

**UM ARCABOUÇO PARA PESQUISAS DE  
OPINIÃO EM REDES SOCIAIS**



RENATO MIRANDA FILHO

UM ARCABOUÇO PARA PESQUISAS DE  
OPINIÃO EM REDES SOCIAIS

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

ORIENTADOR: GISELE LOBO PAPPÀ

Belo Horizonte  
Fevereiro de 2014

© 2014, Renato Miranda Filho.  
Todos os direitos reservados.

Miranda Filho, Renato

M672a Um arcabouço para pesquisas de opinião em redes sociais / Renato Miranda Filho. — Belo Horizonte, 2014

xx, 98 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientadora: Gisele Lobo Pappa

1. Computação – Teses. 2. Redes sociais on-line – Teses. 3. Opinião pública - Pesquisa – Teses. 4. Estudo de usuários – Teses. I. Orientadora. II. Título.

CDU 519.6\*04(043)



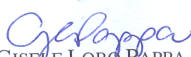
UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS EXATAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO


## FOLHA DE APROVAÇÃO

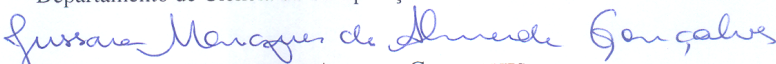
Um arcabouço para pesquisas de opinião em redes sociais

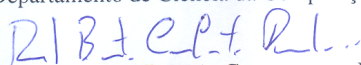
**RENATO MIRANDA FILHO**

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

  
PROFA. GISELE LOBO PAPPÁ - Orientadora  
Departamento de Ciência da Computação - UFMG

  
PROF. FABRÍCIO BENEVENUTO DE SOUZA  
Departamento de Ciência da Computação - UFMG

  
PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES  
Departamento de Ciência da Computação - UFMG

  
PROF. RICARDO BASTOS CAVALCANTE PRUDENCIO  
Centro de Informática - UFPE

Belo Horizonte, 24 de fevereiro de 2014.



# Agradecimentos

Agradeço a Deus por tudo!

O caminho percorrido foi difícil, mas sem algumas pessoas ao meu lado seria impossível.

Agradeço aos meus pais (Renato e Sonia) por sempre acreditarem em mim e me apoiarem em todos os momentos. À minha irmã (Ana), cunhado (Jaurés) e sobrinhos (Gabriel e Isaac), por sempre estarem ao meu lado e tornarem todos os momentos mais divertidos.

Agradeço também aos meus primeiros orientadores: Fabiano C. Botelho, Cristina D. Murta e Adriano C. M. Pereira, por terem me apresentado o mundo das pesquisas e pelos exemplos de vida que são.

Agradeço à minha orientadora Gisele L. Pappa, pela disponibilidade, paciência e confiança. Sua orientação foi fundamental para minha formação. Muito obrigado.

Gostaria de agradecer também aos diversos colegas que me ajudaram a construir este caminho: Alex Guimarães, Michelle Hanne, Pedro Calais, Sílvio Ribeiro, Walter dos Santos, Filipe de Lima, Erica Castilho, Arthur Iperoyg, Guilherme Borges e Gabriel Miranda.

Por fim, agradeço aos membros da banca por terem aceitado de forma tão solícita ao convite para participarem da defesa desta dissertação: Jussara Almeida, Fabrício Benevenuto e Ricardo Prudêncio.





*“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo”*

(Albert Einstein)



# Resumo

Extrapolando a pretensão inicial de servir como meio de comunicação interpessoal para o contato e a discussão dos mais diversificados assuntos, as redes sociais se tornaram sofisticadas fontes de informações que permitem o acompanhamento da opinião e do cotidiano de milhões de indivíduos. Neste contexto, este trabalho tem como principal objetivo utilizar dados públicos de tais redes para realizar pesquisas (análises) de opinião do tipo “IBOPE”. Para tanto, levou-se em consideração fatores inerentes do ambiente virtual como a análise de sentimentos das mensagens, a detecção de *spammers* e de conteúdos jornalísticos aliados a um ponto importantíssimo em análises desta natureza: uma formação de amostra representativa da população alvo da pesquisa. Para a formação desta amostra foram levadas em consideração as seguintes características dos indivíduos: (i) identificação de usuários únicos; (ii) sexo; (iii) idade; (iv) classe social; e (v) localização geográfica. A avaliação experimental deste trabalho foi realizada com dados provenientes de diversos contextos para os quais existem resultados validados por agências, são eles: intenção de votos nas eleições municipais de 2012 em 6 capitais e 14 votações no *Big Brother* Brasil 13. Os resultados alcançados mostram que análises mais elaboradas são capazes de melhorar os números atingidos por metodologias empregadas em outros trabalhos, fundamentadas majoritariamente em técnicas de contagem. Comparando os resultados obtidos com os métodos tradicionais de pesquisa (enquetes realizadas por grandes sítios e pesquisas realizadas por institutos renomados) também alcançamos resultados competitivos, mostrando que a técnica desenvolvida pode ser considerada como uma boa alternativa.

