 6.8.1 City-Based Model Specialization

We start by assessing the benefits of building specialized models for specific cities. To that end, we build models using tips posted at venues located in four selected cities, namely New York (NY), Los Angeles (LA), San Francisco (SF), and Chicago (CHI)<sup>1</sup>. For each city, we compare the specialized model against a general model using the same test set composed of only tips posted at venues of the target city. Moreover, for a fair comparison, the general model is built using a training set of the same size of the one used to learn the specialized model. However, the former consists of tips posted in venues of all cities in the dataset, randomly sampled from the original (global) training set. Specifically, the training sets used to learn the models for the NY, LA, SF and CHI scenarios contain 1141, 293, 308, and 183 tips, respectively. Similarly, the learned models were applied on test sets including 5085, 1551, 1695 and 1443 tips, respectively.

Table 6.8: Geographical Model Specialization: Macro-Average Results

Scenarios	Macro-Average Recall							
	General Model				Specialized Model			
	OLS	SVM	SVR-rbf	SVR-linear	OLS	SVM	SVR-rbf	SVR-linear
NY	<b>0.8123</b> ± 0.0290	<b>0.8175</b> ± 0.0278	<b>0.8286</b> ± 0.0253	<b>0.8267</b> ± 0.0277	0.8147 ± 0.0169	<b>0.8099</b> ± 0.0176	<b>0.8242</b> ± 0.0189	<b>0.8069</b> ± 0.0287
SF	0.7817 ± 0.0271	<b>0.8395</b> ± 0.0313	<b>0.8520</b> ± 0.0197	<b>0.8519</b> ± 0.0167	0.8041 ± 0.0430	<b>0.8315</b> ± 0.0486	<b>0.8185</b> ± 0.0556	<b>0.8177</b> ± 0.0465
CHI	0.7379 ± 0.0295	<b>0.8038</b> ± 0.0370	<b>0.8161</b> ± 0.0328	<b>0.8188</b> ± 0.0337	0.7206 ± 0.0476	<b>0.7701</b> ± 0.0501	<b>0.7767</b> ± 0.0504	<b>0.7734</b> ± 0.0506
LA	0.7860 ± 0.0309	<b>0.8421</b> ± 0.0314	<b>0.8545</b> ± 0.0313	<b>0.8638</b> ± 0.0326	0.7983 ± 0.0504	<b>0.8489</b> ± 0.0383	<b>0.8594</b> ± 0.0326	<b>0.8540</b> ± 0.0330
Scenarios	Macro-Average Precision							
	General Model				Specialized Model			
	OLS	SVM	SVR-rbf	SVR-linear	OLS	SVM	SVR-rbf	SVR-linear
NY	0.5336 ± 0.0067	0.5345 ± 0.0072	0.5332 ± 0.0075	0.5372 ± 0.0083	<b>0.5537</b> ± 0.0090 ↑	<b>0.5499</b> ± 0.0081 ↑	<b>0.5485</b> ± 0.0091 ↑	0.5429 ± 0.0093
SF	0.5121 ± 0.0022	0.5204 ± 0.0065	0.5201 ± 0.0053	0.5244 ± 0.0074	0.5221 ± 0.0040 ↑	<b>0.5338</b> ± 0.0070 ↑	<b>0.5369</b> ± 0.0087 ↑	<b>0.5369</b> ± 0.0094 ↑
CHI	0.5098 ± 0.0015	0.5158 ± 0.0034	0.5163 ± 0.0031	0.5189 ± 0.0036	0.5118 ± 0.0023	<b>0.5259</b> ± 0.0051 ↑	<b>0.5269</b> ± 0.0046 ↑	<b>0.5272</b> ± 0.0050 ↑
LA	0.5169 ± 0.0035	0.5242 ± 0.0049	0.5260 ± 0.0051	0.5314 ± 0.0064	0.5265 ± 0.0072 ↑	<b>0.5407</b> ± 0.0078 ↑	<b>0.5426</b> ± 0.0078 ↑	<b>0.5423</b> ± 0.0083 ↑

Table 6.8 shows macro-average recall and precision results, along with corresponding 95% confidence intervals, for each model and method. For each scenario (city), the best methods (including statistical ties) are shown in bold. A ↑ (or ↓) sign is used to indicate a statistical improvement (or loss) of the specialized model over the corresponding general model. The lack of a sign indicates a statistical tie.

<sup>1</sup>These cities were selected as they have the largest number of tips in our dataset.



We observe that the specialized models outperform (or at least are statistically tied with) the corresponding general models in the vast majority of the cases. The improvements occur in terms of macro-average precision, varying from 1.60% to 3.76%. Such improvements in precision occur in the high popularity class. Indeed, the gains in average precision for that class achieve 84.87% for the SF scenario with SVR (with RBF kernel) model. In terms of macro-average recall, both specialized and general models are statistically tied. Moreover, unlike observed in the previous sections, we do find cases where SVM and SVR (with RBF kernel) significantly outperform the simpler OLS in terms of macro-average precision (up to 3.07%). Note, for example, the difference between these methods for SF, CHI and LA. Such gains are not directly related to the specialization, but rather to the greater robustness of SVM and SVR to smaller training sets [Alwee et al., 2013]. Indeed such gains are observed for both general and specialized models.

In sum, we find that city-based model specialization does bring some improvements, particularly in terms of precision, as specialized models are able to more accurately capture patterns that are specific to the target city, reducing the amount of false negatives. One point to note, though, is that the amount of information available to learn a specialized model is inevitably smaller, if compared to a general model, and may require the use of techniques that are more robust to the lack of training instances.

As a final observation in this analysis, we note that by evaluating specialized models for each city, we are also analyzing the benefits from adding spatial (i.e., geographic) factors to our prediction models. We consider spatial factors at the city level because, on Foursquare, the geographic information associated with each user, one of the central entities of the popularity prediction problem, is available only at the city level.

We also note that spatial information at a

finer granularity (i.e., latitude and longitude coordinates) is available only for venues, and our models do capture patterns associated with particular venues by taking venue-specific features into account. The city-based model specialization captures new factors that may exist due to spatial locality of tip popularity patterns at the city level. We leave to future work a more thorough investigation of other spatial factors as well as other strategies to introduce them to the popularity prediction models.

## 6.8.2 Category-Based Model Specialization

Finally, we analyze the benefits of building specialized models for each venue category. Recall that Foursquare defines nine top level venue categories, namely, “Arts &

Table 6.9: Categorical Model Specialization: Macro-Average Results.

