CLASSIFICAÇÃO DE TÓPICOS ATRAVÉS DE MODELOS DE LINGUAGEM ENRIQUECIDOS COM CONTEXTO

ALEXANDRE GUELMAN DAVIS

CLASSIFICAÇÃO DE TÓPICOS ATRAVÉS DE MODELOS DE LINGUAGEM ENRIQUECIDOS COM CONTEXTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Adriano Alonso Veloso

Belo Horizonte
Fevereiro de 2015

ALEXANDRE GUELMAN DAVIS

SUBJECT CLASSIFICATION THROUGH CONTEXT-ENRICHED LANGUAGE MODELS

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Adriano Alonso Veloso

Belo Horizonte
February 2015

Davis, Alexandre Guelman

D261s Subject Classification Through Context-Enriched Language Models / Alexandre Guelman Davis. — Belo Horizonte, 2015
xxiv, 58 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Adriano Alonso Veloso

1. Computação – Teses. 2. Processamento da linguagem natural (Computação) – Teses. 3. Redes sociais on-line – Teses. 4. Linguística - Processamento de dados. 5. Recuperação da informação. Orientador. II. Título.

CDU 519.6*73(043)

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

FOLHA DE APROVAÇÃO

Subject classification through context-enriched language models

ALEXANDRE GUELMAN DAVIS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Adriano Alonso Veloso - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Renato Martins Assunção
Departamento de Ciência da Computação - UFMG

Prof. Wagner Meira Júnior
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 23 de fevereiro de 2015.

Acknowledgments

This dissertation is far to be a single person work. The long and winding path that we walked to get to this moment would not be possible without the help of many important people in my life. In here, I thank them for not only for their support but also for helping me to become a better person and a better professional.

First of all, I would like to thank my family. To my parents Clodoveu and Elizabeth, who always did their best to provide me with everything that I needed and lead me in the way of excellence, rightousness and knowledge, I owe you everything that I am today. A special thanks to my sister Paula for the company and endless friendship. Without your readiness to help me (even, sometimes, without being able) I would not be completing this degree. I am really lucky of being born in such an amazing family. Thank you, I love you all.

To my greatest companion and loving girlfriend Bruna, who tirelessly helped me and supported me in all my decisions in the last eight years, you have my greatest gratitude. I simply cannot imagine what would my life be without you close to me. I am really gratefull for having such an amazing, big hearted, loving and intelligent life companion. I definetly have still a lot to learn with you, I hope that the rest of my life is time enough. Thank you.

For all the classes and technical learning, a really special thanks to all my professors and teachers. Without your knowledge and patience I would not be finishing this important step in my life. To my advisor Adriano, who always helped me in reaching my goals and to achieve excellence in research, my gratitude.

Thanks to all my friends in e-Speed Lab. To Walter and Coutinho, who were willingly and ready to help anytime, you have my sincerest gratitude. You are one of the most friendly and amazing people I had the pleasure meet in UFMG. Special thanks to Silvio, for hearing my cries during 3 months in that rainy cold Seattle. Another special thanks to Alessandro, for hearing my cries during 5 months in that dry hot Belo Horizonte. Thanks to Itamar for helping me to improve this work even more. A warm thanks also for everyone in Meia Boca Juniors, Tiago Cunha, Luiz,

Osvaldo, Glauber, Herico, Paulo, Leandro, Samuel and Juan, without you Saturday mornings would be a little sadder.

Thanks to my friends in grad 091: Victor, Vellozo, Mattos, Felipe, Anderson. Thanks to my friend Diogo Renno, for being the best partner in groups assignments. To Filipe, Dilson and Henrique, for all the amazing moments in programming contests and in Russia. Yes, I am still not sure how we survived that trip. Thanks, also, to all my school friends: Alexandre Morato, Vitor, Bu, Otoni and Andre. A special thanks to Tereza, to my grandparents, uncles, aunts and cousins.

To my friends at Microsoft and Google, especially to Angelo, Doug Szabo, Ivan Tang, Lucas Vendramin and Ranveer, my thanks for hosting me and teaching me how to be a better professional. Certainly, without these internships my degree would not be so complete.

At last but not least, I would like to thank one of my greatest inspiration sources, Clube Atletico Mineiro. A special thank to the team: Victor, Marcos Rocha, Leo Silva, Rever, Richarlyson, Pierre, Leandro Donizete, Bernard, Ronaldinho, Tardelli and Jo, the 2013 Libertadores title will be a unforgettable reminder in my life that every thing is possible and all of our dreams are achievable if we believe in them. I BELIEVE!

“In the midst of chaos, there is also opportunity”

(Sun Tzu)

Resumo

Ao longo dos anos, humanos desenvolveram um complexo e intricado sistema de comunicação, com diversas maneiras de transmitir informações, que vão de livros, jornais e televisão até, mais recentemente, mídias sociais. No entanto, recuperar eficientemente e entender mensagens de mídias sociais para a extração de informações úteis é desafiador, especialmente considerando que mensagens mais curtas são mais dependentes do contexto. Usuários muitas vezes assumem que o público de suas mídias sociais está ciente do contexto associado e de eventos do mundo real subjacentes. Isso permite que eles encurtem as mensagens sem prejudicar a efetividade da comunicação.

Algoritmos tradicionais de mineração de dados não levam em consideração informações contextuais. Consideramos que explorar o contexto pode levar a uma análise mais completa e precisa das mensagens de mídias sociais. Neste trabalho, portanto, é demonstrado o quão relevantes são as informações contextuais na filtragem de mensagens que são relacionadas a um dado assunto (ou tópico). Também é mostrado que a taxa de recuperação aumenta se o contexto for levado em consideração. Além disso, são propostos métodos para filtrar mensagens relevantes sem utilizar apenas palavras-chave se o contexto for conhecido e datectável. Nesta dissertação, propomos uma nova abordagem para classificação de tópicos em mensagens de mídias sociais que considera tanto informações textuais como extra-textuais (ou contextuais). Essa abordagem propõe e utiliza modelo de linguagem enriquecido com contexto. Técnicas baseadas em conceitos de linguística computacional, mais especificamente na área de Pragmática, são utilizadas. Para avaliar experimentalmente o impacto dessas propostas foram utilizados conjuntos de dados contendo mensagens sobre três importantes esportes americanos (futebol americano, baseball e basquete). Resultados indicam uma melhora de até 50% na recuperação de mensagens sobre estratégias baseadas em texto devido à inclusão de informação contextual.

Palavras-chave: Processamento de Linguagem Natural, Mídias Socias, Linguística Computacional, Recuperação de Informação.

Abstract

Throughout the years, humans have developed a complex and intricate system of communication with several means of conveying information that range from books, newspapers and television to, more recently, social media. However, efficiently retrieving and understanding messages from social media for extracting useful information is challenging, especially considering that shorter messages are strongly dependent on context. Users often assume that their social media audience is aware of the associated background and the underlying real world events. This allows them to shorten their messages without compromising the effectiveness of communication.

Traditional data mining algorithms do not account for contextual information. We argue that exploiting context could lead to more complete and accurate analyses of social media messages. For this work, therefore, we demonstrate how relevant is contextual information in the successful filtering of messages that are related to a selected subject. We also show that recall rate increases if context is taken into account. Furthermore, we propose methods for filtering relevant messages without resorting only to keywords if the context is known and can be detected. In this dissertation, we propose a novel approach for subject classification of social media messages that considers both textual and extra-textual (or contextual) information. This approach uses a proposed context-enriched language model. Techniques based on concepts of computational linguistics, more specifically in the field of Pragmatics, are employed. For experimentally analyzing the impact of the proposed approach, datasets containing messages about three major American sports (football, baseball and basketball) were used. Results indicate up to 50% improvement in retrieval over text-based approaches due to the use of contextual information.

Keywords: Natural Language Processing, Social Media, Computational Linguistics, Information Retrieval.

List of Figures

List of Tables

List of Acronyms

AKC: Agent Knowledge Context.

CMC: Computer Mediated Communication.

MLB: Major League Baseball.

MLE: Maximum Likelihood Estimator.

NBA: National Basketball League.

NFL: National Football League.

NLP: Natural Language Processing.

OCR: Optical Character Recognition.

PDF: Probability Distribution Function.

SC: Situational Context.

Contents

Chapter 1

Introduction

Human communication involves a sender, a medium, a message, and a receiver. Variations and combinations in these four elements allow for a broad set of communication alternatives, that range from personal oral conversations to complex written documents. Over time, the forms of communication between humans have evolved significantly, especially through the use of technology. The rise of new communication technologies have been associated, in the past, to revolutions in human behavior: the invention of the telegraph, of the telephone, and of the television have been decisive in defining the characteristics of modern human societies all over the world. Computer-based communications followed, granting a new and faster way to exchange messages, initially using e-mail and transmitting digital versions of written documents, and then evolving to mixing text with images and video. The increasing ease and reduced cost of electronically sending and receiving messages caused a significant increase in human communication, covering all kinds of subjects, from events of personal concern to news of global importance.

In a recent phenomenon, many people have started using online social media as their primary channel for interaction and communication. Social interaction with other people has been partly transferred to online social networks, in which communication can take place between a single sender and large numbers of receivers, using text messages along with images and video. The use of mobile devices as instruments for this kind of communication allowed the subject of such messages to be associated to events and phenomena of immediate (and usually ephemeral) interest, potentially covering all kinds of themes and topics. There are billions of users of social media services, and the number of messages grows continuously. In only one of the most popular social networks, the number of daily exchanged messages reached 64 billion in late 2014[1]. In another, the number of active users exceeds 1.3 billion (one of every 5 humans), and over 300 million images are uploaded and distributed every day[2].

Since in this kind of communication messages can be digitally encoded, stored, accumulated and analyzed as a stream of data, many interesting research opportunities arise, associated to the challenge of handling large volumes of real-time data and fueled by the overall goal of increasing our understanding of human societies. Suppose that a social scientist becomes interested in analyzing the repercussion of some theme, or subject, of current interest. If it is possible to, routinely, (1) tap into a social media source, (2) filter out messages that are associated to the selected subject, and (3) analyze the set of messages according to parameters such as content, vocabulary, volume, emotional bias, propagation speed, spatial distribution, duration, and others, then the scientist can frequently have instantaneous glimpses on the way society treats that subject and can advance current knowledge for similar situations in the future. Using large volumes of data from social networks, social sciences achieve a scale that has never been possible before, in one of the many facets of the so called *data science* phenomenon.

However, there are numerous difficulties in every step of this process. In the first step, not every social media provider allows broad access to messages, user profiles and user relationships. The second and third steps introduce complex challenges as to accurately selecting messages as being related or not to a given subject, and on analyzing sets of messages related to a given theme. Furthermore, the sheer volumes of messages that are exchanged daily indicate that only automated approaches can be used in a meaningful way, and that the complexity of these approaches is limited by the computational infrastructure that is made available to run them.

Consider, specifically, the case of Twitter. It is the fifth social network as to the number of users, but its characteristics make it a primary source for research initiatives. Twitter offers a comprehensive Application Programming Interface (API), with clear rules for collecting and filtering messages in significant volumes. User profiles can be obtained, and user relationships can be retrieved. On the other hand, messages are limited to 140 characters, and thus issues such as abbreviations, misspelled words and informal lexical and grammatical constructions are common. Twitter users also frequently rely on contextual information[3] as a way to keep messages short. Assuming that the receivers will be able to recognize the context, users frequently omit key parts

[1] WhatsApp, according to http://www.statista.com/statistics/258743/daily-mobile-message-volume-of-whatsapp-messenger/

[2] Source: https://zephoria.com/social-media/top-15-valuable-facebook-statistics/

[3] By *context* we mean all non-written information that is relevant to understand a message, such as real world events and common sense knowledge

of the message, posting incomplete sentences or chaining ideas over several messages spread through time.

Regardless of such difficulties, filtering and analyzing short messages in a medium such as Twitter has been successfully done in the recent past, with significant results regarding problems such as sentiment analysis [Guerra et al., 2011], spatial location [Davis Jr et al., 2011] and epidemics forecasting [Gomide et al., 2011] (all of these are results from our research group at UFMG). A broad range of techniques has been used in those studies, including data mining, machine learning, and natural language processing (NLP). In most cases, filtering has been simplified to use sets of keywords, in a strategy applied in large scale by the Observatory of the Web project[4]. Contextual information has usually not been taken into consideration. In part, the success of such initiatives comes from the fact that the number of messages for which context is not necessary is still large, allowing for the recognition of trends and for the identification of relevant information.

Consider, for example, an individual social media user A, that at some point in time issues a message containing the sentence “And he scores!!”. Without any contextual information, an automated filtering or analysis agent would have trouble associating such a message to any possible subject. However, the receivers of the message, individuals B_i in A’ s social network, know from previous interactions that A is a soccer fan, who roots for a specific team T. Some of those may further know that a game is going on at the moment the message is issued, therefore deducing that some player of T has scored. Others may have read A’s previous messages, in which he criticized a specific player for his disappointing performance so far in the game, and can infer who is the player (“he”) implicitly mentioned in the message. Therefore, a message containing only three words and only 15 characters can convey a lot of information to its receivers, but most of that information has to be gathered from the context, including information on user previous behavior, preferences, habits, and external information on game schedule, news and much more. A keyword-based filtering system would have missed all this information, losing a valid indication of a crucial moment (scoring) in the event (the soccer match)[5]. Humans can do that in a natural way, but transferring this kind of possibility to a computer represents a big challenge.

For this work, therefore, we pose some questions: how relevant is contextual information in the successful filtering of messages that are relevant to a selected subject?

[4] observatorio.inweb.org.br/

[5] More than 35 million tweets were issued during the Brazil vs Germany match in the FIFA World Cup 2014 semifinals, an event of global repercussion, the most tweeted event on record so far – but who knows how many tweets are not included in this figure due to lack of contextual information?

Can the recall rate increase if context is taken into account? Furthermore, if the context is known and can be detected, is it possible to filter relevant messages without resorting only to keywords? In this dissertation, we propose a novel approach for subject classification of social media messages that considers both textual and extra-textual (or contextual) information. Techniques based on concepts of computational linguistics, more specifically in the field of *pragmatics*, are employed. For experimentally analyzing the impact of the proposed approach, different datasets containing messages about three major American sports leagues (football, baseball and basketball) were used.

1.1 Statement

Based on Pragmatics theory, we posit that text is not informative enough for subject classification in social media. Therefore, we need extra-textual information to supplement written message in this scenario to achieve better classification recall. In this dissertation, we propose contextual models (from Pragmatics) and language models that approximate human reasoning to include contextual information in message interpretation. Finally, we claim that messages uttered in personal and informal media require this kind of information to be fully understood.

1.2 Objectives

The main objective of this work is to demonstrate the importance of context to infer the subject of social media messages. We propose a novel method for identifying subjects, considering both textual and extra-textual information. This method is an adapted language model that uses context models (also proposed as part of this work) to supplement written information.

With the proposed techniques, our goal is to show that it is possible to increase the recall of subject-related messages in social media without using a bag-of-words data gathering model. We posit that retrieval based on fixed keyword selection is misleading for some kinds of social media messages, specifically the ones that are deeply affected by pragmatic effects and *conversational implicatures*, i.e., elements of language that contribute to understanding a sentence without being part of what is said, nor deriving logically from what is said.

1.3 Specific objectives

The following specific objectives were pursued in this work:

- To study and organize concepts related to computational linguistics, more specifically about pragmatics, in order to apply them to the problem of filtering social media messages;

- To obtain, organize and annotate a corpus of messages pertaining to popular sports (football, baseball and basketball), with which to analyze and measure the effect of context;

- To conceive and implement filtering algorithms that are able to consider contextual information;

- To experimentally analyze the proposed techniques, quantifying their effect over the test data, and obtaining valuable insights as to the application of the proposed techniques in other applications.

1.4 Challenges and contributions

The challenges and contributions of this work follow:

- **Data gathering without keywords:** Retrieving a tweet stream without using bag-of-words selection was a great challenge. Twitter's API only provides three options: (1) retrieve all messages that contain a set of keywords, (2) retrieve 1% of the entire message stream and (3) retrieve all messages posted by a set of users. Since we want to identify subjects in tweets without resorting to keywords, and in the second option we would get an extremely sparse stream, the only option left was to select by users. We proposed a method and demonstrated some hypotheses for choosing a particular set of users and enabling the use of this stream.