**Palavras-chave:** Pesquisas de Opinião, Redes Sociais, Caracterização de Usuários, Twitter.



# Abstract

Extrapolating the initial intention of serving as means of interpersonal communication, social networks have evolved to be sophisticated sources of information that allow monitoring the opinion and the daily lives of millions of individuals. In this context, the main goal of this work is to use public data available from these networks to conduct opinion analysis in the same way the “NIELSEN” does. In order to do that, we take into account inherent characteristics of virtual environments, including sentiment analysis of messages and the detection of spammers and journalistic content. This first analysis was combined with the generation of a representative sample of the target population of study. For the formation of this sample, the following characteristics of individuals were taken into account: (i) identification of unique users; (ii) gender; (iii) age; (iv) social class; and (v) geographical location. The experimental evaluation was performed with data from various contexts which results were validated by real agencies: intention to vote in municipal elections of 2012, and poll of Big Brother Brasil. The results show that more elaborate analyzes are able to improve the numbers achieved by methods used in other studies, mostly based on counting techniques. Comparing the results achieved with traditional research methods (polls conducted by major sites and research conducted by renowned institutos) also achieved competitive results, showing that the technique developed can be considered as a good alternative.

**Keywords:** Research Opinion, Social Network, Characterization of Users, Twitter.



# Lista de Figuras

1.1	Previsão para as eleições americanas [Institute, 2012] . . . . .	2
1.2	Resultado das eleições americanas [Times, 2012] . . . . .	3
1.3	Indivíduo . . . . .	4
3.1	Esquema para pesquisa de opinião em redes sociais . . . . .	21
3.2	Idade média das contas dos usuários . . . . .	27
3.3	Média do mínimo de <i>tweets</i> postados por semana . . . . .	27
3.4	Fração média de <i>tweets</i> respondidos . . . . .	28
3.5	Fração média de <i>tweets</i> com URLs . . . . .	28
3.6	Média de menções por <i>tweet</i> . . . . .	28
3.7	Algoritmos para análise comportamental . . . . .	30
3.8	Eleições - concordância entre L, RM e RA . . . . .	35
3.9	Eleições - concordância entre L e RM . . . . .	35
3.10	Eleições - concordância entre L e RA . . . . .	35
3.11	Eleições - concordância entre RM e RA . . . . .	35
3.12	BBB - concordância entre L, RM e RA . . . . .	35
3.13	BBB - concordância entre L e RM . . . . .	35
3.14	BBB - concordância entre L e RA . . . . .	36
3.15	BBB - concordância entre RM e RA . . . . .	36
4.1	Média de menção a ídolos até 16 anos por usuário . . . . .	49
4.2	Média de menção a filmes até 16 anos por usuário . . . . .	49
4.3	Média de menção a palavras de diversão noturna por usuário . . . . .	50
4.4	Média de menção a palavras frequentes mais de 45 anos por usuário . . . . .	50
4.5	Média do tamanho das palavras por usuário . . . . .	51
4.6	Média do tamanho dos <i>tweets</i> por usuário . . . . .	51
4.7	Média do número de hashtags por usuário . . . . .	51
4.8	Termos discriminativos para três classes sociais . . . . .	60

4.9	Distribuição do Produto Interno Bruto brasileiro (PIB) per capita em diferentes regiões do país . . . . .	61
4.10	Correlação de Pearson do vocabulário das classes utilizando 100 termos . . . . .	62
A.1	Grafo de amizades na base com 8,477 usuários do Twitter . . . . .	95
A.2	Matrizes de confusão alcançadas com a aplicação da metodologia proposta. A primeira matriz mostra as medidas de recall e a segunda a medida da precisão . . . . .	96
A.3	Matrizes de confusão dos resultados alcançados através da aplicação da metodologia proposta e correção pelo fator Tf-Idf. A primeira matriz mostra as medidas de recall e a segunda a medida da precisão . . . . .	98



# Lista de Tabelas

3.1	Algoritmos de classificação - parâmetros utilizados . . . . .	23
3.2	Resultados para classificação de usuários <i>spammers</i> . . . . .	24
3.3	Comparação atributos selecionados . . . . .	26
3.4	Dicionários Léxico - número de palavras . . . . .	32
3.5	Rotulação manual - Acurácia (%) . . . . .	33
3.6	Lista de <i>Emoticons</i> . . . . .	33
3.7	Rotulação automática dos tweets selecionados com <i>emoticons</i> - Acurácia (%)	33
3.8	Análise de sentimentos - avaliação do método (Métricas de acurácia e cobertura em porcentagem (%)) . . . . .	36
3.9	Tamanho da amostra (95.5% de confiança) . . . . .	38
4.1	Sexo <i>Womens Health</i> - F1 utilizando todos os termos . . . . .	42
4.2	Sexo <i>Mens Health</i> - F1 utilizando todos os termos . . . . .	42
4.3	Sexo Revistas ( <i>Womens</i> e <i>Mens Health</i> ) - F1 utilizando todos os termos . .	42
4.4	Sexo BBB - F1 utilizando todos os termos . . . . .	43
4.5	Sexo Eleições - F1 utilizando todos os termos . . . . .	43
4.6	Sexo Base Completa - F1 utilizando todos os termos . . . . .	43
4.7	Sexo Base Completa - seleção de atributos . . . . .	44
4.8	Acerto do dicionário . . . . .	44
4.9	Acerto das variações do método (%) . . . . .	45
4.10	Principais termos para divisões de idade em 3 e 5 classes, conforme medida de Ganho de Informação . . . . .	52
4.11	Resultados obtidos utilizando o texto completo, atributos textuais selecionados, não textuais selecionados e todos selecionados (textuais com não textuais) - métrica de acurácia mostrada em porcentagem (%) . . . . .	53
4.12	FGV - Definição de classe social de acordo com o rendimento mensal familiar	56
4.13	Número de usuários do Twitter encontrado em cada classe . . . . .	57
4.14	Número médio de interações <i>Foursquare</i> considerando a classe atribuída . .	58