Scenarios	Macro-Average Recall							
	General Model				Specialized Model			
	OLS	SVM	SVR-rbf	SVR-linear	OLS	SVM	SVR-rbf	SVR-linear
Arts	<b>0.8010</b> ± 0.0132	<b>0.7865</b> ± 0.0139	<b>0.7959</b> ± 0.0153	<b>0.7848</b> ± 0.0371	<b>0.7832</b> ± 0.0108 ↓	<b>0.7937</b> ± 0.0124	<b>0.7896</b> ± 0.0112	<b>0.7795</b> ± 0.0351
Food	0.7884 ± 0.0193	0.7972 ± 0.0188	0.7853 ± 0.0196	0.7682 ± 0.0153	<b>0.8140</b> ± 0.0164 ↑	<b>0.8250</b> ± 0.0153 ↑	<b>0.8262</b> ± 0.0137 ↑	0.7627 ± 0.0096
Night	0.7921 ± 0.0102	<b>0.8048</b> ± 0.0079	<b>0.8036</b> ± 0.0081	0.7707 ± 0.0167	<b>0.8065</b> ± 0.0113 ↑	<b>0.8134</b> ± 0.0102	0.8008 ± 0.0086	0.7627 ± 0.0075
Shops	<b>0.8119</b> ± 0.0347	<b>0.8130</b> ± 0.0310	<b>0.8124</b> ± 0.0337	0.7042 ± 0.0359	<b>0.8187</b> ± 0.0331	<b>0.8156</b> ± 0.0353	<b>0.8092</b> ± 0.0355	0.7609 ± 0.0366 ↑
Scenarios	Macro-Average Precision							
	General Model				Specialized Model			
	OLS	SVM	SVR-rbf	SVR-linear	OLS	SVM	SVR-rbf	SVR-linear
Arts	0.5279 ± 0.0059	0.5237 ± 0.0058	0.5255 ± 0.0065	0.5627 ± 0.0026	0.5470 ± 0.0076 ↑	0.5497 ± 0.0089 ↑	0.5494 ± 0.0095 ↑	<b>0.5658</b> ± 0.0025 ↑
Food	0.5258 ± 0.0036	0.5288 ± 0.0048	0.5271 ± 0.0053	0.5234 ± 0.0034	0.5194 ± 0.0026 ↓	0.5186 ± 0.0027 ↓	0.5212 ± 0.0040 ↓	<b>0.5360</b> ± 0.0047 ↑
Night	0.5256 ± 0.0049	0.5275 ± 0.0050	0.5314 ± 0.0064	<b>0.5450</b> ± 0.0067	0.5292 ± 0.0032	0.5328 ± 0.0047	<b>0.5383</b> ± 0.0066	<b>0.5448</b> ± 0.0071
Shops	0.5232 ± 0.0042	0.5244 ± 0.0047	0.5246 ± 0.0054	<b>0.5285</b> ± 0.0094	0.5271 ± 0.0041	0.5261 ± 0.0044	0.5261 ± 0.0044	<b>0.5363</b> ± 0.0061

Entertainment”, “Colleges & Universities”, “Food”, “Great Outdoors”, “Nightlife Spots”, “Travel and Transport”, “Shops”, “Professional & Other Places” and “Residences”. We built specialized models for four selected categories, namely “Arts & Entertaining” (Arts), “Food”, “Shops & Services” (Shops), and “Nightlife Spots” (Night)<sup>1</sup>. As before, we compare each specialized model with the corresponding general model (considering all categories) on the same test set of tips posted at venues of the target category. Moreover, both models are learned with training sets of same size. Specifically, the training sets used to learn the models for the Arts, Food, Shops and Night scenarios consist of 1390, 2093, 1326, 1086 tips, respectively, whereas corresponding test sets include 5790, 49341, 17965, and 9685 tips.

Table 6.9 shows macro-average recall and precision results, along with corresponding 95% confidence intervals, for each model and method. Like in Table 6.8, results for the best methods (including statistical ties) for each category are shown in bold, and ↑ and ↓ signs are used to indicate gains and losses due to specialization.

Our first observation is that category-based model specialization does not bring as large and clear improvements over the general model as the city-based specialization. On one hand, we do observe some statistical improvements in macro-average recall and precision of up to 8.06% and 4.98%, respectively. Yet, statistical losses in the same metrics of up to 2.22% and 1.92% are also observed. Overall, we find that the results of the specialized models are only marginally different (if not tied with) those produced by the corresponding general models.

Such small differences are mainly due to the fact that the venue category is already somewhat explored by our (general) model as a feature. Indeed, as discussed in Section 6.6.2, venue category is one of the top 10 most important features for popularity prediction, implying that different patterns may exist across different categories. Yet,

<sup>1</sup>These categories were selected as they have the largest numbers of tips in our dataset.

since this feature is already part of the general model, specialization for each category does not bring as much new information to the model as the city-based specialization does (which, as discussed, introduces factors related to city-level spatial locality to the model). This is why we observed a tie between general and specialized models in various scenarios. The statistically significant differences (improvements and losses) observed in a few cases (e.g., OLS on Food and Arts categories) are caused by differences in the training set distributions used to build both general and specialized models. These differences in turn are an indirect result of the specialization, as we further explain next.

Recall that in our experiments, for a given category, the training sets used to learn both specialized and general models have the same number of tips. However, the training set of the general model contains tips from all categories. Thus, if we consider only tips of the target category, the number of tips in each class (and the class imbalance) may be different in both training sets.

Take, for example, the case of the Food category. Table 6.9 shows that, for OLS, SVM and SVR (with RBF kernel), the specialization does bring some improvements in macro-average recall but losses in macro-average precision. We manually investigated the tips in the training sets (some folds) used to build both models. Focusing only on tips in Food venues, the training set of the specialized model includes a proportionally much larger number of tips of high popularity than the training set of the general model. This favors the classification of tips in the high popularity class by the specialized model, which leads to improvements in recall for that particular class. However, as a side effect, the number of tips in the low popularity class that are incorrectly classified also increases, which leads to losses in precision (once again for the high popularity class, as it is smaller and more sensitive to changes).

For the Arts category, we observe an opposite trend: specialization leads to losses in macro-average recall but improvements in macro-average precision. Once again, this can be explained by the different numbers of tips in each class in both training sets, considering only venues in the Arts category. Compared to the training set of the general model, the training set of the specialized model includes a proportionally much larger number of tips of the low popularity class. This favors the classification of tips in that class, which hurts recall of the high popularity class but also improves its precision.

We note that different techniques (OLS, SVM, SVR) may be more or less robust to such differences in training set. We also emphasize that it was not expected beforehand that the category-based model specialization would not bring consistent improvements over the general model, despite the latter including venue category as a feature. The two prediction models exploit venue category very differently. In the general model,

this information is another dimension of the feature space considered. Given the high dimensionality of this space (125 features), using only tips of the target category to learn the model could help to reduce noise and significantly improve prediction accuracy. Yet, our results revealed that the specialization is not consistently beneficial and may also hurt prediction accuracy.

Thus, given such inconsistent performance, and considering that the differences, when significant, are small, we argue that category-based model specialization is not worthwhile because the main additional information, venue category, is already considered by the general model (though in a different way).

As a final observation, we also note that the gains of the more sophisticated SVM and SVR over OLS are not as large as observed for the city-based specialization. They are constrained to at most 3.77% and 3.44% for macro-averaged recall and precision, respectively, which limit their benefits over the simpler OLS, from a practical perspective. The limited benefits of SVM and SVR are probably due to the amounts of training examples available in the category-based scenarios, which unlike in some of the city-based scenarios, are large enough for OLS to produce reasonably accurate results.

## 6.9 Summary

In this chapter, we tackled the problem of predicting the popularity level of a tip in Foursquare. To that end, we investigated the use of various classification and regression-based strategies, notably SVM, SVR, and OLS, along with different sets of features to build prediction models.

This is a challenging problem which, in comparison with other types of tasks (e.g., assess the helpfulness of longer reviews, estimate the number of views of a video, or predict the ranking of a group of tips), has unique aspects and inherently different nature, and may depend on a non-trivial combination of various user, venue and content features. Nevertheless, some of our proposed models were able to produce good results (over 80%) in terms of recall, though precision results were limited in particular due to the severe imbalance across different popularity levels.

We note that, despite having covered a very extensive set of features, we may have left out other factors that also influence tip popularity, such as characteristics of the interface design, and other types of interaction (e.g., through search engines). In order to gather more evidence that support this insight and further analyze the complexity of our popularity prediction problem, we evaluated whether the values of the features

considered exhibit “locality”. In other words, we analyze whether the tips can be grouped by their similar patterns and eventually be correlated to a popularity category.