- **Contextual elements in the labeling process:** It is important, by our hypothesis, to consider contextual elements for labeling test messages. Since human annotators are not on the same place, time and may not have the same knowledge as the original receivers of the Tweet, it may get a different interpretation of the message. Therefore, we tried to reproduce as much as possible the original post's context in our label environment by displaying user profile (e.g. avatar photo, background image), previous posts, and considering the time of posts to the human rater to give a better background understanding. This is an uncommon approach in the literature, which usually restricts rating to text only.

- **Pragmatic contextual models:** Contextual models are an innovative way of scoring the likelihood of a message to be related to a subject according to its contextual information. The score generated by these models can be used alone or associated with a language model.
- **Context-enriched language models:** To combine the novel pragmatic contextual model with the text posted in messages, we propose a new language model. The idea of combining non-textual elements with text is a major contribution of this dissertation.

1.5 Structure

The remainder of this dissertation is organized as follows. Chapter 2 covers literature that is relevant to our proposal and explores some basic concepts. Chapter 3 presents a discussion on pragmatic effects in social media messages, and introduces methods for modeling context. Next, a new language model that uses these contextual scores is proposed in Chapter 4. Chapters 3 and 4 contain the major contributions of this dissertation. Chapter 5 contains a description of the datasets used in experimentation, along with demonstrations of the proposed contextual models and empirical demonstrations on the validity of assumptions made in Chapter 3. Chapter 6 presents the results of the experimental evaluation of the proposed language model. Finally, Chapter 7 shows conclusions and discusses future work.

Chapter 2

Related Work

In this chapter, we present some literature related to this dissertation. We start with a section discussing pragmatic effects and conversational implicatures, which are our main motivation to claim the importance of context for social media messages interpretation. Then, we discuss language models, which are the foundations of the proposed technique. Next, we show other works on text categorization, a more general version of the subject classification problem. Finally, we show some initiatives that tried to include information from different sources to enrich text when it is scarce.

2.1 Pragmatic effects and conversational implicatures

Contextual influence over text has been studied by linguists, comprising the so-called *pragmatic effects* [Levinson, 1983]. Pragmatic effects are related to the human ability to use context for changing the uttered message's semantic meaning. They are frequently present in communication and may manifest themselves in several different ways, according to the medium (e.g. written, oral, computer mediated), subjects involved, situation of communication, previous shared knowledge between subjects, and others. Pragmatics effects have a major role in short and informal messages, such as those that are common in social media.

Pragmatics effects may manifest in many different sources of information. In some situations, pragmatics effects are used as a form of generating humor [Yus, 2003]. In this case, the change in the sentence meaning generates an unexpected or ironical interpretation. Another interesting pragmatic effect derives from facial expressions and eyeball movements. Humans are able to detect minimal movements in eyeball that indicate changes in the meaning of uttered messages [Hanna and Tanenhaus, 2004]. However, one of the most common pragmatic effects are the *conversational implicatures* [Cruse, 2006; Levinson, 2000].

Conversational implicatures are an implicit speech act. In other words, they are part of what is meant by a speaker's utterance[1] without being part of what was uttered. Implicatures have a major contribution to sentence meaning without being strictly part of 'what is said' in the act of their enunciation, nor following logically from what was said. Therefore, implicatures can be seen as a method of reducing the gap between what is literally expressed and the intended message [Levinson, 1983]. A good example of implicatures is given by Levinson [1983]. If a speaker A tells receiver B "The cat is on the mat", without any contextual information, this sentence may be interpreted as a simple statement and no action is required. However, if both A and B know that this cat usually goes to the mat when it is hungry, the uttered message may be interpreted by B as "The cat is hungry. Please feed it". This example is interesting because the intended meaning of the message is completely modified once both A and B share a common piece of information.

Conversational implicatures are essential to expand oral language expressiveness, since humans have a tendency to contract spoken messages as much as possible without compromising message comprehension [Levinson, 2000, 1983]. Conversational implicatures, however, are not exclusively used in spoken communication, they are also important on computer mediated communication (CMC), especially on social media (such as blogs, instant messaging, and in social media messages typical of Twitter and Facebook, among others) [Barbulet, 2013]. As previously mentioned, in this kind of media users want to post short and concise messages because of time and space constraints.

Despite of being really important for informal and personal communication, most of the natural language processing (NLP) approaches overlook conversational implicatures. In this thesis, we propose to consider implicatures in language models, one of the most traditional NLP techniques. In the next section, we discuss some previous works on language models.

2.2 Language models

Statistical language models are an important class of algorithms in natural language processing (NLP) literature. The main goal of these models is to estimate the probability of finding a given sequence of words in a set of sentences, commonly called

[1] An uninterrupted chain of spoken or written language.

in this context a *language*. One of the most traditional methods for estimating that probability is the n-gram language model [Pauls and Klein, 2011]. This model assumes that words in the input sequence are not independent. Moreover, it argues that the probability of a word w in the input sequence depends only on its textual context C (usually described as the n words preceding w). Therefore, the n-gram language model estimates the probability $P(w|C)$, which is useful in many applications, such as automatic translation [Pauls and Klein, 2011] and speech recognition [Saluja et al., 2011]. For these applications, however, methods that consider $P(w|C) \neq P(w)$ perform much better than those that assume independence between words in a sentence [Ifrim et al., 2008]. We argue that, in social media text, conversational implicatures make textual context C unreliable to estimate this probability. For this matter, we propose to add extra-textual context to improve the accuracy of language models for this specific kind of data.

Another interesting aspect of language models is that they can be used as an unsupervised learning method, since they use a set of unlabeled sentences in their construction. This characteristic makes them useful for information retrieval and text categorization applications. Many information retrieval techniques, for instance, improve the ranking of Web pages or texts by using language models [Kurland and Lee, 2010]. In these cases, the input word sequence is given by the user while constructing a query. Regarding text categorization, some authors proposed approaches that use language models for improving robustness in texts with misspelled words and syntactic errors [Cavnar and Trenkle, 1994], such as those extracted from e-mails and from documents generated by optical character recognition (OCR). Alternatively, language models may be used for clustering similar documents into categories [Erkan, 2006] and for improving hypotheses in vastly used classifiers, such as Naive-Bayes [Peng et al., 2004] and Logistic Regression [Ifrim et al., 2008]. In the next section, we show some other approaches for text categorization.

2.3 Text categorization

Text categorization (or classification) is a classic problem in Machine Learning and NLP. The goal is to assign previously defined labels, or classes, to fragments of real-world texts. Depending on the applications proposed, those labels may also be called topics. In early works, news corpora and medical records have been used as textual data sources [Hayes et al., 1988; Yang and Pedersen, 1997]. Recently, text categorization approaches have been adapted to challenging datasets, such as blog texts [Mishne, 2005a,b] and spam filtering [Androutsopoulos et al., 2000; Drucker et al., 1999]. Some challenges associated with these new datasets are text sparsity and oralism, which lead to adaptations in traditional machine learning algorithms.

Traditional machine learning text classification approaches have addressed some of these issues by using a bag-of-words model, in which each word is used as a feature for a supervised classifier [Joachims, 1998]. Unfortunately, this model produces an enormous number of dimensions, which may impact in the classifier's accuracy [Sebastiani, 2002]. Therefore, authors have proposed dimensionality reduction techniques [Yang and Pedersen, 1997; Guyon and Elisseeff, 2003] for achieving better results. Other authors argue that generative models are more robust to unstructured text and to situations where a document may have multiple labels [Schwartz et al., 1997; Natarajan et al., 2007]. Such approaches are heavily dependent on textual features and on big training sets. Thus, it would be unfeasible to use them on social media message streams. Our proposal uses contextual features and unsupervised methods (language models). Consequently, our approach is less dependent on text quality and does not require any previous manual labeling effort.

Recently, incremental online classifiers were proposed for solving text categorization in scenarios that require constant updates in training data [Crammer et al., 2012; Guan et al., 2009]. Another possibility is to automatically update online classifiers with techniques such as EM^2 [Davis et al., 2012]. However, in some datasets, such as blogs and social media text, changing the training set may not be enough. Many authors addressed the vocabulary dynamicity and text sparsity problems using semantic analysis [Li et al., 2011; Guo et al., 2010; Qiming et al., 2011]. We believe that this approach is not effective in social media datasets, given the number of grammatical and spelling errors usually found in this kind of text. Therefore, in these cases, we may need to find other information sources to complement or substitute textual information. In the next section, we show some works that used alternative data sources for improving the accuracy of categorization and classification systems in scenarios which the textual information is scarce or incomplete.

2.4 Alternative data sources

In applications for which textual data is sparse, some works resort to external sources (including knowledge bases such as Freebase, WordNet and Wikipedia) to help in the categorization process [Husby and Barbosa, 2012; Lao et al., 2012; Li et al., 2012; Son et al., 2013]. Some authors also use custom ontologies as input for their methods

[Machhour and Kassou, 2013]. Another usage for external knowledge sources is to relate entities using inference rules [Lao et al., 2012; Raghavan et al., 2012]. Although this is a good way of contextualizing information, we observe that external sources are not updated at the same pace as new expressions and terms are created in social media (for instance, slang, nicknames, ironic expressions). Alternatively, we may search for contextual information within the dataset itself [Lam et al., 2001]. One contribution of our proposed approach is to extract this information using mostly meta-information on the messages. Our goal is to use this kind of information to mimic conversational implicatures, as described by Pragmatics Theory [Levinson, 2000], commonly used in human communication.

According to Pragmatics Theory, as texts become shorter and more informal, environment knowledge becomes more important to discover the message's meaning [Yus, 2011]. Moreover, in instant messaging and online chat discourse, users have a tendency to contract the text as much as possible with the objective of typing faster or mimicking spoken language [Herring, 2001]. For instance, character repetition is a very common way of reproducing excitement in this kind of text (e.g "Gooooooal") [Brody and Diakopoulos, 2011]. In social media, the sender's knowledge of real-world events [Howard and Parks, 2012], location [Son et al., 2013] and popularity [Cha et al., 2013] may influence the elaboration of messages, assuming a similar set of knowledge on the part of the receiver. Bayesian logic has been used to extract "common-sense" beliefs from natural language text [Raghavan et al., 2012]. A recent work has proposed combining implicit information such as tone of voice and pauses in sentences for improving the results in speech translation techniques [Saluja et al., 2011]. Our proposal is to model this implicit external influence in social media messages, and to use it to classify each one according to predefined labels, or subjects.

Chapter 3

Pragmatic Contextual Modeling

Previous pragmatics studies argue that context is an important information source for understanding the message's utterance. One of the most common examples is spoken communication, where, the meaning of the speaker's message is frequently not clear in the utterance and, therefore, requires additional knowledge to be interpreted. For better comprehending the dimensions of communication, Grice introduced the *cooperative principle* [Grice, 1975], a major contribution to pragmatics theory, which describes how people communicate with each other. This principle defines assumptions or maxims on the nature of the information that is expected to be exchanged in conversation (Definition 3.1).

Definition 3.1 (Cooperative Principle) *Pragmatic contributions to sentence meaning should follow the accepted purpose or direction of the talk exchange in which the speaker is engaged. Therefore, contributions should usually respect the following maxims:*

Maxim of Quantity: *The author provides only the amount of explicit information necessary for the listeners to complete the communication.*

Maxim of Quality: *There is no false information.*

Maxim of Relevance: *The information needs to be relevant to the conversation topic.*

Maxim of Manner: *The information should be clear.*

Another important contribution to pragmatics is Hirschberg's definition of *proposition* in conversational implicatures [Hirschberg, 1985]. According to Hirschberg, implicatures are composed by propositions, which are part of the inference process that the receiver is required to do if he believes the sender is following the cooperative principle. Therefore, propositions can be seen as bits of information that are considered by the sender to be known by the receiver as well. If this assumption is false, the cooperative principle fails and, usually, there is a communication break. In some cases, however, violations in the cooperative principle are used for humor and irony [Attardo, 1993; Eisterhold et al., 2006]. Because of that, violation in Maxim of Quality is common in social media.

Definition 3.2 (Hirschberg's proposition) *Given a utterance U and a context C, proposition q is part of a conversational implicature of U by agent B if and only if:*

1. *B believes that it is public knowledge in C for all the discourse participants that B is obeying the cooperative principle.*
2. *B believes that, to maintain item 1 given U, the receiver will assume that B believes q.*
3. *B believes that it is mutual and public knowledge for all the discourse participants that, to preserve item 1, one must assume that B believes q.*

Definitions 3.1 and 3.2 can be better explained using an example, as follows. During a football match someone listens agent B uttering *"Oh NO!"* (this is U). Unconsciously, the recipient of U infers that *"Something bad happened with the team B cheers for"* (this is q). The context C, in this example, informs not only that there is a match happening at utterance time, but also the team B cheers for. Therefore, q follows definition 3.2.(1), by which B expects that the discourse participants are aware of the necessary knowledge in C to understand the message.

Both the cooperative principle and Hirschberg's definition guide some properties in conversational implicatures that are essential to the proposed contextual models. The first important property is that implicatures are **non-lexical**; therefore, they cannot be linked to any specific lexical items (e.g. words, expressions). Second, implicatures **cannot be detached from utterances** by simple word substitution. In the football match example, it is clear that implicatures are not linked to any word in U (i.e., the proposition would be the same if the utterance was different, but issued to the same effect). Another important property is **calculability**: the recipient should be able to infer implicatures from utterances. In the football example, the recipient of the message needs to be able to deduce to which team something bad is happening.

Conversational implicatures and the cooperative principle, however, are not exclusively used in oral communication. It is clear that both of these theories apply to computer mediated communication (CMC), especially in social media [Yus, 2011]. Twitter messages provide many clear situations that fit these theories. For instance, the maxim of quantity is usually present in a considerable share of the messages, due to message size limitations - messages must have less than 140 characters. Also, the maxim of relevance is essential in Twitter, since messages are closely related to real time events. Finally, it is important to notice that the sources of contextual information in social media are not as rich as in spoken communication. Consequently, users add new forms of context to increase their expressiveness in this media, such as emoticons, memes, hashtags, videos and photos. All these characteristics make it even harder to understand the meaning of a social media message just by reading the plain text.

Recently, we have seen many approaches for sentiment analysis [Silva et al., 2011] and topic tracking [Phuvipadawat and Murata, 2010] on social media. Unfortunately, most of these proposals completely ignore extra-textual features. Many of those authors, for instance, use as data source a stream of messages selected by keywords. Under the conversational implicature hypothesis, we argue that keyword selection is not enough for filtering a subject-related message stream. Therefore, many of these works are using biased data streams with limited recall. In Chapter 5 (Table 5.3), we show an estimate of how much data is lost using keyword-selection. Our objective is to increase recall in Twitter subject-related data gathering. This approach relies on contextual scores given to each tweet. These contextual scores were proposed to imitate the role of information that is commonly used in conversational implicatures.

In this work, we focus on two major types of contextual information for social media communication: *agent knowledge context* and *situational context*. Agent knowledge context compresses information about an agent that can be accessed by all discourse participants (according to Definition 3.1). For this category, it is expected that we have information such as agent interests in subjects (e.g. american football, baseball, basketball). On the other hand, situational context compresses information about real-world events that are seen as common knowledge between all discourse participants, such as football matches, scoring situations (e.g. touchdowns and goals), team news (e.g. player hiring and injury reports). In the next sections, we propose models to evaluate the relation of each of these types of context to a subject in a Twitter message. These models generate a numeric score that indicates the model's confidence on the relatedness between subject and message context.

3.1 Agent Knowledge Context

A user in social media is generally interested in several subjects. It is straightforward to assume that a user interested in american football will issue messages about this subject many times. Therefore, receivers would often be able to identify the interests of a given user by looking at the distribution of messages related to each subject over time. It is also interesting to notice that knowledge about a user is built based on a longer term relation, needed to increase the quality of the contextual information the receivers have about the user.