4.15	Média da distância máxima (km) entre quaisquer dois lugares visitados pelos usuários, considerando a classe atribuída . . . . .	58
4.16	Distribuição dos usuários em classes sociais conforme sua região . . . . .	62
4.17	Principais termos para lazer e consumo conforme medida de Ganho de Informação . . . . .	64
4.18	Resultados obtidos para 2, 3 e 4 classes sociais, utilizando os atributos textuais na base de dados balanceada e original . . . . .	65
4.19	Resultados para a classificação de classe social usando NBM em duas regiões	65
4.20	Resultados obtidos para 2, 3 e 4 classes sociais utilizando diferentes atributos	67
5.1	Resultados BBB - p@1 (Sen: análise de sentimento; SJ: remoção de usuários <i>spammers</i> e jornalísticos; Sex: sexo; ID: idade; CS: classe social) . . . . .	72
5.2	Perfil do eleitorado em 2012 . . . . .	75
5.3	Perfil dos usuários do Twitter - Eleições 2012 . . . . .	76
5.4	Resultados Eleições - p@1 (Sen: análise de sentimento; SJ: remoção de usuários <i>spammers</i> e jornalísticos; Sex: sexo; ID: idade; CS: classe social) - Quadrados de cor verde representam acertos, pretos a não formação de amostra e, conseqüentemente, brancos os erros. . . . .	79
5.5	Resultados Eleições - Segundo turno (Sen: análise de sentimento; SJ: remoção de usuários <i>spammers</i> e jornalísticos; Sex: sexo; ID: idade; CS: classe social) - Quadrados de cor verde representam acertos, pretos a não formação de amostra e, conseqüentemente, brancos os erros. . . . .	80
A.1	Localização: precisão dos métodos (%) . . . . .	95

# Sumário

Agradecimentos	vii
Resumo	xi
Abstract	xiii
Lista de Figuras	xv
Lista de Tabelas	xvii
<b>1 Introdução</b>	<b>1</b>
1.1 Objetivos . . . . .	4
1.2 Contribuições . . . . .	5
1.3 Organização . . . . .	5
<b>2 Revisão Bibliográfica</b>	<b>7</b>
2.1 Pesquisas de opinião tradicionais . . . . .	7
2.2 Redes sociais no monitoramento de eventos . . . . .	9
2.3 Caracterização pessoal de usuários . . . . .	11
2.4 Caracterização comportamental de usuários . . . . .	13
2.5 Análise de sentimentos . . . . .	15
2.6 Relevância estatística de dados da API do Twitter . . . . .	17
<b>3 PODEReS</b>	<b>19</b>
3.1 Preparação dos dados . . . . .	20
3.2 Caracterização comportamental . . . . .	22
3.2.1 Identificação de usuários <i>spammers</i> . . . . .	22
3.2.2 Identificação de usuários jornalísticos . . . . .	27
3.2.3 Método resultante . . . . .	29
3.3 Análise de sentimentos . . . . .	30

3.3.1	Dicionário Léxico . . . . .	31
3.3.2	Algoritmo supervisionado - Rotulação manual . . . . .	32
3.3.3	Algoritmo supervisionado - Rotulação automática . . . . .	32
3.3.4	Concordância entre os algoritmos de análise de sentimentos . . . . .	34
3.3.5	Avaliação do método . . . . .	36
3.4	Tamanho da amostra . . . . .	37
<b>4</b>	<b>Caracterização pessoal</b>	<b>39</b>
4.1	Sexo . . . . .	40
4.1.1	Dicionário de nomes . . . . .	41
4.1.2	Classificadores . . . . .	41
4.1.3	Avaliação do método . . . . .	43
4.2	Idade . . . . .	45
4.2.1	Base de dados . . . . .	45
4.2.2	Características textuais e não textuais . . . . .	46
4.2.3	Classificação . . . . .	51
4.3	Classe social . . . . .	53
4.3.1	Base de dados . . . . .	54
4.3.2	Características textuais . . . . .	59
4.3.3	Classificação . . . . .	64
<b>5</b>	<b>Avaliação do arcabouço</b>	<b>69</b>
5.1	<i>Big Brother</i> Brasil (BBB) . . . . .	70
5.2	Eleições Municipais 2012 . . . . .	74
5.3	Considerações finais . . . . .	77
<b>6</b>	<b>Conclusões e Trabalhos Futuros</b>	<b>81</b>
6.1	Trabalhos Futuros . . . . .	82
	<b>Referências Bibliográficas</b>	<b>85</b>
	<b>Apêndice A Localização</b>	<b>91</b>
A.1	Metodologia . . . . .	92
A.2	Avaliação do método . . . . .	94

# Capítulo 1

## Introdução

Surgida no início da década de 90, a Web se desenvolveu e hoje podemos dizer que suas ferramentas permeiam praticamente todas as atividades humanas. Diversos mecanismos foram desenvolvidos a fim de proporcionar a seus usuários um maior relacionamento interpessoal, são alguns exemplos: e-mails, sítios para compartilhamento de arquivos e todas as demais ferramentas que surgiram com a evolução da Web 2.0, em que o usuário passou a ser um provedor ativo de informações.