For this experiment, we considered the top-10 most discriminative features according to Information Gain which, as discussed in Section 6.6.2.4, produces prediction results as good as those produced by the complete set of 125 features. We defined a 10-dimensional space based on the selected features. The values of each feature were discretized into four intervals<sup>1</sup>, namely,  $[0; 0.25]$ ,  $(0.25; 0.5]$ ,  $(0.5; 0.75]$ , and  $(0.75; 1]$ . We then divided the 10-dimensional space into 1,048,576 *sectors*<sup>2</sup>, each one defined by a tuple of 10 (discretized) feature values (one of each feature).

Our first analysis focused on how tips are spread across such sectors. Considering only sectors with at least 2 tips<sup>3</sup>, we analyze how the popularity classes are spread across the selected sectors. For this analysis, we considered the true popularity of each tip measured 1 month after posting time. We observe that the majority of the selected sectors (83%) contains only low popularity tips, and the tips in those selected sectors account for over 59% of the total number of tips. This shows that tips from low popularity level have a very small locality in terms of the feature values, suggests that these tips do not have a clear “signature” (i.e., one or few sectors that contain most of them). In contrast, we find that the high popularity are spread across a smaller fraction of sectors (18%). However, the vast majority of those sectors (90%) contain a clear majority of tips of the low popularity level. That is, most sectors containing high popularity tips are indeed dominated by low popularity tips. We find that over 84% of the high popularity tips are in those sectors. These fractions reflect the severe class imbalance that challenges our solutions, and provide further evidence of the complexity of the tip popularity prediction problem.

Our second analysis focused on the accuracy of our predictions. We considered our best prediction model (OLS), running it into two scenarios: a ideal scenario that considered tips from the training set as the testing set (Figure 6.15) and a more realistic scenario in which the model was evaluated using another set of tips for testing set (Figure 6.16). In both cases, we considered predictions to 1 month ahead and performed at posting time ( $\varepsilon = 0$ , and  $\delta = 1$  month). We measured the accuracy rate (fraction of the number of correct classifications done by OLS) for each sector, and we correlate it with the dominant (true) popularity category in the sector.

If the model is capturing the most important aspects of the problem, we expect that it should have higher accuracy in sectors containing the majority of tips of the same

---

<sup>1</sup>Recall that features are already normalized between 0 and 1.

<sup>2</sup>We omitted sectors with less than two tips.

<sup>3</sup>We find 38% sectors with only one tip.

popularity category. Indeed, this is true for various sectors in both scenarios. Take for instance the sectors labeled as Region 1 in both Figures 6.15 and 6.16. These are sectors containing only tips from the same popularity category where the model obtained 100% of accuracy. Note that this happens also if we look at Region 3 (both figures), which consists of sectors containing only highly popular tips. However, there also sectors where the model behaves poorly as they have all tips from the low popularity category but zero accuracy rate (Region 2). When we ran the same experiment for the second scenario (different set of tips for the test set), we observed, in particular for the low popularity tips, a greater variation in the model accuracy when the sectors are composed by only tips from the same category. Interestingly, in the same scenario, if we focus on the high popularity tips, we observed a more clear trend of having a better model accuracy in sectors containing a larger fraction of tips of the same class.

These results illustrate the complexity of the problem we tackled. There seems to be other factors that were not explored in this dissertation or possible hidden factor interactions that also affect tip popularity. Identifying and characterizing such factor and interactions can lead to further improvements in prediction accuracy, motivating further investigations in this direction.

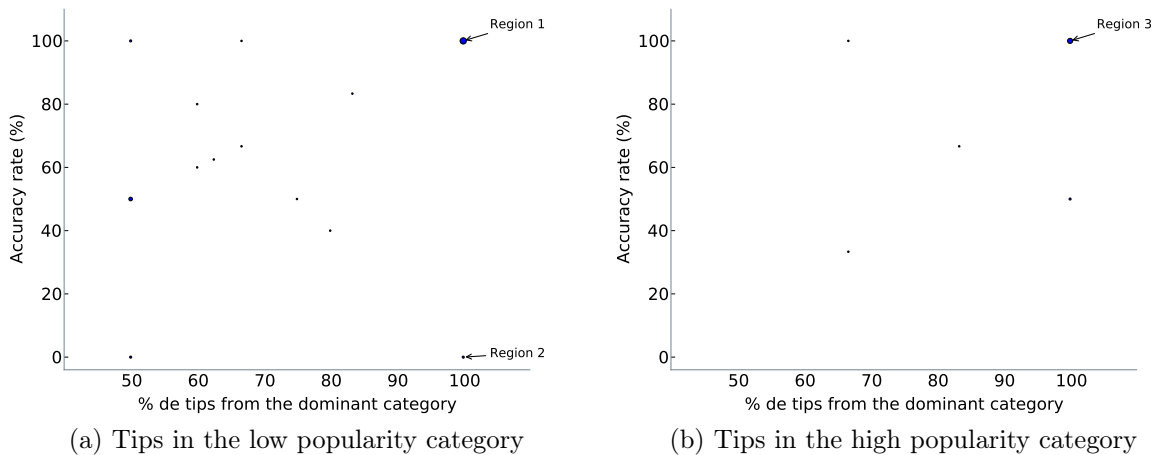


Figure 6.15: Model Accuracy in the Training Set (Each point is a Sector in the 10-Dimensional Space Defined by the Top-10 Features).

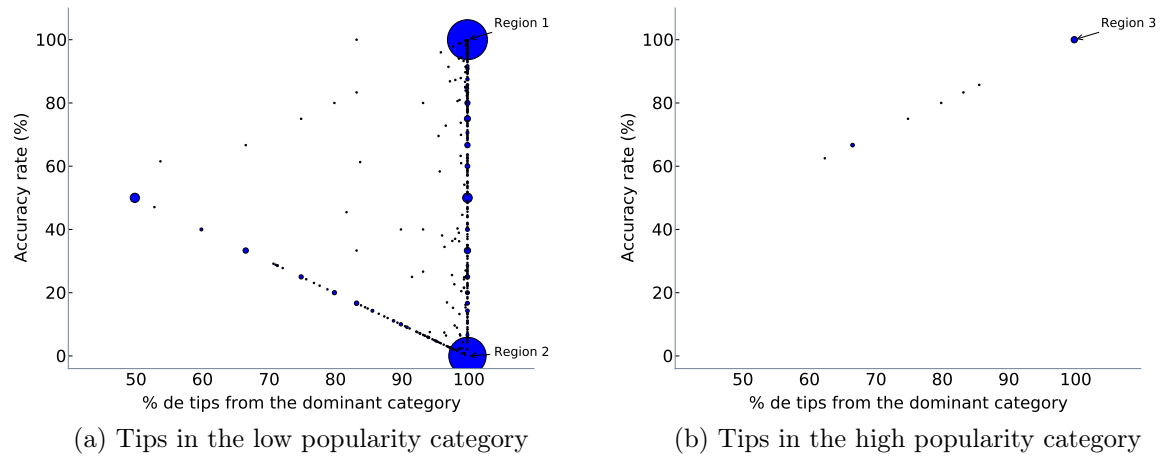


Figure 6.16: Model Accuracy in the Testing Set (Each Point is a Sector in the 10-Dimensional Space Defined by the Top-10 Features).