The proposed agent knowledge model tracks the amount of messages uttered by a user b in a long period training stream related to the subject S. As argued throughout this dissertation, due to some characteristics in communication, it is hard to identify whether a message is related to S. For that matter, we use an heuristic to compute the proposed agent knowledge model that is based on the frequency of manually selected keywords K_s related to S in previous messages from b. We select keywords that are trivially related to the subject S to be included K_s, the criteria will be better explained in Chapter 5 (The exact slected keywords can be further analyzed in Attachment A). This heuristic considers two important hypotheses, as follows:

Hypothesis 3.1 *The chance of a user b being interested in a subject S is proportional to the frequency of keywords from K_S that have been used by b related to S.*

Hypothesis 3.2 *Ambiguous keywords are used by more users (including users that are not interested in S) than unambiguous ones.*

The intuition behind Hypothesis 3.1 is that if a user b frequently utilizes words in K_s, it is probable that b is interested in S. For instance, a user that, over two months, has posted a hundred keywords related to baseball is more likely to be interested in baseball than a user that posted only ten keywords of the same set in that period. However, these keywords can be ambiguous, (i.e., keywords in K_s can also belong to K_t, where T is a subject that is different from S) and Hypothesis 3.2 tries to neutralize this.

For understanding the intuition behind Hypothesis 3.2, consider two sets of users: U_s are the users interested in S and $\overline{U_s}$ are everyone else. It is expected that more ambiguous keywords (such as “NY”, “Boston”) are used by both U_s and $\overline{U_s}$, while unambiguous keywords (such as “Packers”, “Red Sox”) are referenced mostly by users in U_s. Since $|U_s| << |\overline{U_s}|$, ambiguous keywords are referenced at least once by a wider range of users.

With Hypotheses 3.1 and 3.2, we can propose a score for the relation between an agent and a subject S given the agent knowledge context. The score is similar to the traditional TF-IDF (text frequency - inverse document frequency) measures used in information retrieval. Following Hypothesis 3.1, we define a value TF_k^b (Equation 3.2) that represents the normalized number of times that a keyword $k \in K_s$ was used by a user b in the analyzed period. For each keyword, we define a value IDF_k (Equation 3.1), which gives a weight to k according to the ambiguity Hypothesis 3.2. These values are calculated using the following formulas, where N is the number of users, n_k is the number of users that used k at least once in the given stream and f_k^b is the number of times k was uttered by user b.

$$IDF_k = log(\frac{N}{n_k}) \tag{3.1}$$

$$TF_k^b = 1 + log(f_k^b) \tag{3.2}$$

Finally, we define a score AKC^b for each user b in the stream, defined by the product of IDF_k and TF_k^b (Equation 3.3). Information retrieval approaches normalize the TF-IDF score by the document size. As described ahead, in Chapter 5, the stream used for model generation (i.e. training stream) was created by keyword projection. Therefore, it is impossible to normalize this value by the number of messages posted by b, since the dataset contains only messages that include at least one keyword $k \in K$.

$$AKC^b = \sum^{\forall k \in K_s} (IDF_k * TF_k^b) \tag{3.3}$$

Having thus defined the agent knowledge context, which will allow us to quantify the effect of the user's knowledge of a subject, next section approaches the situational context, by which the effects of concurrent events on the utterance of messages will be evaluated.

3.2 Situational Context

Social media messages are often closely related to real world events. We argue that most users expect that their receivers are aware that a real world event at the time of an utterance may be an important contextual information for the posted message. Consequently, we posit that tweets posted during important events of S are more likely to be related to S (Hypothesis 3.3).

Hypothesis 3.3 *Messages are more likely to be related to a subject S if they were posted within a timeframe that contains meaningful real world events related to S.*

In the proposed situational model, our goal is to measure the relation of a message posted during a time window T with subject S[1]. Given Hypothesis 3.3, we want to assign higher scores for messages that were posted during important events related to S. Unfortunately, since most subject-related events are highly dynamic, it is unfeasible to extract such contextual knowledge from external sources. In this section, we define methods for estimating the likelihood of such events to be happening during T given the messages posted in that period of time.

For estimating this likelihood, we hypothesize that events are more likely to occur in timeframes in which a higher frequency of unambiguous keywords related to S is observed (Hypothesis 3.4). For instance, during American football season matches, a much higher frequency of unambiguous football-related keywords are expected, such as "Green Bay Packers", "GoPats" and "New York Jets". If there was no match happening, the frequency of those keywords would be much lower. It is important to notice that matches are not the only important events in this scenario, breaking news, hiring and draft may also impact the Situational Context of a message.

Hypothesis 3.4 *The probability of a meaningful event related to S to be occurring during time window T is correlated with the number of unambiguous keywords related to S that are posted during T.*

Following Hypothesis 3.4, we define a score SC^T (Equation 3.5) for each fixed time window T, to measure the likelihood of a meaningful event related to S to be happening during T. To calculate this score, we use TF_k^T to track the frequency of a keyword k during a time window of T. Finally, we normalize the score with the number of tweets in the training stream that have been issued during that time window.

$$TF_k^T = 1 + log(f_k^T) \tag{3.4}$$

$$SC^T = \sum^{\forall k \in K_s} \frac{IDF_k * TF_k^T}{|T|} \tag{3.5}$$

It is important to notice that, despite of being TF-IDF-based, situational context and agent knowledge are not correlated. Therefore, these contextual sources may be used simultaneously by human cognition in the message interpretation process. Consequently, to improve even further the ability to reproduce cognition using pragmatics,

[1] For simplicity, fixed-length time windows are adopted.

we need methods for combining contextual features. Analyzing the potential in act of mixing multiple contextual features is left for future work. In the next chapter, we propose a novel language model that uses these proposed scores for improving retrieval performance.

Chapter 4

Language Model

As discussed previously, conversational implicatures are a common pragmatic effect in social media text. Space constraints and speech informality increase the importance of context for interpreting short messages. Unfortunately, most of the traditional NLP and Machine Learning techniques ignore extra-textual information, mainly because they were developed for self-contained textual sources, such as news and books. One of the most important classic NLP techniques, Language Models have been used as abstract methods of representing language patterns. This representation involves using probability estimations for the presence of a word within a sentence. We propose a novel approach for estimating such probabiities using extra-textual information. This approach is detailed in Section 4.2, after reviewing the classic n-gram language model.

4.1 n-gram language model

The traditional n-gram language model assumes that the probability of a word w to be part of a language depends only on the previous $n-1$ words. These sequences of words are called n-grams. This way, the goal of a n-gram language model is to estimate the probability $\hat{P}(w_i|w_{i-1}, w_{i-2}, ..., w_{i-n})$, given L, a training set of text (commonly called a *language sample*). The probability of a text fragment, or message, m to be part of a language is given by the following formula:

$$P(m \in L) = \prod^{w_i \in m} P(w_i|w_{i-1}, ..., w_{i-n}) \tag{4.1}$$

A major concern about this model is that if a n-gram is not found in the training examples, the maximum likelihood estimator would attribute a null probability to this n-gram and, consequently the whole text fragment would get no probability of belonging to the language. Many previous works have used smoothing methods to force $P(w_i|w_{i-1}, ..., w_{i-n}) > 0$. For this work, we are using Katz backoff with Lidstone Smoothing, described in the next subsections, to estimate the probability of a n-gram belonging to the language.

4.1.1 Lidstone smoothing

Lidstone smoothing (or additive smoothing) is a technique used by the probability estimator to attribute a probability for items unseen in the training set. The main idea of this method is to add a constant value α for each event in the amostral space. This way, every time the estimator receives an n-gram not found in the training examples, it attributes a very small but greater than zero probability to the n-gram.

Consider c_i as the number of times an event i happens, and C as the amostral space ($\sum c_i = N, c_i \in C$). The estimated probability for the event i is given by a small modification to the Maximum Likelihood Estimator (MLE) (Equation 4.2). Notice that for each event in the amostral space, a value α is added independently from the event's frequency. Therefore, if $c_j = 0$, $\hat{P}(j) = \alpha/(N + \alpha|C|)$.

$$\hat{P}(i) = \frac{c_i + \alpha}{N + \alpha|C|} \tag{4.2}$$

The choice of α is essential for the smoothing quality, since bigger values imply in attributing a larger share of the probability space to unseen examples. Consequently, if α is too large, events with low frequency will be attributed a probability that can be similar to those with high frequency. This is called *smoothing underfitting*. In this dissertation, we chose the best α experimentally.

For the language model, we need to adapt this estimator to be used with conditional probabilities. To estimate the probability $\hat{P}(w_i|w_{i-1}, ..., w_{i-n})$, we also adapt from MLE. Therefore, we have the estimated probability in Equation 4.3, where $f_{w_i,...,w_{i-n}}$ is the frequency of a n-gram and $f_{w_{i-1},...,w_{i-n}}$ is the frequency of the $(n-1)$-gram in conditional probability prior.

$$\hat{P}(w_i|w_{i-1}, ..., w_{i-n}) = \frac{f_{w_i,...,w_{i-n}} + \alpha}{f_{w_{i-1},...,w_{i-n}} + \alpha|C|} \tag{4.3}$$

4.1.2 Katz backoff

When a n-gram cannot be found within a language model, it is interesting to notice that we may generalize it into a simpler $(n-1)$-gram, which is more likely to have occurred at least once in the dataset. The generalization property is an important

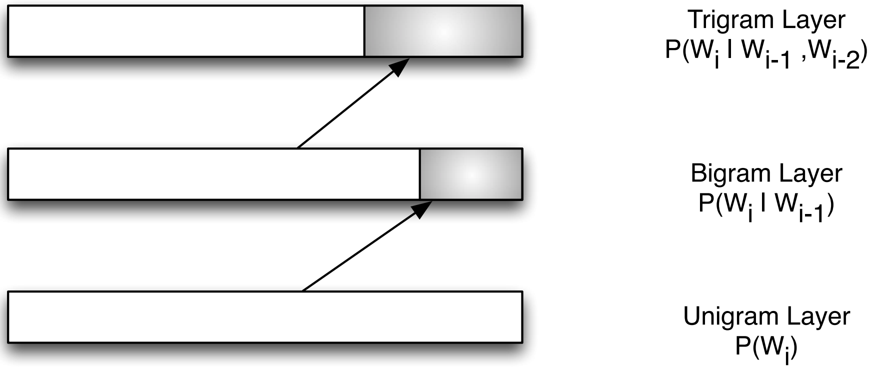

Figure 4.1. Ilustration of the classic Katz backoff Model for n-gram language model

feature of this kind of language model, enabling recursive search for a textual token in simpler models, also known as *backoff*. With the backoff, n-gram based language models are able to attribute a probability for a not found item that is proportional to the probability of finding it in a more general model. This is expected to perform better than attributing a constant probability, as in Lidstone smoothing. For this dissertation, we use Katz backoff, which is a method for estimating the probabilities $\hat{P}(w_i|w_{i-1},...,w_{i-n})$.

Katz backoff defines N levels of language models according to the generality. The first level is the most specific language model, which estimates $\hat{P}(w_i|w_{i-1},...,w_{i-n})$. If $f_{w_i} = 0$ or $f_{i-1,...,i-n} = 0$, backoff to the next level is needed. Similarly, the next level estimates $\hat{P}(w_i|w_{i-1,...,i-n-1})$. In the last level, we get the most general language model that estimates the probability of a unigram (or single token) belonging to the language model: $\hat{P}(w_i)$. The discount and backoff processes are illustrated in figure 4.1.

In order to maintain the main property of conditional probabilities, it is necessary that $\sum \hat{P}(w|w_{i-1},...,w_{i-n}) = 1, \forall w \in L$. Katz backoff accomplishes this by reserving part of the probability space for backing off. In our case, the reserve is similar to the Lidstone probability estimator (as seen on equation 4.3). The method adds a discount factor $\alpha|C^w_{S_n}|$ in the denominator of the MLE estimator, where α is the Lidstone parameter, S_n is a sequence of n words and $|C^w_{S_n}|$ is the number of different words that appear after the sequence of words. Once we do that, we have an discounted value D_{S_n} (equation 4.4) to be distributed in next level of the backoff. This way, we define a value β (equation 4.5), that is a multiplier for the backoff probabilities into this reserved probability space. Notice that, for computing this value, we simply divide D_{S_n} by the amount not discounted in the following backoff level.

$$D_{S_n} = \frac{|C^w_{S_n}|}{f_{w_i-1,...w_n} + \alpha|C^w_{S_n}|} \tag{4.4}$$

$$\beta = \frac{D_{S_n}}{1 - D_{S_{n-1}}} \tag{4.5}$$

Finally, the general Katz backoff formula has two cases: (1) when the n-gram is in the dataset, we run a discounted probability estimation, (2) otherwise we call the backoff model and move it to the new probability space using the factor β (Equation 4.6).

$$P(w_i|S_{n-1}) = \begin{cases} \frac{f_{w_i}}{f_{w_i,S_{n-1}} + \alpha|C^w_{S_n}|}, & \text{if } w_i, ...w_{i-n} \in L \\ \beta P(w_i|S_{n-2}), & \text{otherwise} \end{cases} \tag{4.6}$$

In the last level (unigram level), we apply the classic Lidstone smoothing. Therefore, in a case where $w_i \notin L$, we would still get a non-null probability of $\hat{P}(w_i) = \alpha/(N + \alpha|C|)$. Notice that this only happens in the last level. In all other levels, Katz backoff prefers to search for a more general version of the n-gram in the following levels than atributing the same probaility to all not found items.

4.2 Context-enriched n-gram language model

n-gram models associated with Katz backoff are efficient methods for estimating the probability of sentences belonging to a language. However, based on Pragmatics theory, we believe that the decision whether a token w_i belongs to a language L does not depend only on the sequence of tokens that occurred before it in the sentence. We believe that context may be as important, or even more important, than this sequence in some cases. Therefore, we need new language models that do not rely solely on text but also on features used by human cognition to interpret sentences. In this section, we propose a novel language model that considers a contextual score as one of the priors in the conditional probability.

In this novel language model, we want to estimate the probability $\hat{P}(w_i|C_j, w_{i-1}, w_{i-2}, w_{i-3}, ..., w_{i-n})$, where C_j is the contextual score of one of the proposed contextual models ($j = \{AKC, SC\}$, where AKC is the agent knowledge context and SC is the situational context). For maintaining the same probability estimation method (i.e. Katz backoff and Lidstone smoothing), we use C_j, a discretized value[1] of

[1] We used three bins of equal size

the contextual score. This way, C_j can be considered as an artificial "word" that represents how likely it is for the message to be related to the subject or, in other words, to belong to the language. It is also important to stress that C_j is a message attribute. Therefore, its value is the same for all n-grams extracted in a single message.

The intuition behind the proposed attribute C_j is that we expect that some n-grams may belong to the language only in the presence of important contextual information. For instance, the trigram "And he scores" is much more likely to be related to the subject "football" when there is a football match going on. In this case, C_j would get a higher score during a football match and the language model would return a much higher value than for a contextualized n-gram. Another good example is when there is ambiguity in the n-gram. For example, if the trigram "NY is terrible" was uttered by a basketball fan, this would likely to be related to the NBA team, not to the city. In this scenario, C_j would be high when j is the AKC score and the language model would recognize that "NY is terrible " is a common trigram in this case. Unfortunately, adding the context into the language model is not trivial. In order to fully integrate it with classic n-gram model we had to modify Katz backoff.

To include context in the backoff process, we added new levels to the language model. The first level is the full contextual n-gram, for which the probability, $\hat{P}(w_i|C_j, w_{i-1}, ..., w_{i-n})$ is estimated. In the following level, we maintain the context C_j and remove the word token w_{i-n} from the prior, therefore we estimate the probability $\hat{P}(w_i|C_j, w_{i-1}, ..., w_{i-n-1})$. The contextual unigram is the last level in the contextual backoff. In this level, we maintain only the C_j in the prior, so the estimated probability is $\hat{P}(w_i|C_j)$. If this contextual unigram still cannot be found in the training set, we run the previous n-gram language model ignoring context.

Notice that, in this solution, we fall back to the traditional n-gram whenever context associated with text fails. However, it is possible to prioritize context when this happens. In order to accomplish that, we compute the probability distribution function (PDF) for the contextual scores and attribute it to the the token when we fail to find the contextual unigram. This way, instead of receiving $\hat{P}(w_i|w_{i-1}, ..., w_{i-n})$, which depends only on text, the backoff would get $PDF^j(C_j)$, which relies only on context.