O mais novo fenômeno de popularidade que vem captando cada vez mais usuários na Web são as redes sociais. Tais redes, por meio de ligações de amizades, seguidores ou conhecidos, permitem a seus participantes uma comunicação efetiva e em tempo real sobre os mais diversos assuntos. Dentre as grandes redes sociais da atualidade podemos destacar o Twitter, com mais de 500 milhões de usuários [G1, 2012c], Facebook, com aproximadamente 1 bilhão de usuários [G1, 2012a], Google *plus*, com estimativa de 400 milhões de usuários [Terra, 2012], e o LinkedIn, com 175 milhões de usuários [G1, 2012b].

Um grande diferencial encontrado em tais redes sociais é a finalidade para a qual se destinam. Enquanto uma parte se dedica ao compartilhamento de informações pessoais como, por exemplo, Facebook e Google *plus*, outras se propõem a criar mecanismos relativos ao relacionamento profissional como o LinkedIn ou mesmo ao compartilhamento sobre o que está acontecendo, como pode ser exemplificado pelo Twitter.

Dado o volume de informações disponíveis, um tema que vem despertando o interesse da comunidade científica nos últimos tempos é a identificação de opiniões, gostos e sentimentos que são expressos nessas redes. Assim, esses dados também estão sendo cada vez mais explorados para prever o resultados de eventos no mundo real. Alguns exemplos notórios desse fenômeno são os trabalhos publicados que tentaram prever o

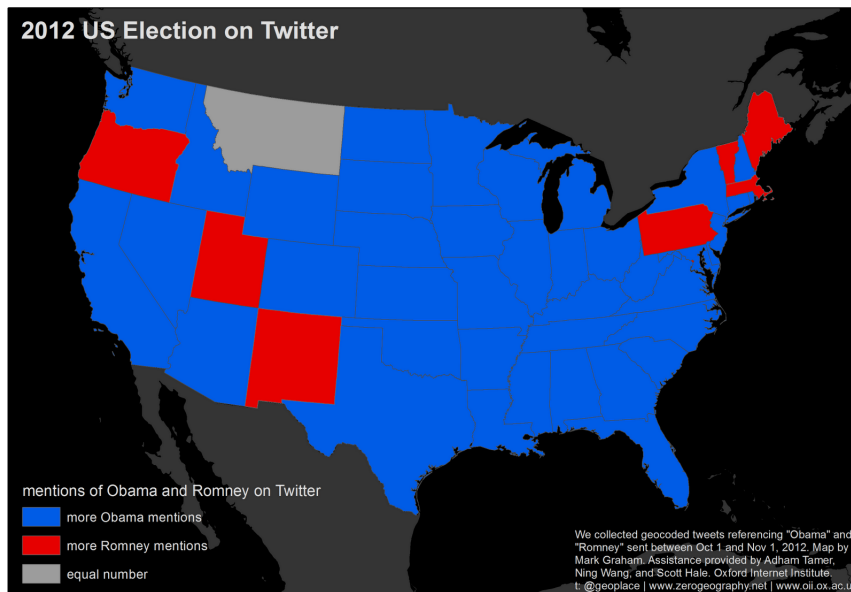


Figura 1.1: Previsão para as eleições americanas [Institute, 2012]

resultado de eleições [Tumasjan et al., 2010; Gayo-Avello, 2012; Jungherr et al., 2012], bilheteria no cinema [Asur & Huberman, 2010] e epidemias [Gomide et al., 2011].

Nessa direção, o grande objetivo deste trabalho é criar um arcabouço teórico e ferramental que permita a realização de pesquisas de opinião por meio de dados coletados em redes sociais e validar sua eficácia diante dos métodos tradicionalmente utilizados.

Pesquisas de opinião são normalmente realizadas por meio de questionários, eletrônicos ou não, que devem ser respondidos por indivíduos com características previamente selecionadas, tais como: idade, localidade, sexo, entre outros. Normalmente, tais pesquisas trabalham com amostras da população, cujos indivíduos escolhidos devem ser os mais representativos possíveis do conjunto total. Quando uma pesquisa é realizada com todos os indivíduos de uma população podemos denominá-la censo.

Algumas limitações embutidas em tais processos estão nos altos custos financeiros que devem ser empregados, na dificuldade de entrevistar grandes volumes de pessoas e em como selecionar indivíduos representativos da população.

Um exemplo desse tipo de estudo, utilizando redes sociais, é mostrado na Figura 1.1. Neste trabalho, pesquisadores ligados ao *Oxford Internet Institute* realizaram a contagem das citações sobre os candidatos à presidência dos Estados Unidos em aproximadamente 30 milhões de *tweets* geo-localizados. Podemos ver que, por grande margem de estados, o candidato democrata Barack Obama seria reeleito. A reeleição realmente ocorreu, mas como podemos ver pela Figura 1.2, que mostra o resultado das eleições, a distribuição de votos pelos estados foi muito diferente da prevista.