# Chapter 7

## Conclusions and Future Work

Online reviews have enabled customers to interact and share information and opinions about products and services they experience. With the diffusion of smartphones, reviews were expanded to the mobile environment, incorporating new forms, features, and addressing new challenges. Micro-reviews or tips, a more concise type of review (usually, up to 200 characters), have emerged from this environment. Tips often capture the immediate reaction from users, while the information is still fresh in the user's mind, and usually contain much more subjective and informal content.

As in longer review systems, the abundance of micro-reviews makes it hard for customers to find helpful reviews, especially on a mobile device. Thus, rather than helping users in their purchasing decision, these large volume of reviews can be make the user experience overwhelming and misleading. As the soon as a micro-review is posted, it has the potential to influence other customers on their purchase decisions via systems like Yelp and Foursquare. So, it is very important to understand how tips are explored by users in order to understand the consequences for the businesses. That can provide valuable inputs to learn a model for automatically predict the quality of a review, which can greatly benefit the future design of content filtering and recommendation methods.

Our dissertation presents a combination of (a) an extensive characterization of the main entities that may impact tip popularity, and analysis of popularity dynamics over time, (b) identification of a rich set of features related to tip popularity, and (c) investigation of solutions to two popularity prediction tasks using the set of features. Next, we summarize our main conclusions, and present some possible directions for future work. We then wrap up with a list of publications derived, directly or indirectly, from this dissertation.

## 7.1 Main Conclusions

The main focus of our dissertation is to investigate the problem of predicting the popularity of micro-reviews. To support our study, we collected two datasets from Foursquare, consisting of over 10 million of tips posted by 13 million users. We presented a large-scale characterization of the three main entities related to Foursquare system that may impact tip popularity – the user who posted the tip, the venue where it was posted, and its textual content. Our analyses also uncovered four different user behavior profiles and identified the presence of spamming activity. Towards better understanding user interactions, notably the presence of influential users, we also investigated methods to automatically infer a user’s influence level on Foursquare. Moreover, we studied how the popularity of a tip evolves over time by performing an extensive analysis on the popularity dynamics of Foursquare.

After performing the characterization, we were able to identify a rich set of over 120 features related to the user, venue and tip content that may impact tip popularity. We then investigated the potential benefits from exploiting these aspects to estimate the popularity of a tip in the future. To that end, we formulated the prediction problem as two different prediction tasks.

The first task addressed the prediction problem as a ranking task aiming at ranking a group of tips based on their predicted popularity at a given future time. Towards addressing this task, we first evaluated the stability of the tip popularity ranking over time, assessing to which extent the current popularity ranking of a set of tips can be used to predict their popularity ranking at a future time. Overall, the ranking is stable corroborating the characterization results, but we observed opportunities for improvement, exploiting a multidimensional set of predictors. Our results showed that the use of the richer set of features can indeed improve the prediction accuracy, provided that enough data is available to train the regression model. Moreover, our experimental results showed that the use of only four features (total number of likes received by the tips previously posted by the user, total number of check-ins at the venues, the type of the Foursquare user, and the current popularity of the tip) produces results that are as good as when all features are used as predictors.

The other prediction task we tackled relates to the problem of predicting the popularity level of a single tip. This turned out to be a more challenging task, as it focuses on absolute values of popularity, as opposed to relative measures (i.e., ranking). Moreover, we focused first on predictions at tip posting time, when the only information about the tip’s textual content and historical patterns related to the user and the venue associated with it can be exploited as predictor variables. In particular, unlike

the vast majority of previous content popularity prediction efforts, no early measures of popularity of the tip is considered. Nevertheless, some of our proposed models were able to produce good results (over 80% in terms of recall). We also observed that the top-10 most discriminative features, which include features that capture the prior popularity of the user who posted the tip and the venue where it was posted as well as characteristics of the user's social network, produce results that are as good as when all 125 features considered are used as predictors. We found significant improvements for all prediction methods if we relax the restriction of performing predictions at posting time. In other words, by monitoring the tip for a (short) time, we are able to gather information about its early popularity, which improves prediction accuracy. Moreover, although state-of-the-art prediction methods that use early popularity measures as the only predictors do perform reasonably well in such scenarios, our models are more robust, as they can be applied to any tip, at or after posting time (unlike the other methods), besides producing much higher recall for the high popularity class. Finally, we found that model specialization does bring some (limited) improvements if performed at the city-level, whereas category-based specialization does not bring clear and consistent gains.

Predicting the popularity of micro-reviews (tips in particular) is a challenging problem which, in comparison with previous related efforts, has unique aspects and inherently different characteristics, and may depend on a non-trivial combination of various user, venue and content features. We expect that the knowledge derived from the present effort may bring valuable insights into the design of more cost-effective automatic tip filtering and recommendation strategies.

## 7.2 Directions for Future Work

We envision several directions into which our work can be extended in the future.

One possible direction consisting in analyzing and investigation the introduction of new features to the prediction models. For example, we have explored the geographical location of the venues to build specialized models, finding some improvements in the prediction accuracy. We believe that there is room for other spatial data analyses. For example, the model could have a feature that captures the correlation between the tip geographic location and the geographical location of previous likes received by the tip's author. Other factors that were not explored in this dissertation and may also influence tip popularity, such as the characteristics of the interface design and referrals from other external source (e.g., search engines, impact of newspaper articles,

other recommendation sites) should also be investigated towards improving the prediction accuracy. Moreover, based on our experimental results, we observed that textual features used for longer reviews were not effective to improve the accuracy of our prediction models. We believe that a study of new textual features, more specific for short and informal texts is a promising approach to develop an understanding about which information attracts more people.

In this study, we focused on tips written in English. A possible direction to be explored is to investigate the tips written in other languages which would allow us to analyze the patterns for different countries or cultures. That would require the use of specific textual tools for each language or the analysis of these tips using only the other features (i.e., user and venue features), since the textual features used in our models were not as relevant for the popularity prediction tasks. It is also interesting to evaluate if there are invariants in the text of the most popular tips.

Another group of possible directions is related to how we model the problem. In this work, we observed that most of our features were non-stationary features (i.e., features defined by their varying statistical properties with time) and we explored them as aggregate numbers (e.g., total, average, median). One extension of this work is to consider the evolution of these metrics over time during the monitoring time. This is similar to approaches that explore time series for predicting queries in a search engine [Radinsky et al., 2013] or trends in financial market [Ruiz et al., 2012]. Exploring such models, we are also able to capture seasonal or trend components of the prediction associated with events (e.g., Christmas or Thanksgiving holidays) that were not accounted when we use models based on aggregate metrics. Other interesting investigations are studying other ways to represent historical data and the possibility of a specific model for each type of prediction (e.g., for the next hour, day, week or to predict the number of likes received in a single day). Another way to formulate our popularity prediction problem is to build multivariate models that takes into account several dependent variables simultaneously. A multivariate study may show other variables that are related to the tip popularity. For example, the total number of likes received by a tip may be not the only important variable related to popularity, an interesting feature to be investigated is the mean number of likes received by the user's tips.

We also envision further investigations on technical aspects of the prediction techniques to help improving our predictions. Recall that to cope with the effects of class imbalance in the training data, we used the under-sampling technique. However, in some scenarios, this strategy is not ideal, since it reduces the size of training dataset. Thus, there are other approaches that should be investigated, such as over-

sampling (e.g., SMOTE [Chawla et al., 2002]), boosting [Seiffert et al., 2007] and bagging [Breiman, 1996] techniques. Moreover, active learning approaches [Harpale and Yang, 2008] can also be used to identify the most informative set of training instances. Even though we use state-of-the-art methods, other classification and ranking techniques could be also investigated.