The whole process is described in Figure 4.2. In this figure, the shadowed part represents the discounted probability space that maps to the following level model. Remember that there are two options once the language model fails to find the contextual n-gram, one is attributing the contextual PDF (context emphasis) and the other backs off to the non-contextual n-gram (text emphasis). Both these options will be evaluated separately in Chapter 6.

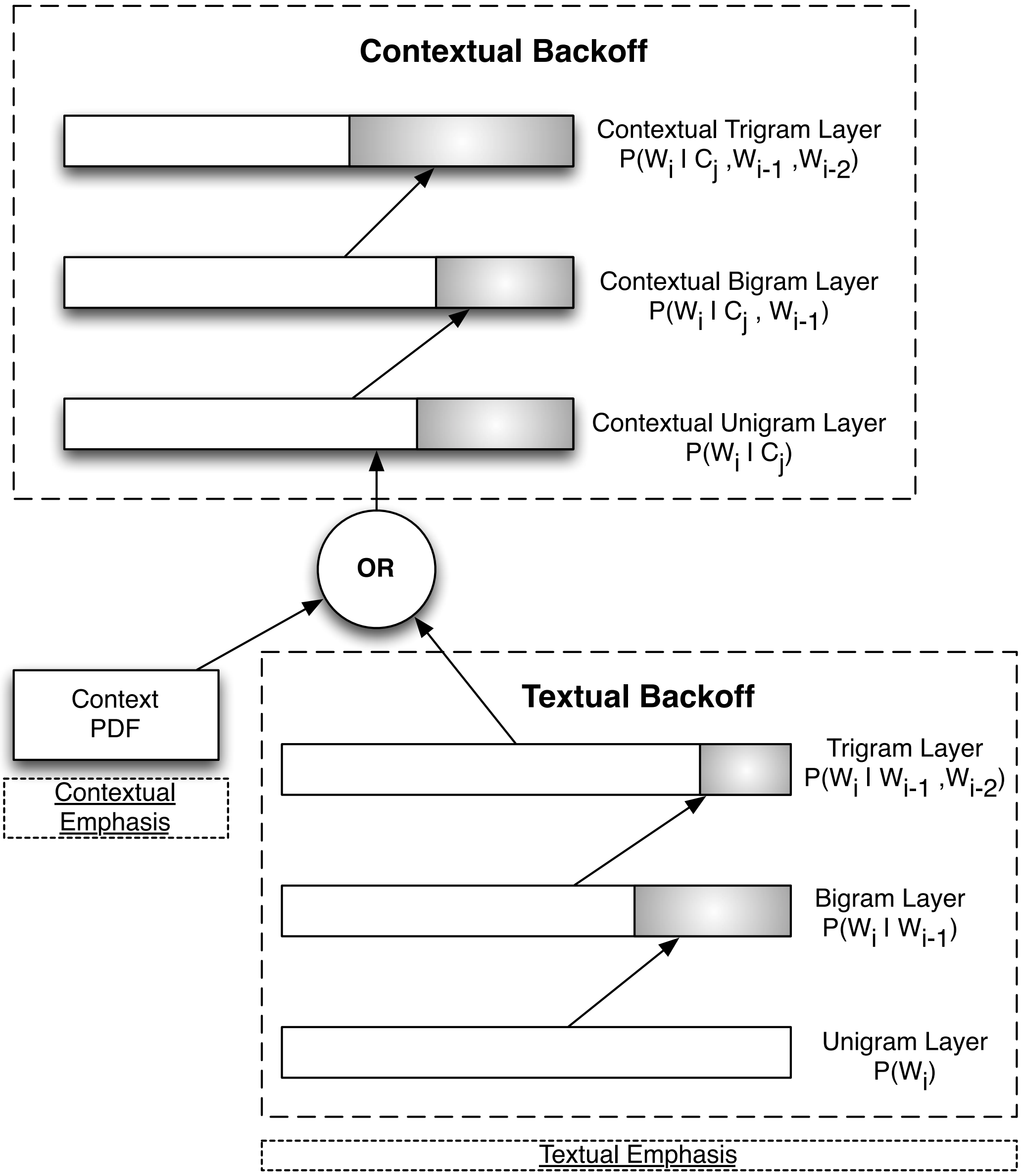

Figure 4.2. Ilustration of the context-enriched language model backoff

Language models are efficient techniques for subject classification when there is no negative label. In this section, we defined methods for using extra-textual information in these models for performance improvement. In the next chapter, we show some characteristics of the dataset and demonstrate some of the defined hypotheses.

Chapter 5

Dataset

For the experimental evaluation of the proposal in this dissertation, a dataset was gathered using Twitter's streaming API. This dataset comprises tweets which could be related to three popular American sports and their leagues: american football (NFL), baseball (MLB) and basketball (NBA). Our goal is to define if a given message references one of these subjects. For that, we define two subsets of messages: training stream and test stream. The training stream is used exclusively for model generation, while the test stream is the one we extract messages for labeling. In the next sessions, we are describe the process of data gathering and some characterizations of the dataset.

5.1 Dataset information

For evaluating our hypotheses and the proposed language model we use three training streams (one for each subject) and one test stream collected using Twitter's streaming API. The training streams were collected by keyword-selection approach. Since the keywords were manually choosen focusing on recall, they may include some ambiguous keywords. Consequently, the training stream may contain tweets that are not related to the subject.

For the test stream, we employed a user-selection approach. The set of users was chosen from each training stream according to the method we presented in the previous chapter. Then, we collected all tweets posted by these users.

All training streams were collected from October 17th, 2013 to November 4th, 2013. As for the test stream, we collected messages from October 27th, 2013 to November 4th, 2013. The user set used for the test stream was generated from tweets gathered from Octber 17th to October 27th. The number of collected tweets can be seen in Table 5.1.

Dataset	Number of Tweets	Period
MLB Training Stream	20M	10/17/2013 - 11/04/2013
NFL Training Stream	16M	10/17/2013 - 11/04/2013
NBA Training Stream	19M	10/17/2013 - 11/04/2013
Test Stream	100K	10/27/2013 - 11/04/2013

Table 5.1. Information about the collected streams

During the test set period, the following events, related to the target subjects, took place:

- **Baseball (MLB):** World Series 2013 - St Louis Cardinals vs. Boston Red Sox - Matches happened on 10/23, 10/24, 10/26, 10/27, 10/28 and 10/30. In the end, the Red Sox won the series and became the 2013 champions.
- **American Football (NFL):** Mid Regular Season 2013 - Weeks 8 and 9 - All regular season matchups on NFL happen on Thursdays, Sundays and Mondays. Therefore, all matches happened on 10/27, 10/28, 10/31, 11/03 and 11/04.
- **Basketball (NBA):** Preseason and early season - Week 1 - There are matches every day since the start of the preseason.

It is expected that characteristics of these events have a major impact on contextual features and on the language model result. We will explore this in the next sections.

Labeling messages from the test dataset was not a trivial task. Since we wanted to improve the recall in messages that do not have textual references to the subject, we could not rely on traditional methods of annotation that only consider the text of the message. Therefore, to annotate a tweet we consider user profile, previous messages, time of post, mentions to other users and other contextual information. Of course, such annotations had to be performed manually, thus limiting the size of the training set.

In the next sections, we show some interesting information about the collected training stream and test stream. Since they were collected using different approaches, we discuss characteristics of these streams separately.

5.2 Training stream

To generate the training stream R, we select all messages that contain at least one keyword $k \in K$ related to the subject s. The set of keywords K was chosen manually (for more information look at the appendice). To increase recall in the training stream, we chose to include many ambiguous keywords in K. Sport teams in the United States are trditionally named after a location and a nickname, for instance, "New York Giants" and "Baltimore Ravens". We chose to include in K, the location and the nickname of all teams separetely (following the last example we would include "New York", "Giants", "Baltimore" and "Ravens") and some other keywords referencing to the sport (such as "touchdown", "nfl" and "field goal"). Notice that some nicknames and all locations are ambiguous. Fortunately, the proposed algorithms are robust towards ambiguity and are able to generate good contextual models even if many tweets in R do not reference subject s, as the experimental results in Chapter 6 show.

5.2.1 Characterization

5.2.1.1 Users

In Chapter 3, we proposed using weights inspired in TF-IDF for modeling the agent knowledge context (AKC). In our model, users are characterized by the set of keywords $k \in K$ used by them in R. The model proposes an analogy, in which users play the role of documents in traditional TF-IDF. In order to use TF-IDF-like weights for the AKC model, we need to show that the distribution of the TF component is similar to the IDF one [Baeza-Yates and Ribeiro-Neto, 1999]. It is intuitive to think that if one of them grows faster than the other, one term would be dominant. Figure 5.1 shows that this is not case in our database. In this figure, we plot the TF and IDF values for each keyword. We can see that many words have similar TF and IDF values in the center portion of the graph. This happens because of the similarity between keywords (most are team names and nicknames). The words with high IDF and low TF are the least ambiguous ones. We show an example of those in Table 5.2.

In Figure 5.2, we plotted the TF-IDF weights for each keyword as a function of TF. This plot has a similar pattern to TF-IDF of terms in Web documents [Baeza-Yates and Ribeiro-Neto, 1999], in which the words that are given best TF-IDF scores are those with average TF. On Web documents, however, there is a concentration of terms with high frequency and low IDF that cannot be observed in our dataset because we never use stopwords in the K set. Therefore, we can see a low concentration of points in the right end of the horizontal axis. It is interesting to notice that this technique was able to give higher scores to keywords that are less ambiguous while giving lower score to more ambiguous words. For instance, terms such as "cheeseheads" and "diamondbacks" are almost exclusively used to refer indirectly to teams in NFL and MLB. Meanwhile, terms such as "NY" and "SF" are a lot more ambiguous. In Table 5.2 there are some

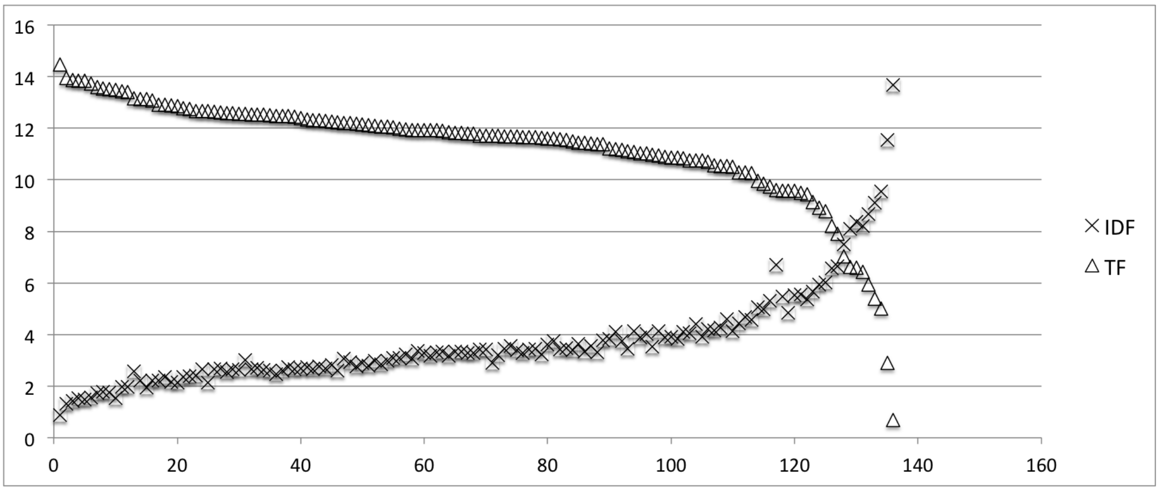

Figure 5.1. Score of TF and IDF for each keyword ordered by the reverse TF order

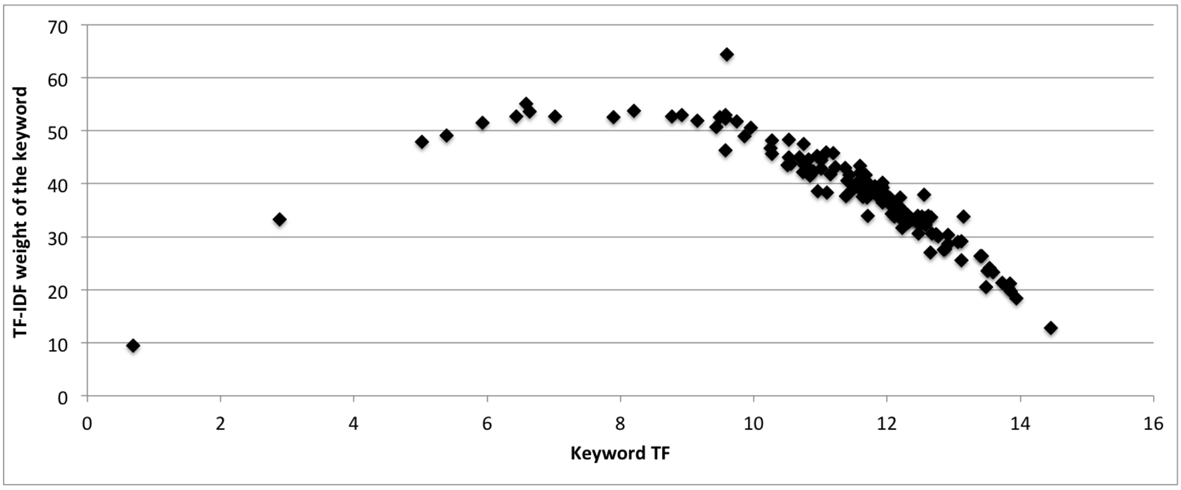

Figure 5.2. Sum of all TF of a keyword by TF-IDF of that word

examples of keywords that can be considered ambiguous and unambiguous and their IDF score. Moreover, both Figure 5.1 and Table 5.2 are empirical demonstrations of Hypothesis 3.2, which posits that the ambiguity degree of a word is related with its frequency.

As expected, users behave differently in the training stream. One of the major differences is the number of keywords used during the analyzed period. By grouping the users according to the variety of keywords used in the training stream we noticed that the size of these groups follows a power law. Therefore, few users post a wider variety of keywords while the majority used only a few.

Unambiguous Keywords		Ambiguous Keywords	
Keyword	**IDF**	**Keyword**	**IDF**
cheesehead	52.51	SF	27.50
nyj	44.98	miami	21.19
white sox	52.75	boston	20.29
SF giants	52.70	NY	12.76
diamondbacks	53.76	new york	21.27

Table 5.2. Examples of ambiguous and unambiguous keywords

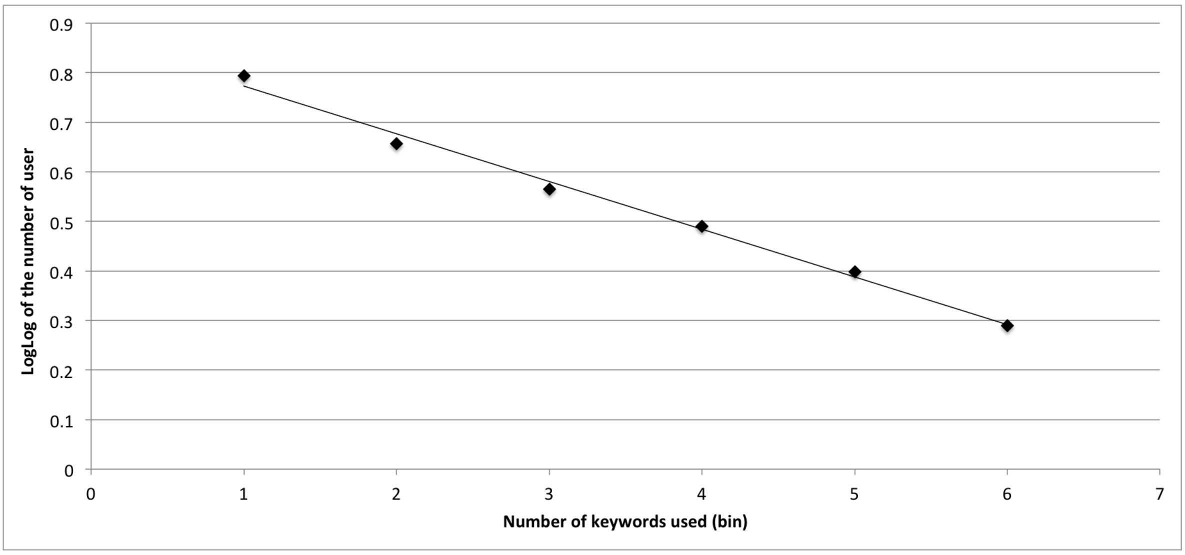

Figure 5.3. LogLog of the number of users in each bin according to the keywords used by them

5.2.1.2 Time Frame

For estimating the likelihood of a message to be related to a subject during an arbitrarily defined frame of one hour, we consider the frequency of keywords posted during that timeframe. By Hypothesis 3.3, we believe that this likelihood is related to the amount of messages containing keywords posted during the timeframe. For penalizing ambiguous keywords, our proposed model uses the IDF score computed in the AKC model.