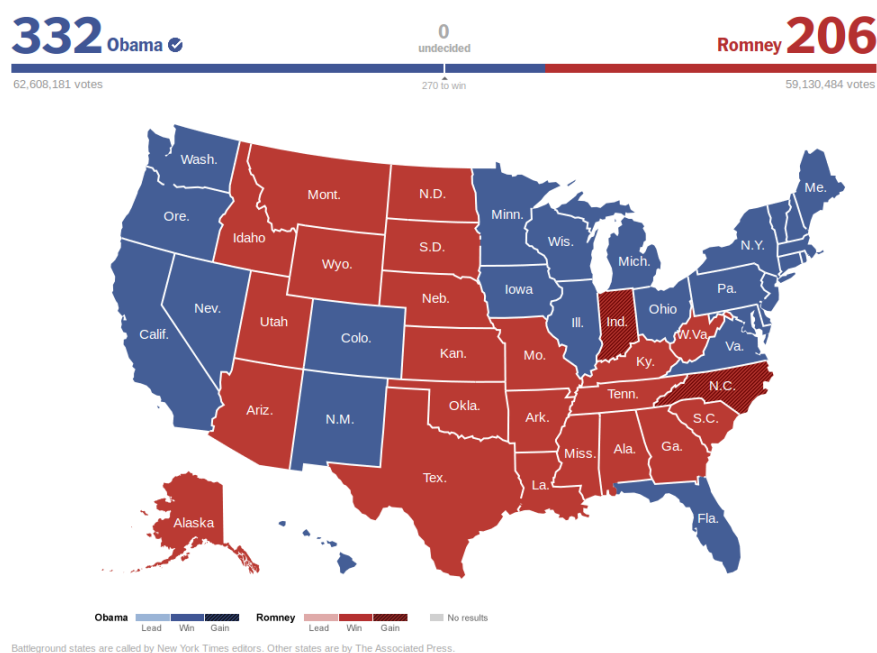


Figura 1.2: Resultado das eleições americanas [Times, 2012]

Outro trabalho nesse sentido foi o de Tumasjan et al. [2010]. Imerso nas eleições parlamentares da Alemanha, este estudo realizou a contagem de menções a candidatos ou partidos políticos e identificou que o volume coletado é um bom preditor para os resultados de eleições. Porém, em Jungherr et al. [2012], onde o mesmo procedimento foi adotado, os autores não conseguiram prever o resultado do pleito.

Assim, identificamos que tais estudos têm trabalhado sob consideráveis limitações como, por exemplo:

1. Considerar que a simples menção será um voto a um candidato ou um gosto particular;
2. Ignorar a localização geográfica de parte dos usuários, ou, como no caso do estudo de *Oxford*, desconsiderar mensagens sem localização;
3. Desconsiderar critérios de formação de amostra para pesquisas com, por exemplo, a distribuição de idade e sexo da população;
4. Considerar que todas as mensagens postadas são confiáveis, ou seja, não retiram da amostra possíveis usuários *spammers* ou propagandistas.
5. Considerar que todas as mensagens postadas são opiniões de indivíduos, desconsiderando, por exemplo, conteúdos jornalísticos.

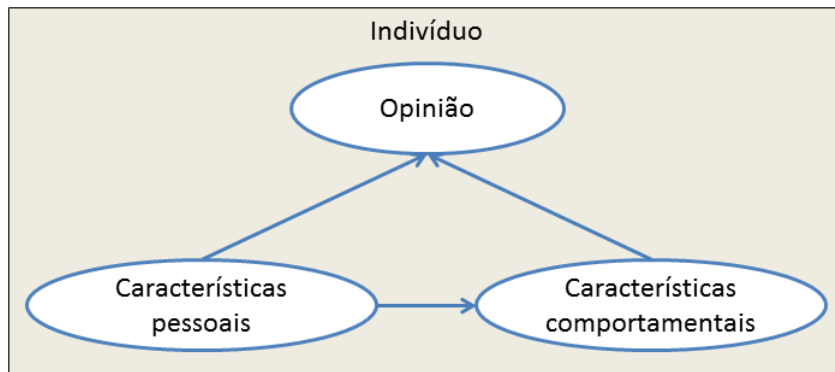


Figura 1.3: Indivíduo

Dada as limitações das abordagens atuais, este trabalho propõe uma estratégia que leva em consideração o tratamento de todos os pontos descritos anteriormente. Para tanto, propomos inicialmente caracterizar os indivíduos. Um indivíduo é representado por uma conta/perfil na rede social e, como mostrado na Figura 1.3, possui características pessoais capazes de interferir em seu comportamento e com um conjunto de características pessoais e comportamentais que podem nos auxiliar a inferir sua opinião.

## 1.1 Objetivos

O grande objetivo deste trabalho é criar um arcabouço teórico e ferramental que permita a realização de pesquisas de opinião por meio de dados coletados em redes sociais e validar sua eficácia diante dos métodos tradicionalmente utilizados.

Como partes constituintes da face teórica do trabalho podemos citar o estudo de características sociológicas capazes de diferenciar, por exemplo, diferentes faixas etárias, comportamentos de classes sociais e condutas típicas dos diferentes gêneros. Já a face ferramental pode ser representada pelos métodos, códigos e análises desenvolvidas.

Os principais objetivos específicos são:

- Levantamento das principais metodologias empregadas pelos institutos de pesquisa tradicionais, incluindo métodos estatísticos para formação de amostra;
- Identificar as principais adaptações que devem ser consideradas em pesquisas realizadas em redes sociais frente às pesquisas tradicionais;
- Propor um *framework* com as etapas identificadas para realização de pesquisas em redes sociais e verificar os benefícios alcançados com cada parte constituinte do processo.