Another possible direction for future developments consists of exploring our prediction models on various applications and services including recommendation/filtering of tips. Systems can make use of such predictions to instantly identify reviews that are expected to be helpful to users, and adjust their presentation efficiently, ultimately improving the user experience. Moreover, both reviews systems and businesses owners would like to be able to predict if a tip have the potential to become popular. For example, such information can be used to estimate the revenue of promoted tips ahead of time as well as to provide a quickly feedback about the used marketing strategy. An evaluation with real users is also important to better define the popularity measure and understanding the mapping between helpfulness and popularity with Foursquare users as done for Amazon reviews in Liu et al. [2007]. This evaluation can be performed as mechanical turk. Moreover, a deeper analysis of particular categories or subcategories of venues tips can be used to understand the impact of public services on the population from a given city. For example, monitoring the citizens reviews about buses stations give an idea of what the citizens expect from the provided service. This analysis can also point problems with the service or underutilization of the service for part of the users. Moreover, since Foursquare is a world-wide social network, this analysis can also be used to compare the most common problems faced by different countries (e.g., developed vs. undeveloped countries).

Finally, an interesting direction would be to evaluate our prediction models in other micro-reviews systems (e.g. Yelp). Such evaluation would require additional work to map our proposed features to the new system as well as investigating new features that are specific to the application. It would be interesting to investigate whether the same features would have a similar explanatory power in the new system as they have in Foursquare, and whether other features emerge as relevant to the popularity prediction in those systems. Furthermore, since our data was collected in 2011, it will be interesting if we can assess the impact on the tip popularity of the current Foursquare system. From 2015, Foursquare app was divided in two apps, separating the check-in functionality from the recommendation tool. It will be interesting to evaluate how the popularity of a tip is evolving in the new app comparing with our results.

## 7.3 Publications

The main results of this dissertation generated the following publications:

- [Vasconcelos et al., 2012b] published in ACM 5th ACM International Conference on Web Search and Data Mining (WSDM'12).
- [Vasconcelos et al., 2012a] published in the XXX Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'12).
- [Vasconcelos et al., 2014b] published in ACM 29th Symposium On Applied Computing (SAC'14).
- [Vasconcelos et al., 2014c] published in ACM Conference on Online Social Networks (COSN'14)
- [Vasconcelos et al., 2014a] submitted to Elsevier Information Sciences (2nd round of review).

During the development of this dissertation, we were also involved in other studies that are indirectly related to its topic. They generated the following publications:

- [Moraes et al., 2013b] published in 19th Brazilian Symposium on Multimedia and the Web (WebMedia'13). In this work, we evaluated the effectiveness of four polarity classification strategies on subsets of our Foursquare dataset.
- [Moraes et al., 2013a] published in 5th International Conference on Social Informatics (SocInfo'13). In this work, we compared the same four polarity detection methods with a hybrid approach that combines all techniques using stacking.
- [Pontes et al., 2012b] published in 4th International Workshop on Location-Based Social Networks (LBSN'12). In this work, we investigated how much information about a user can be inferred from her tips, likes and mayorships.
- [Pontes et al., 2012a] published in the International Workshop on Privacy in Social Data (PinSoDa'12). In this work, we performed a large-scale inference study in three of the currently most popular social networks: Foursquare, Google+ and Twitter.

# Bibliography

- Abbasi, M. A. and Liu, H. (2013). Measuring User Credibility in Social Media. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP)*.
- Adal, S., Liu, T., and Magdon-Ismail, M. (2012). An Analysis of Optimal Link Bombs. *Theoretical Computer Science*, 437:1--20.
- Adamic, L., Zhang, J., Bakshy, E., and Ackerman, M. (2008). Knowledge Sharing and Yahoo Answers: Everyone Know Something. In *Proceedings of the 17th International World Wide Web Conference (WWW)*.
- Agarwal, N., Liu, H., Tang, L., and Yu, P. S. (2008). Identifying the Influential Bloggers in a Community. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*.
- Aggarwal, A., Almeida, J., and Kumuraguru, P. (2013). Detection of Spam Tipping Behaviour on Foursquare. In *Proceedings of the 2nd International Workshop on Mining Social Network Dynamics (MSND)*.
- Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding High-Quality Content in Social Media. In *Proceedings of the First International Conference on Web Search and Data Mining (WSDM)*.
- Akoglu, L., Chandy, R., and Faloutsos, C. (2013). Opinion Fraud Detection in Online Reviews by Network Effects. In *Proceedings of the 7th International AAAI Conference on Weblogs (ICWSM)*.
- Allen, A. (1990). *Probability, Statistics, and Queueing Theory with Computer Science Applications*. Academic Press Professional, Inc., San Diego, CA, USA. ISBN 0-12-051051-0.

- Alwee, R., Shamsuddin, S., and Sallehuddin, R. (2013). Hybrid Support Vector Regression and Autoregressive Integrated Moving Average Models Improved by Particle Swarm Optimization for Property Crime Rates Forecasting with Economic Indicators. *The Scientific World Journal*, 2013(951475).
- Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Discovering Value from Community Activity on Focused Question Answering Sites: a Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- Ante, S. (2009). Amazon: Turning Consumer Opinions into Gold. [http://www.businessweek.com/magazine/content/09\\_43/b4152047039565.htm](http://www.businessweek.com/magazine/content/09_43/b4152047039565.htm).
- Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Bakshy, E., Karrer, B., and Adamic, L. A. (2009). Social Influence and the Diffusion of User-Created Content. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC)*.
- Bandari, R., Asur, S., and Huberman, B. (2012). The Pulse of News in Social Media: Forecasting Popularity. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM)*.
- Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Barbierri, C. (2011). Foursquare Reveals New “Promoted” Check-in for Super Bowl Sunday. <http://venturebeat.com/2011/02/03/foursquare-super-bowl/>.
- Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Ross, K. (2009). Video Interactions in Online Video Social Networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(4):1–25.
- Benevenuto, F., Rodrigues, T., Veloso, A., Almeida, J., Gonçalves, M., and Almeida, V. (2012). Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems. *IEEE Transactions on Systems, Man and Cybernetics - Part B*, PP(99):1–14.



- Berjani, B. and Strufe, T. (2011). A Recommendation System for Spots in Location-Based Online Social Networks. In *Proceedings of the 4th Workshop on Social Network Systems (SNS)*.
- Bermingham, A. and Smeaton, A. F. (2010). Classifying Sentiment in Microblogs: is Brevity an Advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*.
- Borghol, Y., Ardon, S., Carlsson, N., Eager, D., and Mahanti, A. (2012). An Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- Breiman, L. (1996). Bagging predictors. *Journal Machine Learning*, 24(2):123--140.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5--32.
- Brodersen, A., Scellato, S., and Watternhofer, M. (2012). YouTube Around the World: Geographic Popularity of Videos. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*.
- Castillo, C., Mendoza, M., and Poblete, B. (2011). Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*.
- Catalyst Marketers Blog (2010). Catalyst Marketers Blog. <http://www.catalystmarketers.com/foursquare-spam>.
- Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of 4th International AAAI Conference on Weblogs and Social (ICWSM)*.
- Cha, M., Mislove, A., and Gummadi, K. P. (2009). A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*.
- Chang, C. and Lin, C. (2001). LIBSVM: a Library for Support Vector Machines.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321--357.