In Figure 5.4, we plot the frequency for some keywords in each 1-hour interval (from Oct 28 to Nov 3) in the baseball dataset. It is easy to see that the keyword “NY” , which has a low IDF, only brings noise to the model. Words that are more related to the finals (“red sox” and “cardinal”) clearly show two big peaks of frequency and a smaller one. These big peaks occurs exactly during the last two matches of

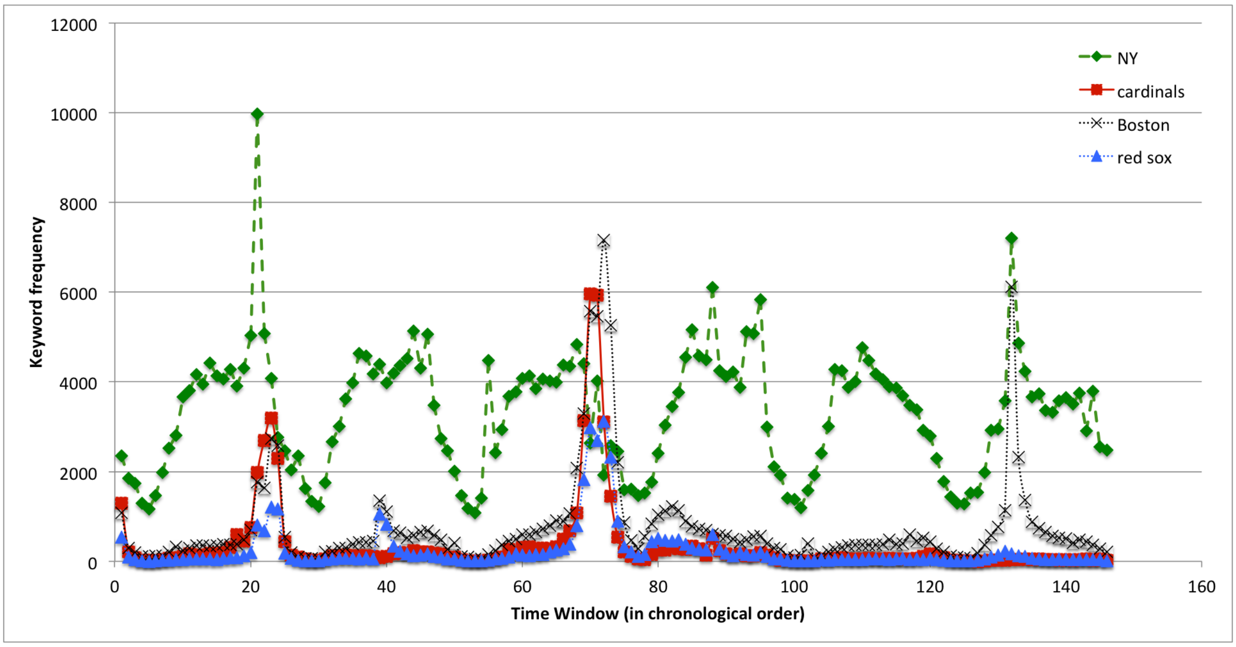

Figure 5.4. Frequency of keywords in each one-hour time window of the training stream

the World Series. Therefore, this is a good example of Hypothesis 3.3: the increase in the frequency of these unambiguous keywords shows that there is an important event happening by that time. On the other hand we can see an ambiguous keyword ("boston") that followed the event peaks, but also had a really big peak on the last frames. Since "boston" has a low IDF, this probably offtopic peak does not have a major impact in the proposed model.

Figure 5.5 shows the Situational Context value (defined in Equation 3.5) for each time frame in the MLB dataset. It is easy to see that it followed the peak pattern from Figure 5.4, and also that noisy ambiguous words such as "NY" did not reduce the quality of the model. It is interesting to notice also that the higher values seen in this figure occur in late night periods with low traffic of messages in the United States. This is a good result, since we do not expect to have many messages related to the topic late at night. Another interesting observation is that our proposed situational context modeling is only sensitive for messages posted during major events. Therefore, when there is nothing relevant happening, we need to rely in other forms of context.

5.3 Test stream

By our main idea in this work, in a message we may only have an implicit reference to a subject. Therefore, simple keyword selection may get only a small subset of messages

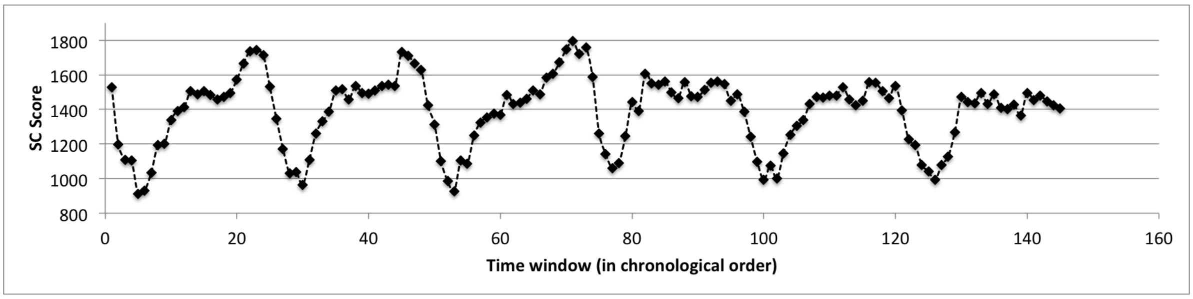

Figure 5.5. Situational Context score in each time window of the training stream

that reference the subjects. It is important for our purposes to have messages that make both kinds of references (implicit and explicit) to s in the test stream, since one of the major objectives of this dissertation is to improve the recall in messages with implicit references to the subject. To collect the test stream we chose users that posted at least one message in the training stream.

Sampling the users to be collected is not an easy task. As shown in the previous section, the estimated interest degree of a user towards a subject follows a power law. Consequently, if we simply select users at random in the training stream, it is likely that we will only get users that are not really interested in the subject. This would result in a test stream with too few positive examples. For solving this problem, we discretized our user set in 10 buckets with different sizes, according to their AKC context. Then we sampled the same amount of users in each of them. IWe collected all tweets posted by a set of 300 sampled users.

Despite of the fact that we used different methods for collecting messages in each stream, we expect that all messages mentioning s implicitly or explicitly are under the same context and the same pragmatics effects. Consequently, we assume that using only messages that contain explicit references to s for training does not affect the perfomance of contextual models.

5.3.1 Characterization

One of the most important characteristics of the test stream is the presence of messages that are related to a given subject without having any of the keywords used for generating the training stream. In Table 5.3, we can see that more than half of the tweets that reference a subject, according to human annotators, did not include any of the keywords in K. This information shows us that relying on simple bag-of-words approaches for extracting subject-related messages may lead to low recall. Despite of that, we believe that messages that do not contain keywords are under the same

Subject	Positive Examples	% without keywords	% positives
Football	16	56.3	2.0
Baseball	14	71.4	2.0
Basketball	72	47.2	10.0

Table 5.3. Positive examples in the test stream

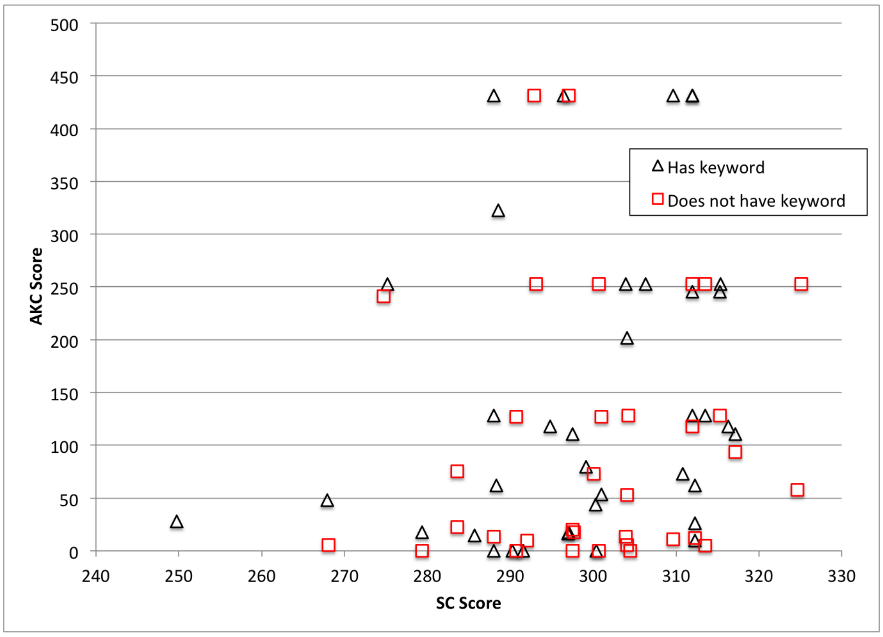

Figure 5.6. Scatter plot showing the differences between AKC and SC

context.

Figure 5.6 shows a scatter plot in which each point is a tweet, annotated as being related to a subject. The horizontal axis is the *SC* score of the message, while the vertical axis is the *AKC* score of the message. One important observation we can get is that there is no concentration of points with or without keywords in any region of the graph. This way we can conclude that, for our dataset, the contextual scores are independent from the presence of keywords. Another interesting side conclusion we can draw from this graph is that *SC* and *AKC* scores do not seem to be correlated.

In this chapter, we empirically demonstrated many hypotheses that we formulated throughout this dissertation. In the next chapter, we show the experimental results for the proposed language model and demonstrate the remaining hypotheses.

Chapter 6

Results

In this chapter, we discuss the results of our proposed contextual-enriched language model for our three datasets. In these experiments, we add context scores for each tweet in both training and test stream according to our models. Once we have annotated all messages with contextual scores, we generate the language model using the training stream and then we evaluate it with messages in the test stream.

In this chapter, we denote messages referring to the target subject as positive examples, while the unrelated ones as negative. In the following sections, we discuss the evaluation method, and present experimental results.

6.0.2 Evaluation Workflow

For evaluating both the context-enriched language model and the pragmatic contextual models, we use the pipeline represented in Figure 6.1. In this pipeline, we start with two separate datasets: training stream and test stream. As we argue in the Chapter 5, training stream is generated by keyword selection of messages while test stream is generated by all messages posted by a predefined set of users. The training stream is then used for generating the contextual model, Agent Knowledge or Situational. Once the model is generated, we attribute the contextual score for each message in both training and test stream.

Then we sample examples fron the contextual-enriched training stream to generate the proposed language model. This sampling proccess is also being detailed in the next chapter. With the model generated we use the language model to give a score to each message in the contextual-enriched test stream. Once we have that score, we can run the following evaluation method to measure the effectiveness of our technique.

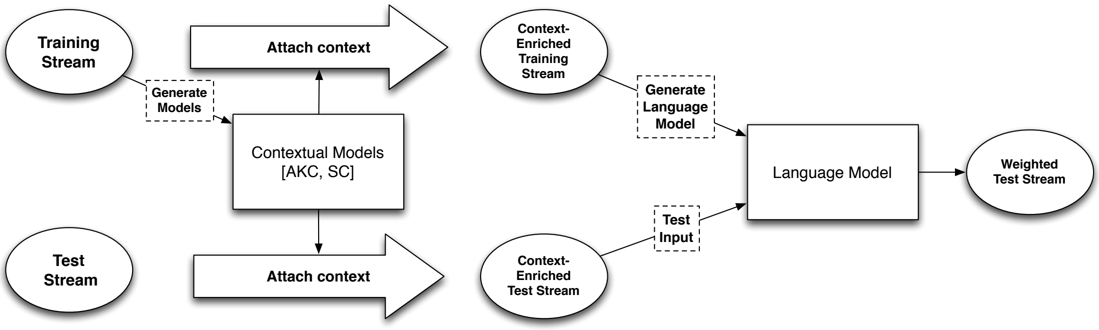

Figure 6.1. Ilustration of the evaluation pipeline

6.1 Evaluation Method

For evaluating the language model we used a test stream being all messages posted by a subset of users, as explained in the previous chapter. Unfortunately, positive examples are scarce on the test stream (Table 5.3). Since less than 10% of the messages are positive examples, we use a ranking evaluation method [Cremonesi et al., 2010].

The evaluation method assumes that most of the unlabeled examples are negative. Given this assumption, we sample 100 unlabeled examples into a probe set. It is expected that most of those 100 examples are negative ones. Then, we add a positive example from the annotated ones into the probe set. Next, we rank all these examples according to the proposed model and baseline scores. Once this ranking is complete, we expect the positive example to figure at one of the top positions. We run this procedure 200 times for each positive example. Finally, we plot the average number of positives found up to the nth position. Since we want to improve recall, our goal is to have a higher percentage of examples seen for every n positions. In other words, we expect the line for the proposed method to be always over the baseline.

6.2 Language model evaluation

Language models require a set of sentences (in our case, Twitter messages plus context) for training. It is expected that in the training stream we have many messages unrelated to the subject because of ambiguous keywords. To address this issue, we selected messages that are more likely to be positive in the training stream according to their contextual values and used them in model generation. In othere words, we selected only the top two thirds of messages according to the Contextual Models and used them for language model generation. In the baseline, we randomly picked examples in the training stream so that we would get the same number of messages as in this sample.

Therefore, in both training generation approaches we use the same number of messages to generate the language model.

For all experiments, we discretized the contextual scores given by the models in three bins that contain the same number of examples. We found out that dividing the dataset into more bins leads to worse results, since it fragments the training set.

We ran the experimental evaluation on five different methods. Each one uses context in a different way. The evaluated methods follow:

- **Baseline-n-gram (NG):** For this method we used the classic n-gram model generated with randomly picked examples from the training stream. Therefore, this method *completely ignores context* and *focuses only in the written message.*

- **Baseline-n-gram + contextual training (NGCT):** For this method we used the classic n-gram model generated with messages that were selected by their contextual values (the top two thirds of the messages, according to their contextual scores). This method *ignores context for model generation.* However, the training examples are supposedly better chosen than the previous method.

- **Context-enriched language model with textual emphasis (CELM-TE):** For this method we used the proposed context-enriched language model generated with messages that were selected by their contextual values (the top two thirds of the messages, according to their contextual scores). If there is not a contextual n-gram in a test example, this method backs off to the tradional n-gram and checks if it can be found there. Moreover, this method *only ignores context when contextual backoff fails.*

- **Context-enriched language model with contextual emphasis (CELM-CE):** For this method we used the proposed context-enriched language model generated with messages that were selected by their contextual values (the top two thirds of the messages, according to their contextual scores). If there is not a contextual n-gram in a test example, this method backs off to the raw contextual value. Therefore, this method *considers textual information only initialy* if associated with the context value.