- Estudar, adaptar e implementar os principais algoritmos existentes para as etapas de análise de sentimentos, identificação de *spammers* e conteúdo jornalístico;
- Desenvolver métodos para, sem violarmos a privacidade dos indivíduos, identificar a faixa etária, sexo, localização e classe social dos usuários;
- Comparar e discutir os resultados obtidos pelo método proposto, em diferentes cenários, com os resultados ocorridos no mundo real.

## 1.2 Contribuições

Os objetivos propostos inicialmente neste trabalho foram inteiramente cumpridos. Além disso, diferentes etapas desenvolvidas para o arcabouço apresentado como, por exemplo, a caracterização do sexo, classe social, idade, análise de sentimentos e identificação de usuários *spammers* e jornalísticos estão sendo integrados as ferramentas do Observatório da Web<sup>1</sup>.

## 1.3 Organização

Este trabalho encontra-se dividido da seguinte forma: O Capítulo 2 apresenta alguns trabalhos que serviram de base para o estudo. Os Capítulos 3 e 4 propõem o método para a realização de pesquisas de opinião em redes sociais, sendo o capítulo 4 destinado exclusivamente a caracterização pessoal do usuário. No Capítulo 5 encontram-se os resultados alcançados. Finalmente, as conclusões e propostas de trabalhos futuros são apresentadas no Capítulo 6.

---

<sup>1</sup><http://observatorio.inweb.org.br/>



# Capítulo 2

## Revisão Bibliográfica

Neste capítulo, faremos uma breve revisão sobre como são realizadas as pesquisas de opinião tradicionais. Em seguida, focaremos nossa atenção nas redes sociais. Como a tarefa de realizar pesquisas de opinião nestas redes pode ser subdividida em várias linhas de pesquisa, optamos por organizar esta parte da revisão bibliográfica em quatro diferentes categorias, são elas: (i) Redes sociais no monitoramento de eventos; (ii) Caracterização pessoal de usuários; (iii) Caracterização comportamental de usuários; e (iv) Análise de sentimento/opinião.

### 2.1 Pesquisas de opinião tradicionais

Atualmente as pesquisas de opiniões tradicionais coletam informações sobre os mais diversificados assuntos, dentre eles: intenção de votos, comportamento de consumo, audiência da TV, preferência de marcas, entre outros. Entender o processo que guia a realização destas pesquisas é de fundamental importância para identificarmos os critérios que deverão ser levados em consideração em pesquisas realizadas *online*.

No Brasil, podemos destacar como principais institutos de pesquisa o IBOPE e o Data Folha. Já nos EUA podemos destacar o Nielsen, na aferição de audiência de TV, e a *Public Policy Polling* (PPP), em realização de pesquisas eleitorais.

Tais institutos possuem como grande diferencial a metodologia de pesquisa que é empregada em seus levantamentos como, por exemplo, meio de realização e formação da amostra. Tais metodologias também podem diferir dependendo do objetivo da pesquisa.

Abaixo são mostrados maiores detalhes para o caso de pesquisas eleitorais:

(a) IBOPE: Utiliza dados oficiais provenientes, por exemplo, do Instituto Brasileiro

de Geografia e Estatística (IBGE) e do Tribunal Superior Eleitoral (TSE) para definir cotas proporcionais ao sexo, idade, grau de instrução e setor de dependência econômica na amostra do universo de pessoas entrevistadas [IBOPE, 2012b,a].

Além disso, as pesquisas são realizadas em domicílio, após uma seleção probabilística dos setores censitários do IBGE onde serão realizadas as entrevistas, pelo método PPT (Probabilidade Proporcional ao Tamanho) [IBOPE, 2012b,a].

- (b) Data Folha: O Datafolha [Folha, 2012] utiliza cotas proporcionais de sexo e idade de acordo com dados obtidos junto ao IBGE e TSE. Além disso, também é observada a região geográfica dos entrevistados. Variáveis como escolaridade ou renda familiar mensal não são utilizadas como critério para formação de amostras, pois o Data Folha considera que não existem dados atualizados disponíveis para os municípios brasileiros.

As entrevistas são realizadas pessoalmente nas ruas das cidades. Métodos como, por exemplo, entrevistas via telefone ou e-mail são descartadas, pois, segundo o instituto, isso formaria uma amostra que não representa toda a população.

Em uma pesquisa eleitoral nacional o tamanho das amostras do Datafolha tem entre 2,000 e 2,500 entrevistas, mas não é estabelecido um tamanho mínimo, o mais importante é sua representatividade.

- (c) *Public Policy Polling*: diferentemente dos institutos brasileiros o PPP realiza suas entrevistas automaticamente via telefone. São apontados como vantagem deste método a eliminação de possíveis interferências humanas e a redução de custos [PPP, 2012].

A amostra dos entrevistados é formada com base em um cadastro de eleitores em que cotas e pesos para sexo, idade, raça, localização, entre outros, são estabelecidos.

Desta forma, podemos concluir que um ponto fundamental das pesquisas é a escolha adequada de uma amostra que represente a população. Em nosso trabalho esta etapa será realizada por meio da caracterização dos usuários com informações como faixa etária, localização, classe social e sexo. Além disso, dependendo do objetivo da pesquisa, dados oficiais do IBGE e TSE podem ser utilizados para definir a proporção de cada uma destas características nas amostras construídas.

## 2.2 Redes sociais no monitoramento de eventos

Dado o grande volume de informação disponível, as redes sociais estão sendo cada vez mais exploradas para monitorar e prever eventos no mundo real. Assim, conhecer trabalhos neste âmbito é de suma importância para identificarmos, principalmente, como são coletadas as informações sobre tais eventos e quais critérios são utilizados para avaliar os dados obtidos.