- Chen, B.-C., Guo, J., Tseng, B., and Yang, J. (2011). User Reputation in a Comment Rating Environment. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- Chen, P.-Y., yi Wu, S., and Yoon, J. (2004). The Impact of Online Recommendations and Consumer Feedback on Sales. In *Proceedings of the International Conference on Information Systems (ICIS)*.
- Chen, Y. and Xie, J. (2008). Online Consumer Review: Word-of-Mouth as A New Element of Marketing Communication Mix. *Management Science*, 54(3):477–491.
- Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can Cascades Be Predicted? In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*.
- Cheng, Z., Caverlee, J., Kamath, K., and Lee, K. (2011a). Toward Traffic-Driven Location-Based Web Search. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*.
- Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011b). Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*.
- Chevalier, J. A. and Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, 43(3):345–354.
- Cho, E., Myers, S., and Leskovec, J. (2011). Friendship and Mobility: User Movement In Location-Based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- Coleman, M. and Liau, T. (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284.
- Costa, H., Benevenuto, F., and de Campos Merschmann, L. H. (2013). Detecting Tip Spam in Location-based Social Networks. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC)*.
- Crane, R. and Sornette, D. (2008). Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System. In *Proceedings of the National Academy of Sciences (PNAS)*.

- Dalip, D., Gonçalves, M., Cristo, M., and Calado, P. (2011). Automatic Assessment of Document Quality in Web Collaborative Digital Libraries. *J. Data and Information Quality*, 2(3):14:1--14:30.
- Dalip, D., Lima, H., Gonçalves, M., Cristo, M., and Calado, P. (2014). Quality Assessment of Collaborative Content With Minimal Information. In *Joint Conference on Digital Libraries and the Theory and Practice of Digital Libraries*.
- Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2013). Exploiting User Feedback to Learn to Rank Answers in Q&#38;a Forums: A Case Study with Stack Overflow. In *Proceedings of the 36th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.
- Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., and Lee, L. (2009). How Opinions are Received by Online Communities: a Case Study on Amazon.com Helpfulness Votes. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1--38.
- Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vladimir, V. (1997). Support Vector Regression Machines. In *Advances in Neural Information Processing Systems 9*.
- Du, Y., Shi, Y., and Zhao, X. (2007). Using Spam Farm to Boost PageRank. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, pages 417--422.
- Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing Top K Lists. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

- Fellbaum, C. (1998). *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.
- Figueiredo, F., Almeida, J., Matsubara, Y., Ribeiro, B., and Faloutsos, C. (2014a). Revisit Behavior in Social Media: The Phoenix-R Model and Discoveries. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML/PKDD)*.
- Figueiredo, F., Gonçalves, M., and Almeida, J. (2014b). Improving the Effectiveness of Content Popularity Prediction Methods using Time Series Trends. In *ECML/PKDD Predictive Analytics Challenge*.
- Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221 – 233.
- Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., and Treinen, M. (2001). What Makes Web Sites Credible?: a Report on a Large Quantitative Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*.
- Foursquare (2011). Foursquare House Rules. <http://support.foursquare.com/entries/386768-foursquare-house-rules>.
- Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). An Overview of Ensemble Methods for Binary Classifiers in Multi-Class Problems: Experimental Study on One-vs-One and One-vs-All Schemes. *Pattern Recognition*, 44(8):1761--1776.
- Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. (2010). Detecting and Characterizing Social Spam Campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC)*.
- Georgiev, P., Noulas, A., and Mascolo, C. (2014). The Call of the Crowd: Event Participation in Location-based Social Services. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*.
- Ghose, A. and Ipeirotis, P. G. (2007). Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews. In *Proceedings of the 9th International Conference on Electronic Commerce (ICEC)*.

- Gionis, A., Lappas, T., Pelechrinis, K., and Terzi, E. (2014). Customized Tour Recommendations in Urban Areas. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Gladwell, M. (2002). *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books.
- Gonçalves, P., Araujo, M., Benevenuto, F., and Cha, M. (2013). Comparing and Combining Sentiment Analysis Methods. In *Proceedings of ACM Conference on Online Social Networks (COSN)*.
- Grier, C., Thomas, K., Paxson, V., and Zhang, M. (2010). @Spam: the Underground on 140 Characters or Less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CSS)*.
- Grinter, R. and Eldridge, M. (2003). Wan2Tlk?: Everyday Text Messaging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, CHI '03.
- Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information Diffusion Through Blogspace. In *Proceedings of the 13th International World Wide Web Conference (WWW)*.
- Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill, New York.
- Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004). Combating Web Spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data bases (VLDB)*.
- Harpale, A. S. and Yang, Y. (2008). Personalized Active Learning for Collaborative Filtering. In *Proceedings of the 31st International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.
- Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society (Applied Statistics)*, 28(1):100--108.
- Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *Proceedings of the 11th International World Wide Web Conference (WWW)*.
- He, H. and Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

- Henzinger, M. R., Heydon, A., Mitzenmacher, M., and Najork, M. (1999). Measuring Index Quality using Random Walks on the Web. In *Proceedings of the 8th International World Wide Web Conference (WWW)*.
- Hong, L., Dan, O., and Davison, B. D. (2011). Predicting Popular Messages in Twitter. In *Proceedings of the 20th International World Wide Web Conference (WWW)*.
- Hong, Y., Lu, J., Yao, J., Zhu, Q., and Zhou, G. (2012). What Reviews are Satisfactory: Novel Features for Automatic Helpfulness Voting. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.
- Hovland, C. I. and Weiss, W. (1951). The Influence of Source Credibility on Communication Effectiveness. *The Public Opinion Quarterly*, 15(4):635--650.
- Howley, T., Madden, M. G., O'Connell, M.-L., and Ryder, A. G. (2006). The Effect of Principal Component Analysis on Machine Learning Accuracy with High-Dimensional Spectral Data. *Knowledge-Based Systems*, 19(5):363--370.
- Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking Comments on the Social Web. In *Proceedings of the International Conference on Computational Science and Engineering (CSE)*.
- Hua, S. and Sun, Z. (2001). Support Vector Machine Approach for Protein Subcellular Localization Prediction. *Oxford Journal of Bioinformatics*, 17(8):721--728.
- Inquiry, L. and Count, W. (2007). Liwc2007 output variable information. <http://www.liwc.net/descriptiontable1.php>.
- Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley.
- Janecek, A., Gansterer, W. N., Demel, M., and Ecker, G. (2008). On the Relationship Between Feature Selection and Classification Accuracy. *Journal of Machine Learning Research - Proceedings Track*, 4:90--105.
- Jindal, N. and Liu, B. (2008). Opinion Spam and Analysis. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*.
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*.

- Katz, E. (1957). The Two-Step Flow of Communication: An Up-to-Date Report of an Hypothesis. *Public Opinion Quarterly*, 21(1):61–78.
- Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*. A Charles Griffin Title, 5 edition.
- Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically Assessing Review Helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys (CSUR)*, 31(4es). ISSN 0360-0300.
- Konagurthu, A. and Collier, J. (2013). An Information Measure for Comparing Top k Lists. *CoRR*, abs/1310.0110.
- Korfiatis, N., García-Bariocanal, E., and Sánchez-Alonso, S. (2012). Evaluating Content Quality and Helpfulness of Online Product Reviews: The Interplay of Review Helpfulness vs. Review Content. *Electronic Commerce Research and Applications*, 11(3):205--217.
- Krapivsky, P. L., Redner, S., and Leyvraz, F. (2000). Connectivity of Growing Random Networks. *Physical Review Letters*, 85(21):4629--4632.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International World Wide Web Conference (WWW)*.
- Lappas, T. (2012). Fake Reviews: The Malicious Perspective. In *Natural Language Processing and Information Systems*, volume 7337 of *Lecture Notes in Computer Science*, pages 23–34. Springer Berlin Heidelberg.
- Lee, S. and Choeh, J. Y. (2014). Predicting the Helpfulness of Online Reviews Using Multilayer Perceptron Neural Networks. *Expert Systems with Applications: An International Journal*, 41(6):3041--3046.
- Lerman, K. and Hogg, T. (2010). Using a Model of Social Dynamics to Predict Popularity of News. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*.
- Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The Dynamics of Viral Marketing. *ACM Transactions on the Web (TWEB)*, 1(1).