- **Contextual Only:** For this method we ignore all message text. We *consider only the context value of the message.*

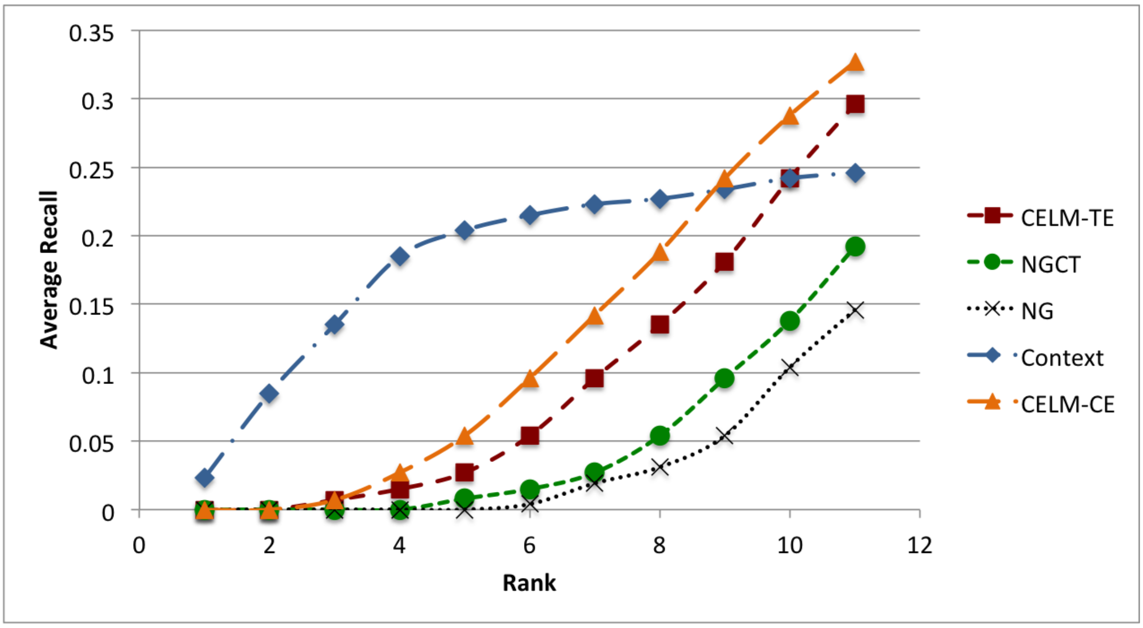

Figure 6.2. American Football (AKC): average recall by ranking position considering all positive examples that do not contain keywords

6.3 American football data

American football is the most influential of the American sports. With seasons scheduled from early September to late February, American football is known to have the most avid fan base. In the following experiments we test how important each context is for message understanding.

6.3.1 Agent Knowledge Context

Since the NFL football championship, during the analyzed period, was in mid-season, we believed that most of the tweets would be posted by fans and interested people. Moreover, we believe that in playoffs and finals (Superbowl), the championship may attract the attention of people that usually are not interested in this sport.

As expected, we can see that context is really important, especially for messages without keywords (Figure 6.2). In those cases, it is interesting to see that when we consider only context we get higher recall in the first positions. This means that there are some messages that have a high AKC score and bad textual information. Also, notice that the context-only approach stabilizes after 0.2 and from that point on the context-enriched language models perform better. We can conclude from this observation that around 20% of the messages texts were poor and relying only on context leads to improved results. Also, we observe that in most of the messages a

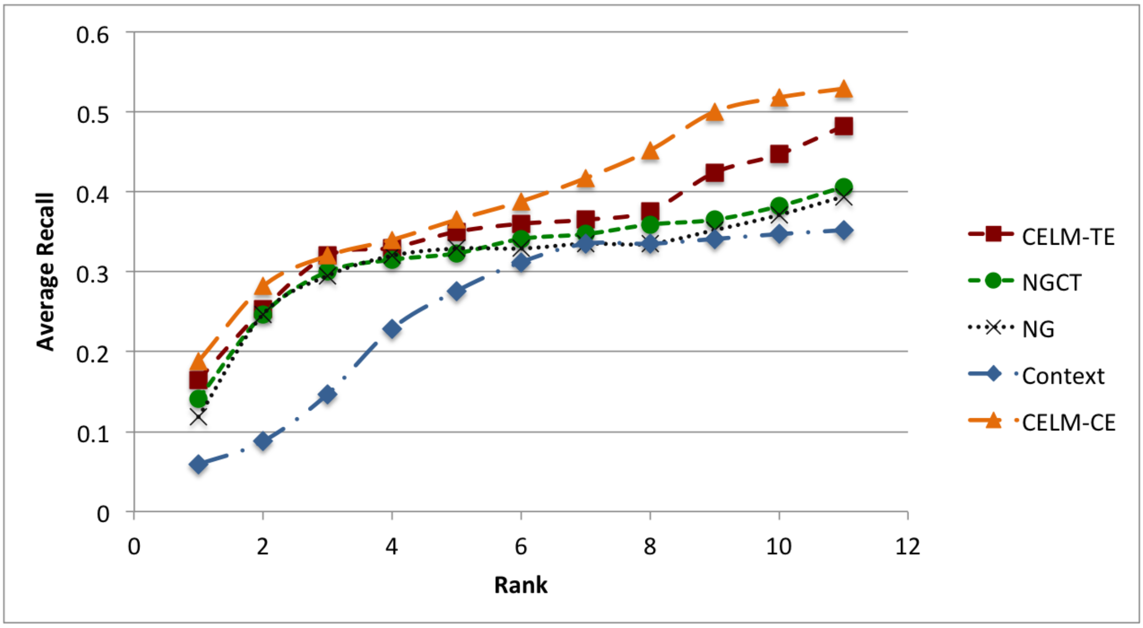

Figure 6.3. American Football (AKC): average recall by ranking position considering all positive examples

combination of context with written message provides better recall.

Analyzing results from the dataset with all positive messages (with and without keywords, Figure 6.3) we can see a slight improvement of context-enriched language models over the baseline. Messages that already contain keywords are more likely to have strong text related to the subject. Therefore, contextual models add more information in this case. Since our goal is to increase the recall of messages, this result is less important than the last one, because, realistically, we would already have all messages that contain a given set of keywords. However, it is good to see that we do not have a worse result than the baseline even when text is supposedly strong.

6.3.2 Situational Context

All matches in the NFL championship are traditionally concentrated in a few days of the week (Thursday, Sunday and Monday). However, during the analyzed period we did not have any highly localized temporal event such as the Superbowl, playoff matches, or important news. Therefore we expected a result slightly worse than *AKC* Context.

The results, shown in Figures 6.4 and 6.5, followed our expectations. The recall for messages without keywords follows a pattern similar to the *AKC*. However, we got better recall ratios at rank 10, which means that the Situational Context is slightly

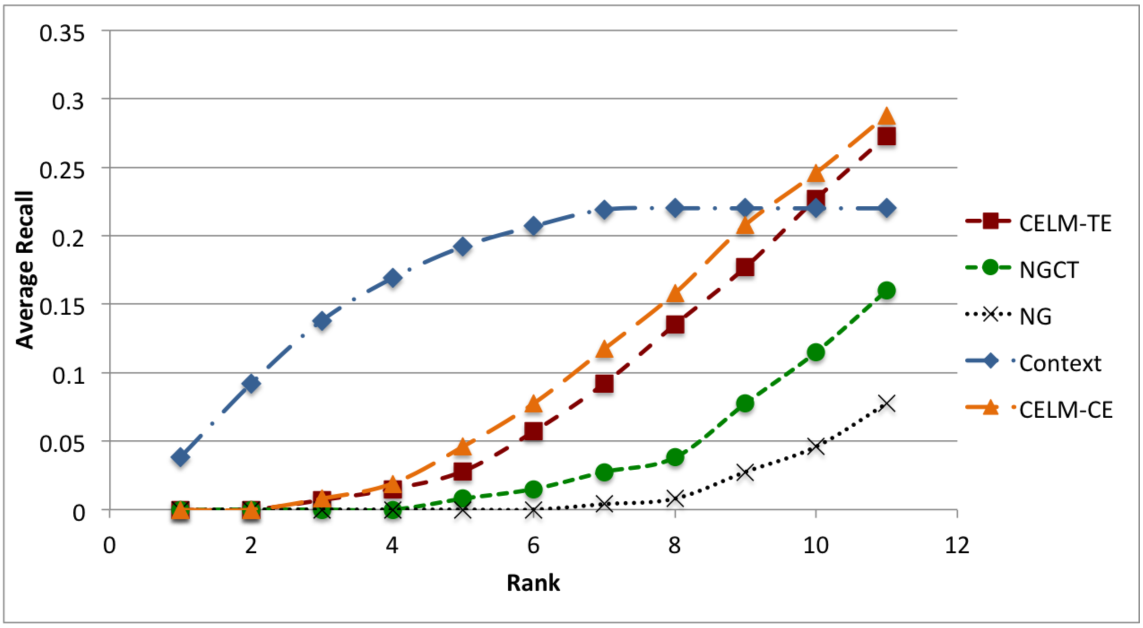

Figure 6.4. American Football (SC): average recall by ranking position considering all positive examples that do not contain keywords

better than the *AKC* for the NFL during the analyzed period. This is an interesting conclusion that leads us to think that we can not rely on a single contextual source all the time. Since human cognition usually chooses between several context sources in order to decode the actual meaning of an utterance, we cannot assume that our simplified contextual model is able to make definitive decisions considering a single one.

For the messages with and without keywords (Figure 6.5). We can see that contextual information improved recall about 5% over the baseline. Once again, messages with keywords have better textual information, reducing the recall improvement in the proposed language model. Despite that, we can see that the contextual model with context backoff achieved better results. This is because messages with good textual results were already retrieved in the first positions.

6.3.3 Summary

Another important observation from this result is that the top messages retrieved considering only *SC* and *AKC* are different one from the other. This shows the complementariness of the proposed model. The top examples in *SC* were clearly related to matches, for instance "Golden Tate is THAT GUY!!", while *AKC* examples are more related to news and NFL-related histories, such as "Did LZ just correct AC again?

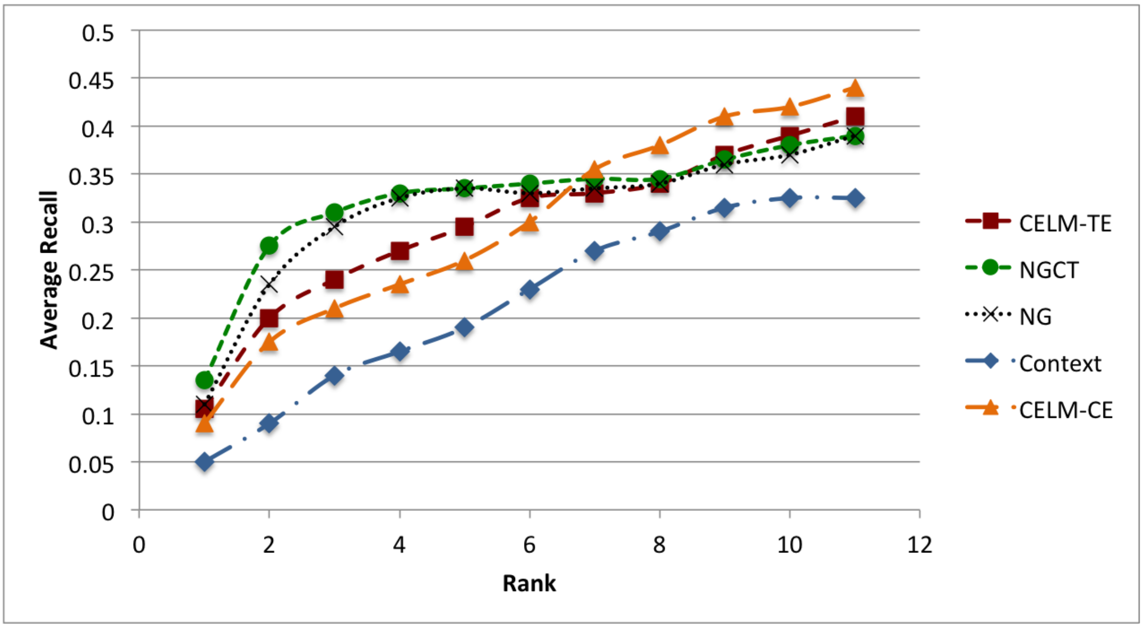

Figure 6.5. American Football (SC): average recall by ranking position considering all positive examples

"The old Robert Horry v Peyton Manning" comparison".

A conclusion that we can draw from this set of experiments is that there are some examples in the dataset that have really strong text. This commonly happens when we have a tweet that was retweeted many times. When this happens, we have good recall in the baseline for the first positions of the ranking (as seen on Figure 6.5). However, the contextual model is more general and gets better results for the following positions.

6.4 Baseball data

Baseball is a really important sport in the U.S., especially because of the time of the year in which the season happens. The MLB championship occurs mostly during summer, and for the majority of the season it does not compete with other American sports. Since a team plays more than 150 matches during the season, the importance of a single match event is low. However, the finals (World Series) happened during the analyzed period, and these were the most important matches of the season. We explore the impacts of context during this specific period in the next results.

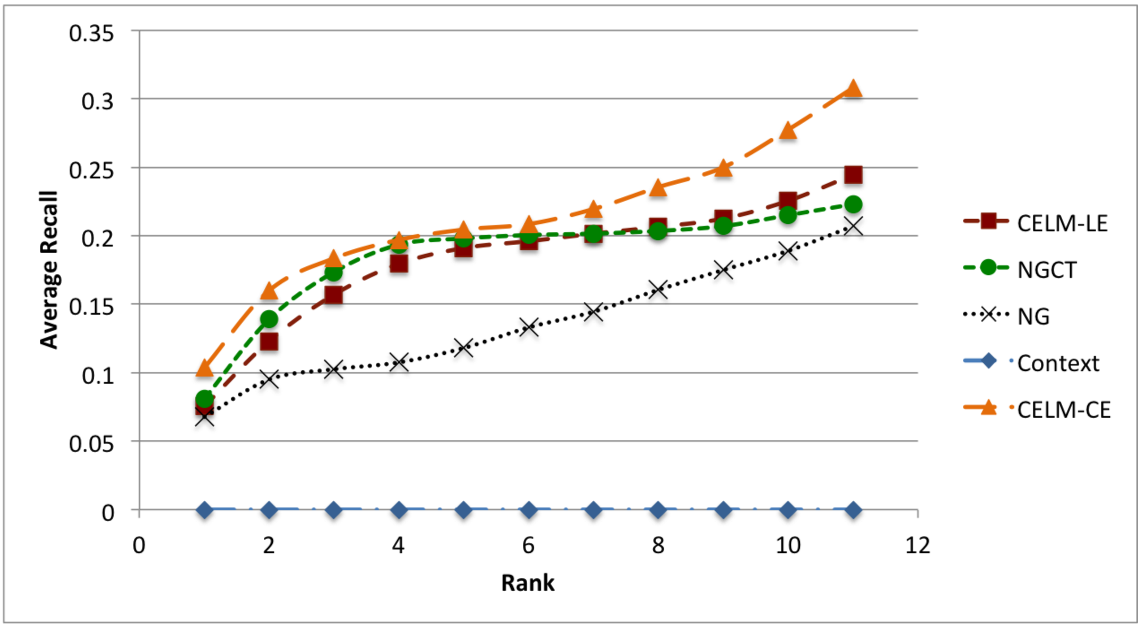

Figure 6.6. Baseball (AKC): average recall by ranking position considering all positive examples that do not contain keywords

6.4.1 Agent Knowledge Context

During the finals, it is common for people who are not usually interested in baseball to post messages about the subject. Therefore, many positive examples were posted by users with average to low AKC. This generates low recall for the context-exclusive model. However, we can see that context-enriched language models achieved good results (Figures 6.6 and 6.7) with recall ratios 8% to 10% better than the baseline at position 10.

The poor performance of the context-exclusive model and the good results of the context-enriched language model means that the activity of users with average AKC actually helped the proposed language model to get better results. Therefore, the language used by those users was actually a good representation for subject related messages, in that moment. It is interesting to notice that despite many positive examples that were posted by non-authorities, those users still had an important role for improving in our context-enriched language model.