O estudo realizado por Asur & Huberman [2010] construiu um modelo de regressão linear e utilizou mensagens no Twitter para prever a receita de bilheteria de filmes. Com isso, os autores mostraram que existe relação entre a atenção dada a um tópico, por parte dos usuários, com público que assistirá os filmes.

Alguns trabalhos também estudaram a relação de epidemiologias no mundo real com o que é dito nas redes sociais. Gomide et al. [2011] estudou como a Dengue é discutida no Twitter e como as mensagens coletadas poderiam ser usadas para vigilância da doença. Em Culotta [2010] os autores também analisaram mensagens postadas no Twitter, mas para verificar a existência de correlação entre mensagens que falavam sobre a Gripe com estatísticas divulgadas por centros de controle de doenças.

Neste trabalho iremos focar nossas atenções em mensagens relacionadas às disputas eleitorais. Neste contexto, imerso no cenário da eleição parlamentar da Alemanha, ocorrida em 2009, o trabalho realizado por Tumasjan et al. [2010] estuda se mensagens postadas no Twitter possuem deliberações políticas e como tais conteúdos podem refletir o resultado de eleições. A metodologia empregada consistiu na coleta de *tweets* que mencionavam um dos seis candidatos da eleição ou seus respectivos partidos políticos.

Após a coleta, as mensagens foram traduzidas para o inglês e submetidas ao *software* LIWC2007 (*Linguistic Inquiry and Word Count*), onde análises de sentimentos, baseados na frequência de palavras empiricamente categorizadas em termos psicológicos e estruturais, foram realizadas.

Os autores observaram que o Twitter é realmente utilizado como uma plataforma para deliberações políticas demonstrando, inclusive, características de discussões e trocas de ideias por meio do uso de *retweets*. No entanto, também foi observado que existe uma considerável concentração nos emissores de mensagens, ou seja, um pequeno número de pessoas, 4%, foram responsáveis por grande parte dos *tweets* postados neste cenário, mais de 40%.

Pela análise de sentimentos das mensagens foi observado que o comportamento dos eleitores se aproxima às ideias difundidas pela campanha eleitoral como, por exemplo, retratando coalisões ou ataques. Como preditor do resultado das eleições, a simples contagem do volume de menções aos partidos ou candidatos foi capaz de acompanhar

pesquisas eleitorais tradicionais e acertar com exatidão o resultado final da eleição.

Foram limitações deste estudo: (i) a falta de cuidado com a escolha da amostra, pois dados demográficos capazes de retratar o eleitorado alemão, tais como, idade, sexo e região não foram utilizados, (ii) a coleta das mensagens se restringiu apenas ao nome dos partidos e candidatos, possibilitando perda de informações, (iii) uso de *software* para análise de sentimentos que não foi desenvolvido para trabalhar com mensagens curtas, além da possível perda de significância dos termos durante o processo de tradução alemão-inglês e (iv) a análise de apenas uma eleição pode ser considerada insuficiente para caracterizar o comportamento de todos os possíveis cenários.

Já em Jungherr et al. [2012], os autores realizaram um experimento sobre as eleições federais da Alemanha de 2009 com a técnica proposta por Tumasjan et al. [2010], que consiste em contabilizar a frequência de citações a partidos políticos, e não validaram a afirmativa de que é possível prever o resultado final do pleito. Assim, os autores apontaram algumas limitações de tal técnica como, por exemplo, a falta de regras bem fundamentadas para a coleta, da escolha dos partidos e do período de tempo correto que deviria se avaliar.

O trabalho realizado por Gayo-Avello et al. [2011] também coloca em xeque o poder preditivo de eleições pelas redes sociais. Neste trabalho os autores utilizaram *tweets* coletados sobre as eleições de 2010 para o congresso americano, aplicaram algumas técnicas normalmente encontradas na literatura e não observaram nenhuma relação entre os resultados obtidos e os resultados eleitorais.

Em Gayo-Avello [2012] é realizada uma pesquisa bibliográfica sobre trabalhos que têm como objetivo a previsão eleitoral com base em dados coletados do Twitter. O autor foca sua análise em citar falhas e limitações das abordagens já empregadas.

São exemplos de pontos considerados como falhos nos trabalhos analisados: (i) falta de previsão de resultados futuros e não somente em eleições já ocorridas; (ii) limitação nas abordagens de contagem de voto, em que muitas vezes é utilizado somente o volume bruto de *tweets* ou de usuários únicos que citaram os candidatos; (iii) análise de sentimentos realizado de forma ingênua; (iv) falta de emprego de técnicas para detectar boatos e informações falsas; (v) negligência na verificação e consequente utilização de dados demográficos como, por exemplo, idade e sexo dos eleitores, (vi) considerar que a opinião emitida por parte dos usuários das redes sociais, em que normalmente são apenas pessoas politicamente ativas que produzem o conteúdo, representa a opinião de toda a população e (vii) definição imprecisa do que é considerado como “voto”. Assim, o autor conclui que a tarefa proposta não deve ser encarada de forma simplista e que nem sempre é possível prever os resultados de eleições.

Podemos concluir que mensagens postadas no Twitter podem nos ajudar a identi-

ficar diversos acontecimentos da vida real, mas que as técnicas utilizadas para detectar tal evento e mais ainda os métodos para inferir a opinião dos indivíduos não podem ser tratados de forma simplista, onde, por exemplo, somente a menção a um determinado termo é levada em consideração.