- Li, B., Jin, T., Lyu, M., King, I., and Mak, B. (2012). Analyzing and Predicting Question Quality in Community Question Answering Services. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*.
- Li, F., Huang, M., Yang, Y., and Zhu, X. (2011). Learning to Identify Review Spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*.
- Li, N. and Chen, G. (2009a). Analysis of a Location-based Network. In *Proceedings of the International Conference on Computational Science and Engineering (CSE)*.
- Li, N. and Chen, G. (2009b). Multi-layer Friendship Modeling of Location-based Mobile Social Networks. In *Proceedings of the 6th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*.
- Li, X. and Hitt, L. M. (2008). Self-Selection and Information Role of Online Product Reviews. *Information Systems Research*, 19(4):456--474.
- Li, Y., Steiner, M., Wang, L., Zhang, Z., and Bao, J. (2013). Exploring Venue Popularity in Foursquare. In *Proceedings of the 5th IEEE International Workshop on Network Science for Communication Networks (NetSciCom)*.
- Lin, Y., Zhu, T., Wang, X., Zhang, J., and Zhou, A. (2014). Towards Online Review Spam Detection. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*.
- Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. (2011). I'm the Mayor of my House: Examining why People use Foursquare - a Social-Driven Location Sharing Application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*.
- Liu, B. (2010). Sentiment Analysis and Subjectivity. In Indurkha, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group.
- Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.



- Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory Undersampling for Class-Imbalance Learning. *Transactions on Systems, Man and Cybernetics - Part B*, 39(2).
- Liu, Y., Huang, X., An, A., and Yu, X. (2008). Modeling and Predicting The Helpfulness of Online Reviews. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*.
- Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. (2010). Exploiting Social Context for Review Quality Prediction. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*.
- Ma, Y. and Li, F. (2012). Detecting Review Spam: Challenges and Opportunities. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*.
- Manevitz, L. M. and Yousef, M. (2002). One-class SVMs for Document Classification. *Journal of Machine Learning Research*, 2:139–154.
- Martin, L. and Pu, P. (2014). Prediction of Helpful Reviews Using Emotions Extraction. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.
- Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2012). Rise and Fall Patterns of Information Diffusion: Model and Implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
- McLaughlin, H. G. (1969). SMOG Grading - A New Readability Formula. *Journal of Reading*, 12(8):639–646.
- Moghaddam, S., Jamali, M., and Ester, M. (2012). ETF: Extended Tensor Factorization Model for Personalizing Prediction of Review Helpfulness. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Momeni, E., Cardie, C., and Ott, M. (2013). Properties, Prediction, and Prevalence of Useful User-generated Comments for Descriptive Annotation of Social Media Objects. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM)*.
- Moraes, F., Vasconcelos, M., Prado, P., Almeida, J., and Gonçalves, M. (2013a). Polarity Analysis of Micro Reviews in Foursquare. In *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia)*.

- Moraes, F., Vasconcelos, M., Prado, P., Dalip, D., Almeida, J., and Gonçalves, M. (2013b). Polarity Detection of Foursquare Tips. In *Social Informatics*, volume 8238 of *Lecture Notes in Computer Science*, pages 153–162. Springer International Publishing.
- Ngo-Ye, T. L. and Sinha, A. P. (2012). Analyzing Online Review Helpfulness Using a Regression Relief-Enhanced Text Mining Method. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):10:1–10:20.
- Nguyen, T.-S., Lauw, H. W., and Tsaparas, P. (2013). Using Micro-reviews to Select an Efficient Set of Reviews. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (CIKM)*.
- Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS one*, 7(5):e37027.
- Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011a). An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011b). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proceedings of 3rd Workshop Social Mobile Web (SMW)*.
- O’Brien, R. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity: International Journal of Methodology*, 41(5):673–690.
- O’Mahony, M. and Smyth, B. (2009). Learning to Recommend Helpful Hotel Reviews. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys)*.
- O’Mahony, M. P. and Smyth, B. (2010). Using Readability Tests to Predict Helpful Product Reviews. In *Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO)*.
- Page, L., Brin, R. M., and Winograd, T. (1998). The Pagerank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project.
- Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Park, J., Cha, M., Kim, H., and Jeong, J. (2012). Managing Bad News in Social Media: A Case Study on Domino's Pizza Crisis. In *Proceedings of 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Pelleg, D. and Moore, A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*.
- Pinto, H., Almeida, J. M., and Gonçalves, M. A. (2013). Using Early View Patterns to Predict the Popularity of YouTube Videos. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., and Almeida, V. (2012a). Beware of What You Share: Inferring Home Location in Social Networks. In *International Workshop on Privacy in Social Data (PinSoDa)*.
- Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., and Almeida, V. (2012b). We Know Where you Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of 4th International Workshop on Location-Based Social Networks (LBSN)*.
- Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. (2011). In the Mood for Being Influential on Twitter. In *Proceedings of the 3th IEEE International Conference on Social Computing (SOCIALCOM)*.
- Radinsky, K., Svore, K. M., Dumais, S. T., Shokouhi, M., Teevan, J., Bocharov, A., and Horvitz, E. (2013). Behavioral Dynamics on the Web: Learning, Modeling, and Prediction. *ACM Transactions on Information Systems (TOIS)*, 31(3):16:1--16:37.
- Ressler, S. (1993). *Perspectives on Electronic Publishing - Standards, Solutions, and More*. Prentice Hall.
- Romero, D., Galuba, W., Asur, S., and Huberman, B. (2011). Influence and Passivity in Social Media. In *Proceedings of the 20th International World Wide Web Conference (WWW)*.
- Rossi, L. and Musolesi, M. (2014). It's the Way you Check-in: Identifying Users in Location-Based Social Networks. In *Proceedings of ACM Conference on Online Social Networks (COSN)*.
- Rubin, V. L. and Liddy, E. D. (2006). Assessing Credibility of Weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

- Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. (2012). Correlating Financial Time Series with Micro-blogging Activity. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Sarah Best (2012). The Building Blocks of a Great Foursquare List. [http://www.mightybytes.com/blog/entry/the\\_building\\_blocks\\_of\\_a\\_great\\_foursquare\\_list/](http://www.mightybytes.com/blog/entry/the_building_blocks_of_a_great_foursquare_list/).
- Scellato, S. and Mascolo, C. (2011). Measuring User Activity on an Online Location-based Social Network. In *Proceedings of 3rd International Workshop on Network Science for Communication Networks (NetSciCom)*.
- Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. (2010). Distance Matters: Geo-social Metrics for Online Social Networks. In *Proceedings of the 3rd Workshop on Online Social Networks (WOSN)*.
- Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011a). Socio-spatial Properties of Online Location-based Social Networks. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Scellato, S., Noulas, A., and Mascolo, C. (2011b). Exploiting Place Features in Link Prediction on Location-based Social Networks. In *Proceedings of 17th ACM Conference on Knowledge Discovery and Data Mining (KDD)*.
- Scherer, K. (2005). What are Emotions? And How Can They be Measured? *Social Science Information*, 44(4):695–729.
- Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., and Napolitano, A. (2007). Mining Data with Rare Events: A Case Study. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*.
- Senter, R. and Smith, E. A. (1967). Automated Readability Index. Technical report AMRL-TR-6620, Wright-Patterson Air Force Base.
- Seward, Z. (2011). Checking In to the Snowpocalypse. <http://blogs.wsj.com/digits/2011/05/19/checking-into-snowpocalypse/>.
- Shi, J. and Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.
- Siersdorfer, S., Chelaru, S., Nejdl, W., and San Pedro, J. (2010). How Useful are Your Comments?: Analyzing and Predicting YouTube Comments and Comment Ratings. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*.