6.4.2 Situational Context

Since the finals are the most important event of the year, we expected that the Situational Context would perform especially well. The results have shown in (Figures 6.8 and 6.9) we would get a higher recall at rank 10 if we considered only the contextual

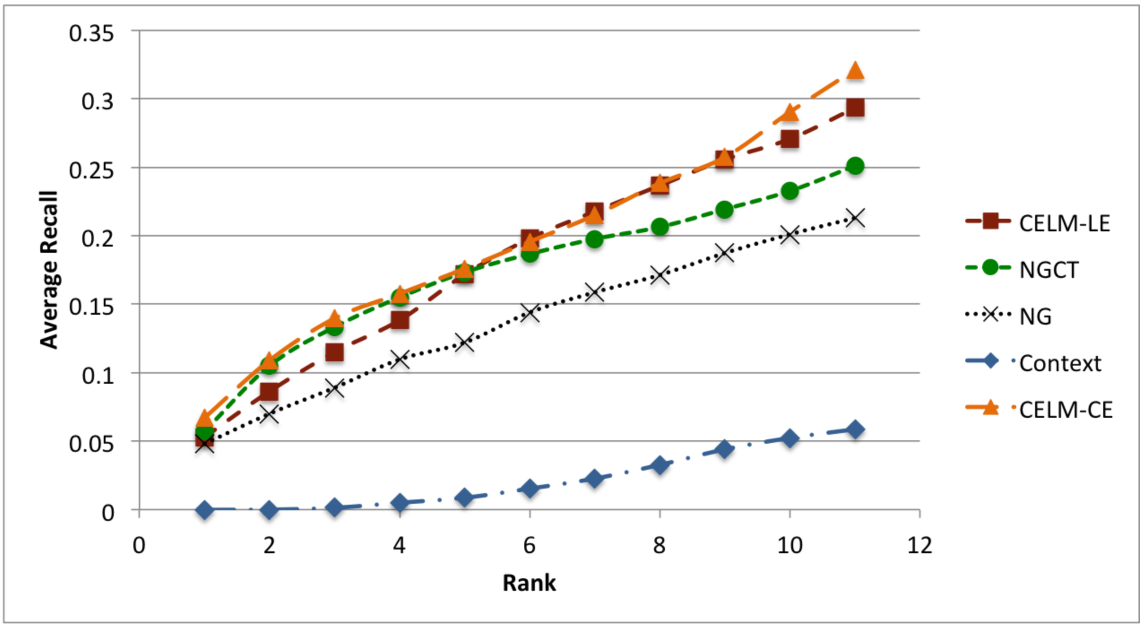

Figure 6.7. Baseball (AKC): average recall by ranking position considering all positive examples

information. It is interesting to notice that in cases in which we have a really strong context, text may reduce the model's performance. Still, we notice that in Figure 6.9 the proposed model still improves the baseline by 5%.

Another interesting observation is that the contextual-exclusive model gets low recall in the first positions. This can be explained by the number of posts during the finals that were not related at all with baseball. Despite the importance of the event, there are still many posts that are unrelated to the subject during the match. Those messages may frequently get the top positions.

6.4.3 Summary

The baseball dataset had the unique characteristic of being gathered during the finals. This enabled us to analyze the results in cases where there is a very strong situational context. A large number of people were interested in the final matches and commented about the outcome right after each one was over. This generated an overly strong situational context, and in this case the text actually reduced the performance of the proposed language model. In this cenario it would be more interesting to have a model that gave less importance to the text.

Another interesting observation from these experiments is that even when authorities are not the only ones posting messages related to the subject, we can still get

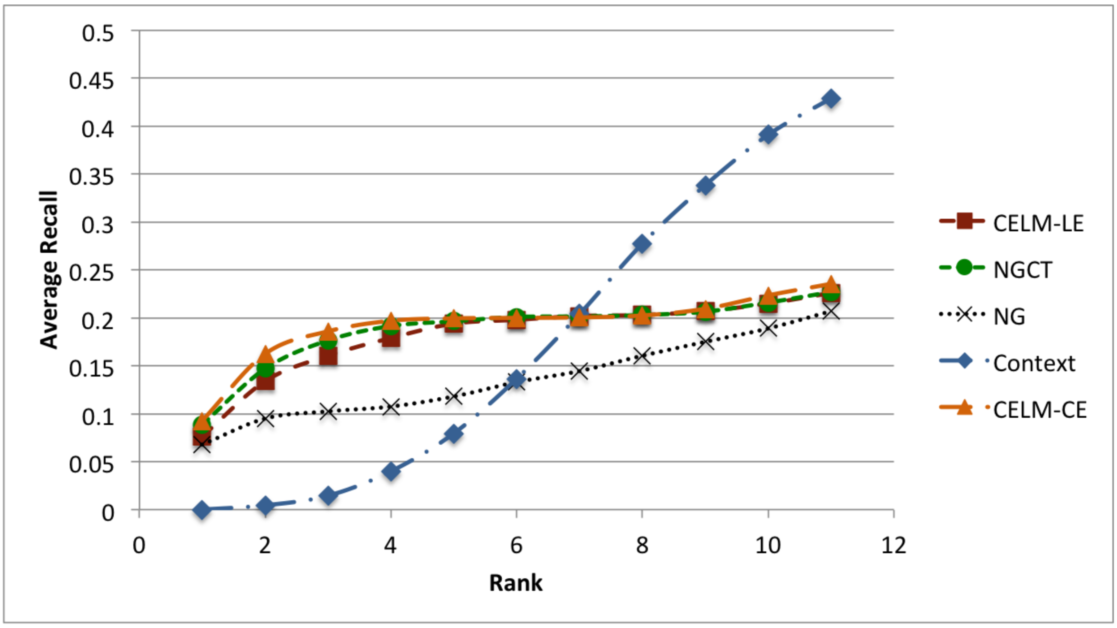

Figure 6.8. Baseball (SC): average recall by ranking position considering all positive examples that do not contain keywords

good recall improvements from context-enriched language models. This happens because the vocabulary of all users related to the subject was similar. Therefore, learning from users with average AKC generated a good language model. A good example of the kind of language that was learned by the contextual model is in this tweet: "RT BrianCostaWSJ: theyre pulling tonight's Silverado Strong promotion.". The MLB officials decided to veto the polemic advertisement "Silverado Strong", which played with Red Soxs slogan "Boston Strong". Since many users with average and high AKC were commenting on this decision, the language model worked well despite that this specific message was not posted by a user with high AKC.

6.5 Basketball data

Basketball is one of the most popular American sports outside of the U.S. It also is known to have the highest engagement levels in Twitter among all American sports [1]. It is not only the fans activity that is famous in NBA, the league is also known for having matches every day. All these factors combined allowed us to have more messages related to basketball than to all other sports, even when the analyzed period is mostly during preseason.

[1] http://mashable.com/2013/04/25/nestivity-engaged-brands/

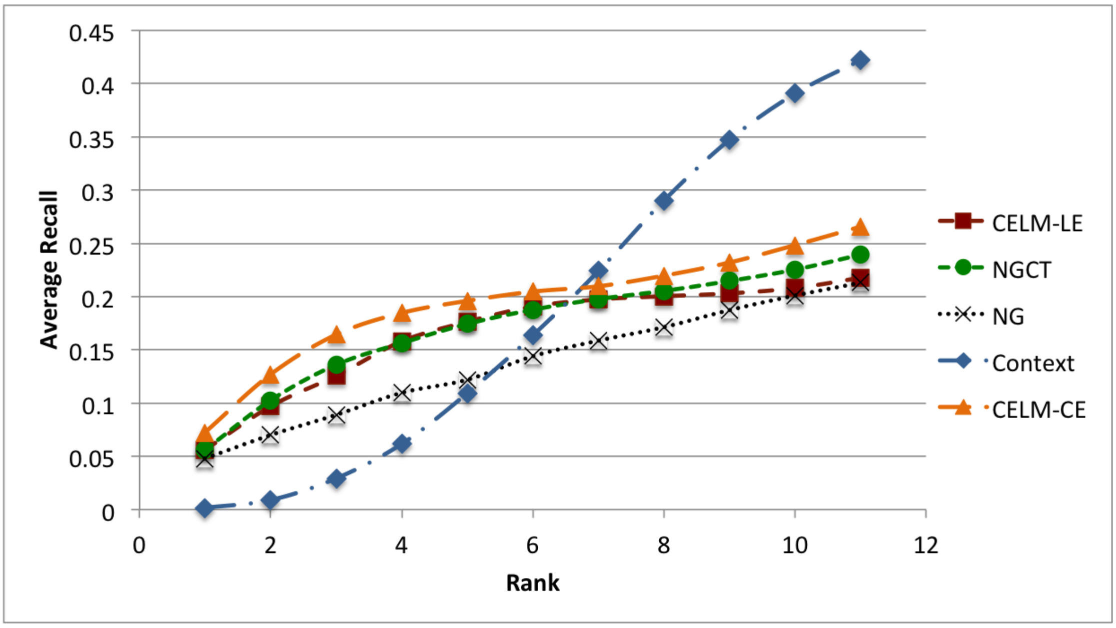

Figure 6.9. Baseball (SC): average recall by ranking position considering all positive examples

6.5.1 Agent Knowledge Context

The NBA, during the analyzed period, was under the preseason. During this period matches do not affect the final season score, they are mostly for practice. Consequently, we expected that most of the messages related to this subject in our test stream would be posted by users that are interested in the subject with high AKC. This is exactly what happened in both scenarios. In the one without keywords, we managed to achieve an improvement of over 13% in recall. In the other one, we had a more modest improvent of only 4%, because of the same reason, stated previously: messages with keywords have strong textual evidence that help baseline language models.

The exclusively contextual model, however, had bad results. Therefore, we can assume that we had a reasonable activity of non-authority users. This leads us to a conclusion similar to the one we had in the baseball dataset: even when context alone does not generate good recall, we get better results in our language models, which tend to be more general.

Figure 6.10, shows that, on the first ranking positions, all methods are equivalent. This happens because of the results with strong textual information. In those cases, the context, even if it is strong, adds few information for the top ranked messages.

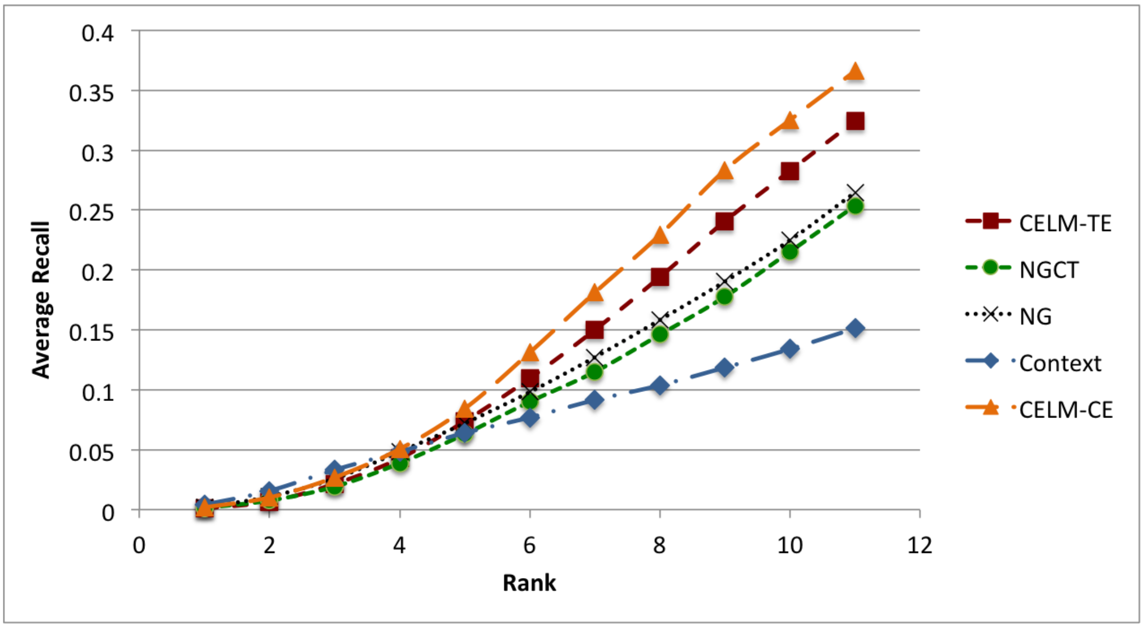

Figure 6.10. Basketball (AKC): average recall by ranking position considering all positive examples that do not contain keywords

6.5.2 Situational Context

Since we have many unimportant NBA matches at each of the days analyzed, we expected a poor result of contextual models for this dataset. Results show that the situational context does not add a significant contribution to the n-gram baseline. When we consider all positively labeled messages, we can see that the addition of context in this case only increased the recall at rank 10 by 3%. In the case where we consider only positive examples without keywords (Figure 6.13), we can see that the n-gram baseline got the second best result. This means that context was not useful. On the contrary, context even reduced the performance of the model.

One interesting observation in both Figures 6.12 and 6.13, is that these are the only cases in which the contextual enriched n-gram with textual backoff got a better result than the one with contextual backoff, showing that the extensive use of context, in this case, reduces the performance of the language. It is also important to emphasize that even with bad contextual information the proposed model was better than the baseline. This shows the robustness of the model and of the backoff technique.

In Figure 6.13, the context-enriched n-gram with textual backoff gives a better recall ratio up to rank 9. After that, there is a draw with one of the baselines. It is an interesting conclusion that the context-enriched language models always get better results in the first positions, showing that they have the expected behavior of boosting the likelihood of messages that were uttered under higher contextual scores. This shows

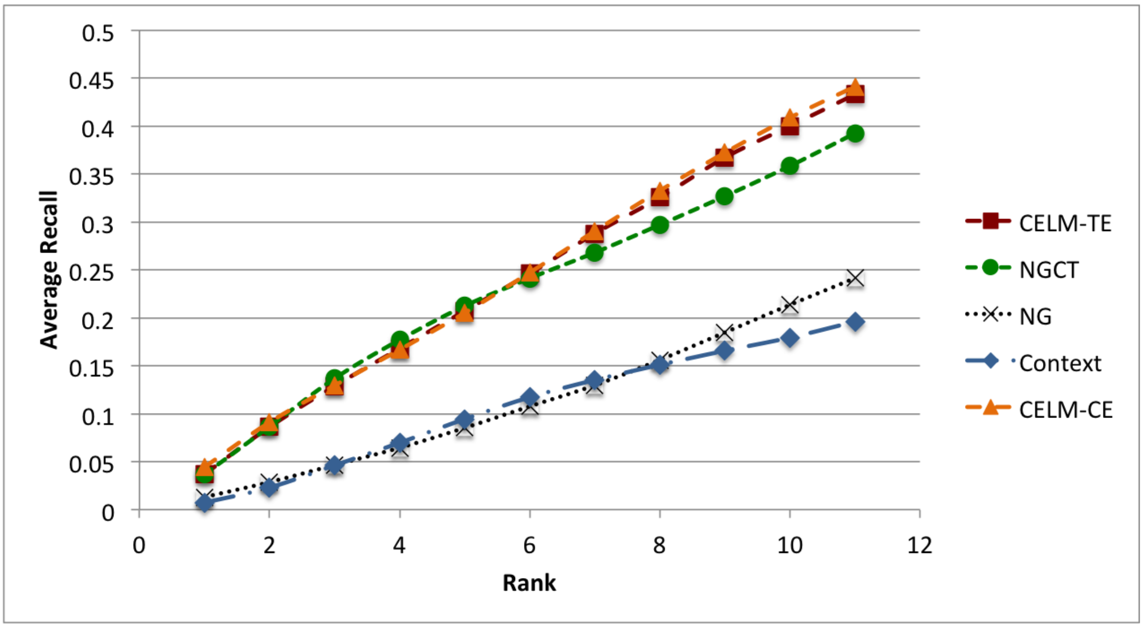

Figure 6.11. Basketball (AKC): average recall by ranking position considering all positive examples

that our contextual models were actually able to identify messages that have a clear conversational implicature that indicates the reference to the message.

6.5.3 Summary

With a weak situational context and a stronger agent knowledge influence, the basketball dataset demonstrated that combining different contexts is an important future work, since the models perform differently in each dataset. However, despite of being worse, in both cases the context-enriched language model performed better than the baseline, demonstrating that even with a weak information gain, context is still a valuable information.

For the agent knowledge context, we achieved a big performance gain over the baseline, retrieving 13% more examples. This is an improvement of about 50% and can be explained by the characteristics of the dataset. Early in the season, the only people who post messages are passionate about this sport and are interested in the event. These results show us how context influences the way people communicate in social media.