- Siersdorfer, S., Chelaru, S., Pedro, J. S., Altingovde, I. S., and Nejd, W. (2014). Analyzing and Mining Comments and Comment Ratings on the Social Web. *ACM Transactions on the Web (TWEB)*, 8(3).
- Silva, T., de Melo, P., Almeida, J., Salles, J., and Loureiro, A. (2014). Revealing the City that We Cannot See. In *Transactions on Internet Technology*.
- Smith, W. (2013). Brands and the New View Of Social Influence. <http://www.brandingstrategyinsider.com/2013/06/brands-and-the-new-view-of-social-influence.html>.
- Stanford NLP Group (2012). Stanford Part-Of-Speech Tagger. <http://nlp.stanford.edu/software/tagger.shtml>.
- Stanford NLP Group (2013). Stanford Named Entity Recognizer. <http://nlp.stanford.edu/software/CRF-NER.shtml>.
- Stevens, J. (2002). *Applied Multivariate Statistics for The Social Sciences*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA. ISBN 0-898-59568-1.
- Suh, B., Hong, L., Pirolli, P., and Chi, E. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Proceedings of the 2nd IEEE International Conference on Social Computing (SOCIALCOM)*.
- Szabo, G. and Huberman, B. A. (2010). Predicting the Popularity of Online Content. *Communications of the ACM*, 53(8):80--88.
- Tang, J., Gao, H., Hu, X., and Liu, H. (2013). Context-aware Review Helpfulness Rating Prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*.
- Tang, Y., Zhang, Y.-Q., Chawla, N., and Krasser, S. (2009). SVMs Modeling for Highly Imbalanced Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):281–288.
- Tatar, A., Antoniadis, P., de Amorim, M. D., and Fdida, S. (2014). From Popularity Prediction to Ranking Online News. *Social Network Analysis and Mining*, 4(1).
- Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M. D., and Fdida, S. (2011). Predicting the Popularity of Online Articles Based on User Comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS)*.

- Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24--54.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in Short Strength Detection Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544--2558.
- Thomas, K., Grier, C., Ma, J., Paxson, V., and Song, D. (2011). Design and Evaluation of a Real-Time URL Spam Filtering Service. In *IEEE Symposium on Security and Privacy (S&P)*.
- Thurlow and Brown (2003). Generation Txt? The Sociolinguistics of Young People's Text-Messaging. *Discourse Analysis Online*.
- T.Ngo-Ye and Sinha, A. (2014). The Influence of Reviewer Engagement Characteristics on Online Review Helpfulness: A Text Regression Model. *Decision Support Systems*, 61:47--58.
- Tsur, O. and Rappoport, A. (2009). RevRank: a Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews. In *Proceedings of the 3th International Conference on Weblogs and Social Media (ICWSM)*.
- Valente, T. (1995). *Network Models of the Diffusion of Innovations*. Hampton Press, Cresskill, NJ.
- van Zwol, R. (2007). Flickr: Who is Looking? In *ACM International Conference on Web Intelligence (WIS)*.
- Vasconcelos, M., Almeida, J., and Gonçalves, M. (2014a). Predicting the Popularity of Micro-Reviews: A Foursquare Case Study. Elsevier Information Sciences (2nd review round).
- Vasconcelos, M., Almeida, J., and Gonçalves, M. (2014b). What Makes your Opinion Popular? Predicting the Popularity of Micro-Reviews in Foursquare. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC)*.
- Vasconcelos, M., Almeida, J., Gonçalves, M., Souza, D., and Gomes, G. (2014c). Popularity Dynamics of Foursquare Micro Reviews. In *Proceedings of the ACM Conference on Online Social Networks (COSN)*.

- Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012a). Caracterizacao e Influencia do Uso de Tips e Dones no Foursquare. In *Proceedings of the 30th Brazilian Symposium on Computer Networks and Distributed Systems (SBRC)*.
- Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012b). Tips, Dones and Todos: Uncovering User Profiles in Foursquare. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Wagner, C., Rowe, M., Strohmaier, M., and Alani, H. (2012). Ignorance isn't Bliss: An Empirical Analysis of Attention Patterns in Online Communities. In *Proceedings of the 4th IEEE International Conference on Social Computing (SOCIALCOM)*.
- Walther, J., Carr, C., Choi, S., DeAndrea, D., Kim, J., Tong, S., and Heide, B. V. D. (2010). Interaction of Interpersonal, Peer, and Media Influence Sources Online. *A Networked Self: Identity, Community, and Culture on Social Network Sites*, pages 17–38.
- Wang, Z., Zhang, D., Zhou, X., Yang, D., Yu, Z., and Yu, Z. (2014). Discovering and Profiling Overlapping Communities in Location-Based Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4):499–509.
- Watts, D. and Dodds, P. (2007). Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4):441–458.
- Weerkamp, W. and de Rijke, M. (2012). Credibility-inspired Ranking for Blog Post Retrieval. *Information Retrieval*, 15(3-4):243–277.
- Weka Machine Learning Project (2012). Weka. <http://www.cs.waikato.ac.nz/~ml/weka>.
- Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). TwitterRank: Finding Topic-Sensitive Influential Twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM)*.
- Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011a). Who Says What to Whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*.
- Wu, S., Tan, C., Kleinberg, J. M., and Macy, M. W. (2011b). Does Bad News Go Away Faster? In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

- Yang, H., Zhou, Y., and Liu, H. (2010). Chaos Optimization SVR Algorithm with Application in Prediction of Regional Logistics Demand. In *Proceedings of the First International Conference on Advances in Swarm Intelligence (ICSI)*.
- Yang, J. and Leskovec, J. (2011). Patterns of Temporal Variation in Online Media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Yang, Y. and Pedersen, J. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*.
- Ye, M., Yin, P., and Lee, W.-C. (2010). Location Recommendation for Location-based Social Networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*.
- Yin, P., Luo, P., Wang, M., and Lee, W.-C. (2012). A Straw Shows Which Way the Wind Blows: Ranking Potentially Popular Items from Early Votes. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*.
- Yu, B., Chen, M., and Kwok, L. (2011). Toward Predicting Popularity of Social Marketing Messages. In *Proceedings of the 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP)*.
- Yu, X., Liu, Y., Huang, X., and An, A. (2010). A Quality-aware Model for Sales Prediction Using Reviews. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*.
- Zhang, J., Ackerman, M., and Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th International World Wide Web Conference (WWW)*.
- Zhang, R. and Tran, T. (2008). An Entropy-Based Model for Discovering the Usefulness of Online Product Reviews. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.
- Zhang, Z. and Varadarajan, B. (2006). Utility Scoring of Product Reviews. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM)*.
- Zwillinger, D. and Kokoska, S. (2000). *CRC Standard Probability and Statistics Tables and Formulae*. Chapman & Hall.