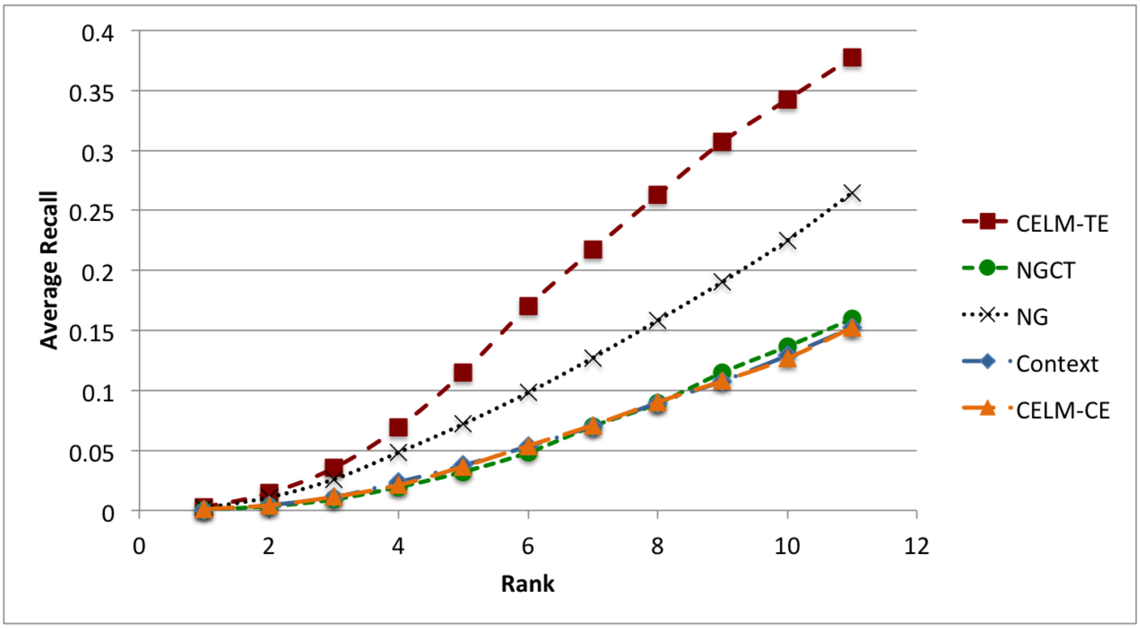

Figure 6.12. Basketball (SC): Average recall by ranking position considering all positive examples that do not contain keywords

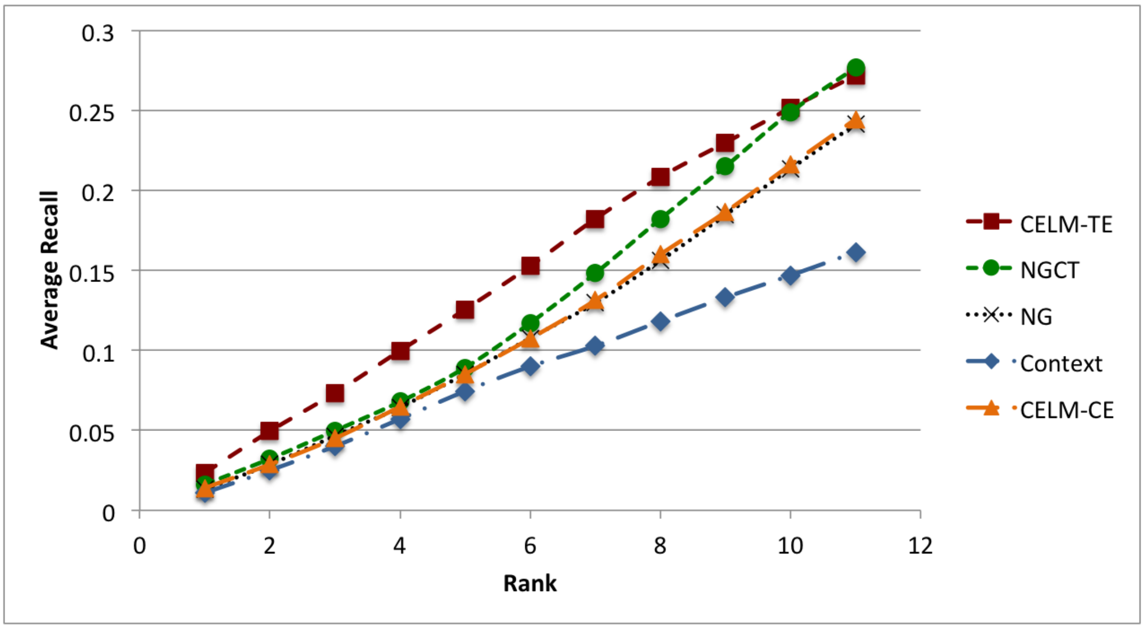

Figure 6.13. Basketball (SC): Average recall by ranking position considering all positive examples

Chapter 7

Conclusions and Future Work

In this dissertation, we argued that considering only text is not enough for social media message comprehension. According to Pragmatics theory, speakers expect that their receivers consider contextual, or extra-textual, information for decoding and understanding messages. Given this assumption, speakers tend to omit from the uttered message information that they expect their receivers already hold. We showed that this implicit method of compression and channel optimization in human communication cannot be overlooked in Twitter messages, since about 70% of subject-related messages in our test stream do not contain any keywords trivially related to the subject.

To address the insufficiency of text in the analyzed scenario, we proposed models based on TF-IDF for attributing scores to the relationship between context and target subject. Our models focus on two different extra-textual elements: the degree of interest of a user towards a subject (Agent Knowledge) and the likelihood that an important event related to that subject is happening (Situational). We demostrated that these models are independent and that messages with and without keywords have similar contextual scores. These models had good performance, especially when there is a strong external signal, such as the MLB World Series matches. However, in many cases, they needed to be combined with text to achieve better results.

In order to simultaneously use text and pragmatic context, we proposed a novel language model that considers both the scores computed by our contextual techniques and the words in the message. We proposed two strategies to be used when a word cannot be found within the given contextual level: (1) ignore all extra-textual information and (2) use exclusively them. The results show that usually the second option achieves better results, except when we have very weak contextual information, such as in the basketball situational context. We also concluded that the performance of the proposed pragmatics models is not constant, as their results improve in the presence of real world events. Therefore, exploring this variation of importance in pragmatics discourse elements constitutes an interesting future work.

The contributions in this dissertation are not restricted to the pragmatics and language models. We also developed a novel framework for analyzing Tweets in a non-keyword driven retrieval approach. In this framework, we propose a new method for collecting messages without keywords in Twitter, given all API constraints. We also show the importance of not considering exclusively text in the labeling process. It is required for the annotator to check the profile, past messages and understand about the subject before labeling the message. Finally, we demonstrate the validity of our hypothesis that both messages selected by keyword and messages selected by user are under the same pragmatic context influence. This hypothesis opens the possibility to use two different streams for training and testing, which is an important perspective in this new retrieval approach.

7.1 Future Work

The verification of the high importance of non-textual elements in social media leads to a deterioration of the performance of classical keyword retrieval approaches. One important question that we need to answer in the future is how to evolve this classic retrieval model to one that is closer to the human intrisic cognition ability to interpret text. To reach this goal, we need to identify other relevant sources of contextual information in this form of communication. Then, another future work would be to create methods of combining all contextual information and text to improve recall. Finally, there is the need of adapting these techniques to work on a stream for creating a continuous subject-related messages source.

Indirectly, this work affects all techniques that rely solely on text for doing retreival and/or classifying social media messages. Since messages that contain keywords may be written differently from those that do not, we believe that techniques which use only text may get biased results. We believe that all techniques of sentiment analysis and trend detection in social media need to use contextual elements for retrieval and for their methods, otherwise they are falling in the pitfall of analyzing just an small subset of messages related with the desired target. Therefore, another possible future work is to create a tool for smart subject-related data gathering to be used by those techniques.

Bibliography

Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., and Spyropoulos, C. (2000). An Evaluation of Naive Bayesian Anti-Spam Filtering. In *Proc. of the workshop on Machine Learning in the New Information Age*.

Attardo, S. (1993). Violation of conversational maxims and cooperation: The case of jokes. *Journal of pragmatics*, 19(6):537--558.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Barbulet, G. (2013). Social Media- A Pragmatic Approach: Contexts & Implicatures. *Procedia - Social and Behavioral Sciences*, 83:422--426.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 562--570, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cavnar, W. B. and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161--175.

Cha, Y., Bi, B., Hsieh, C.-C., and Cho, J. (2013). Incorporating Popularity in Topic Models for Social Network Analysis. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 223--232.

Crammer, K., Dredze, M., and Pereira, F. (2012). Confidence-weighted linear classification for text categorization. *J. Mach. Learn. Res.*, 13(1).

Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pages 39--46, New York, NY, USA. ACM.

Cruse, A. D. (2006). *A Glossary of Semantics and Pragmatics*. Edinburgh University Press, Edinburgh.

Davis, A., Veloso, A., da Silva, A. S., Laender, A. H. F., and Meira Jr, W. (2012). Named Entity Disambiguation in Streaming Data. In *ACL'12*, pages 815--824.

Davis Jr, C. A., Pappa, G. L., Oliveira, D. R. R. d., and Arcanjo, F. d. L. (2011). Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS*, 15(6):735--751.

Drucker, H., Wu, D., and Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048--1054.

Eisterhold, J., Attardo, S., and Boxer, D. (2006). Reactions to irony in discourse: Evidence for the least disruption principle. *Journal of pragmatics*, 38(8):1239--1256.

Erkan, G. (2006). Language Model-based Document Clustering Using Random Walks. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 479--486, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gomide, J., Veloso, A., Meira Jr, W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd International Web Science Conference*, page 3. ACM.

Grice, P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and semantics. 3: Speech acts.*, pages 41--58. New York: Academic Press.

Guan, H., Zhou, J., and Guo, M. (2009). A Class-feature-centroid Classifier for Text Categorization. In *Proceedings of the 18th International Conference on World Wide Web*, pages 201--210, New York, NY, USA. ACM.

Guerra, P. H. C., Veloso, A., Meira Jr, W., and Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150--158. ACM.

Guo, Y., Shao, Z., and Hua, N. (2010). Automatic text categorization based on content analysis with cognitive situation models. *Information Sciences*, 180(5):613--630.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, 3:1157--1182.

Hanna, J. E. and Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, 28(1):105--115.

Hayes, P. J., Knecht, L. E., and Cellio, M. J. (1988). A News Story Categorization System. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 9--17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Herring, S. C. (2001). Computer-mediated discourse. *The handbook of discourse analysis*.

Hirschberg, J. (1985). *A Theory of Scalar Implicature*. PhD thesis, University of Pennsylvania.

Howard, P. N. and Parks, M. R. (2012). Social Media and Political Change: Capacity, Constraint, and Consequence. *Journal of Communication*, 62(2):359--362.

Husby, S. D. and Barbosa, D. (2012). Topic Classification of Blog Posts Using Distant Supervision. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 28--36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ifrim, G., Bakir, G. o. k., and Weikum, G. (2008). Fast Logistic Regression for Text Categorization with Variable-length N-grams. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 354--362, New York, NY, USA. ACM.

Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137--142.

Kurland, O. and Lee, L. (2010). PageRank Without Hyperlinks: Structural Reranking Using Links Induced by Language Models. *ACM Trans. Inf. Syst.*, 28(4):18:1--18:38.

Lam, W., Meng, H. M. L., Wong, K. L., and Yen, J. C. H. (2001). Using contextual analysis for news event detection. *International Journal on Intelligent Systems*.

Lao, N., Subramanya, A., Pereira, F., and Cohen, W. W. (2012). Reading the Web with Learned Syntactic-semantic Inference Rules. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational*

Natural Language Learning, pages 1017--1026, Stroudsburg, PA, USA. Association for Computational Linguistics.

Levinson, S. C. (1983). *Pragmatics (Cambridge textbooks in linguistics).* Cambridge Press.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT Press.

Li, C. H., Yang, J. C., and Park, S. C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. *Expert Systems With Applications*, 39(1):765--772.

Li, Z., Xiong, Z., Zhang, Y., Liu, C., and Li, K. (2011). Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, 32(3):441--448.

Machhour, H. and Kassou, I. (2013). Improving text categorization: A fully automated ontology based approach. In *Communications and Information Technology (ICCIT), 2013 Third International Conference on*, pages 67--72.

Mishne, G. (2005a). Blocking Blog Spam with Language Model Disagreement. In *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb).*

Mishne, G. (2005b). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access.*

Natarajan, P., Prasad, R., Subramanian, K., Saleem, S., Choi, F., and Schwartz, R. (2007). Finding structure in noisy text: topic classification and unsupervised clustering. *International journal on document analysis and recognition*, 10(3):187--198.

Pauls, A. and Klein, D. (2011). Faster and Smaller N-gram Language Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 258--267, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peng, F., Schuurmans, D., and Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317--345.

Phuvipadawat, S. and Murata, T. (2010). Breaking News Detection and Tracking in Twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 120--123.

Qiming, L., Chen, E., and Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems With Applications*, 38(10):12708--12716.

Raghavan, S., Mooney, R. J., and Ku, H. (2012). Learning to read between the lines using Bayesian Logic Programs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 349--358. Association for Computational Linguistics.

Saluja, A., Lane, I., and Zhang, Y. (2011). Context-aware Language Modeling for Conversational Speech Translation. *Proceedings of Machine Translation Summit XIII, Xiamen, China.*

Schwartz, R. M., Imai, T., Kubala, F., Nguyen, L., and Makhoul, J. (1997). A maximum likelihood model for topic classification of broadcast news. In Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors, *Eurospeech*. ISCA.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, 34(1):1--47.

Silva, I. S., Gomide, J., Veloso, A., Meira Jr, W., and Ferreira, R. (2011). Effective Sentiment Stream Analysis with Self-augmenting Training and Demand-driven Projection. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 475--484, New York, NY, USA. ACM.

Son, J.-W., Kim, A.-Y., and Park, S.-B. (2013). A location-based news article recommendation with explicit localized semantic analysis. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 293--302.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412--420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yus, F. (2003). Humor and the search for relevance. *Journal of pragmatics*, 35(9):1295--1331.

Yus, F. (2011). *Cyberpragmatics: Internet-mediated Communication in Context*. John Benjamins Publishing Company.

Attachment A

Chosen Keywords

For this dissertation, we were interested in three sbujects: american football, basketball and baseball. We gathered a training stream for each subject in which all messages contain at least one keyword from the following:

- **American Football:** nfl, new york, jets, nyj, new england, patriots, bufalo, bills, miami, dolphins, baltimore, ravens, cleveland, browns, pittsburgh, steelers, jacksonville, jaguars, indianapolis, colts, tenessee, titans, san diego, chargers, denver, broncos, oakland, raiders, kansas city, chiefs, dallas, cowboys, washington, redskins, philadelfia, eagles, giants, nyg, detroit, lions, chicago, bears, green bay, packers, minnesota, vikings, tampa bay, buccaneers, atlanta, falcons, carolina, panthers, new orleans, saints, arizona, cardinals, san francisco, 49ers, seattle, seahawks, st louis, rams, cheesehead, touchdown, fumble, sack, fumbles, interception, score

- **Baseball:** boston, red sox, redsox, tampa bay, rays, baltimore, orioles, new york, yankees, NY, toronto, blue jays, detroit, tigers, cleveland, indians, kansas city, royals, minnesota, twins, chicago, white sox, chi white sox, texas, rangers, oakland, seatle, mariners, los angeles, angels, LA angels, houston, astros, atlanta, braves, washington, nationals, mets, NY mets, philadelphia, phillies, miami, marlins, pittsburgh, pirates, st louis, cardinals, cincinnati, reds, milwaukee, brewers, cubs, chi cubs, dodgers, LA dodgers, arizona, diamondbacks, dbacks, colorado, rockies, san diego, padres, san francisco, giants, SF giants, SF

- **Basketball:** nba, boston, celtics, dallas, mavericks, brooklyn, nets, houston, rockets, new york, knicks, memphes, grizzlies, philadelphia, 76ers, new orleans, pelicans, toronto, raptors, san antonio, spurs, chicago, bulls, denver, nuggets, cleveland, cavaliers, minnesota, timberwolves, detroit, pistons, portland, trail, blazers, indiana, pacers, oklahoma city, thunder, milwaukee, bucks, utah, jazz, atlanta, hawks, golden state, warriors, charlotte, bobcats, los angeles, clippers, miami, heat, los angeles, lakers, orlando magic, phoenix, suns, whashington, wizards, sacramento, kings