## 2.3 Caracterização pessoal de usuários

Como visto anteriormente, inferir características pessoais dos usuários, tais como faixa etária, sexo e localidade, é fundamental para a formação de uma amostra que seja significativa do conjunto total da população e, portanto, deve ser alvo de um estudo cuidadoso por parte desta pesquisa.

Peersman et al. [2011] realiza um estudo exploratório, com uma base coletada da rede social belga Netlog, em que se aplica uma abordagem de caracterização de texto para previsão de idade e sexo. No quesito idade, o principal objetivo dos autores é classificar adultos e adolescentes e, com isso, auxiliar na tarefa de identificar possíveis usuários pedófilos. Já o quesito sexo é de interesse dos autores, porque existe uma predominância em pessoas do sexo masculino entre os indivíduos pedófilos. Assim, os autores utilizaram um classificador SVM e a técnica para escolha de atributos  $\chi^2$  (*Chi Square*) e observaram que a escolha de palavras, tais como “bro”(brother) e “grts” (*greetings*) parece ser mais importante para a previsão da idade do que a forma como tais termos são combinados. O SVM conseguiu obter resultados promissores na identificação de adolescentes e adultos, principalmente em faixa etária maiores. A informação de gênero também se mostrou como útil na construção de um classificador de idade mais preciso.

Para identificar a idade e sexo de blogueiros, Goswami et al. [2009] realizou um estudo sobre diferenças de estilo entre as mensagens postadas pelos usuários. Dentre as características propostas para serem usadas são o uso de gírias e a variação no comprimento médio das sentenças, além de outras características já normalmente utilizadas como “palavras de conteúdo”, ou seja, termos que são encontrados uma enorme variação entre sexo e grupos etários. Assim, após montar um conjunto de treinamento, houve uma classificação pelo método *Naive Bayes*, em que se obteve uma precisão de aproximadamente 80% para a identificação de gênero e de aproximadamente 90% para a identificação de idade. Também se observou que os grupos mais jovens utilizam mais gírias além de postarem mensagens mais curtas.

Um estudo sobre a relação existente entre o uso da linguagem e a previsão da idade dos usuários do Twitter, utilizando contas holandesas, é realizado em Nguyen et al.

[2013]. O principal fator abordado são as mudanças que ocorrem com a diferença etária. Neste trabalho, os autores trabalham com a classificação de idade em três níveis: (i) faixas etárias; (ii) fases da vida; e (iii) idade exata. Algumas das observações realizadas foram: a relevância de identificar o sexo dos indivíduos em trabalhos para previsão de idade, o comportamento de pessoas mais jovens, com alongamento de palavras (repetição de caracteres) e uso constante da primeira pessoa, e o de pessoas mais velhas, com tweets mais longos e maior uso de preposições.

Partindo de uma análise de mensagens postadas em milhares de blogs o trabalho realizado por Schler et al. [2006] estuda a existência de diferenças significativas no estilo de escrita entre os sexos masculino/feminino e entre os diferentes grupos de idade. Os autores identificaram diferenças tanto no conteúdo das mensagens como, por exemplo, homens escrevem mais sobre política, tecnologia e dinheiro e mulheres mais sobre vida pessoal, quanto no estilo de escrita como, por exemplo, mulheres usam mais pronomes e palavras de negação e concordância enquanto homens usam mais artigos e preposições.

Outras características, como o uso de *hiperlinks* e número de palavras, também se mostraram como fator de diferenciação entre sexos, onde o primeiro ocorre mais entre os homens e o segundo mais entre as mulheres. Estas diferenças também se refletem nas faixas etárias, em que de forma geral observa-se que as mensagens tendem a adquirir um comportamento mais masculino entre os grupos mais velhos.

Em relação a localização do usuário, no intuito de aumentar o número de *tweets* com informação de localidade o trabalho proposto por Davis Jr. et al. [2011] consiste na expansão da rede de seguidor-seguido dos usuários coletados do Twitter, por meio de uma busca em profundidade, para inferir a localidade de usuários sem esse tipo de informação. Assim, utilizando um esquema de votação no caminamento da rede eles conseguiram aumentar em 45% a quantidade de mensagens localizáveis. Outros métodos de classificação para dados dispostos em grafos já foram apresentados na literatura e uma revisão ampla sobre o assunto pode ser obtida em Macskassy & Provost [2007].

Uma outra fonte de informação que pode ser utilizada para se fazer inferência sobre a localização do usuário são as mensagens que ele publica. Assim, utilizando algoritmos de classificação, em Mahmud et al. [2012] os autores são capazes de inferir a localização para diferentes níveis de granularidade: cidade, estado e zona temporal.

Também com base no texto, Cheng et al. [2010] fazem uso do fato de que os *tweets* postados pelos usuários podem conter alguma informação sobre sua posição geográfica. Essa informação pode vir de nomes específicos ou expressões que tenham maior probabilidade de estarem associadas a um determinado local. Por exemplo, a expressão *Howdy*, que em português significa “olá”, é tipicamente utilizada no estado do Texas.



Ao se utilizar esse tipo de informação no processo de inferência, aparecem uma série de complicações. Algumas delas são apontadas pelos autores. As mensagens postadas apresentam muito ruído. Elas abordam os mais diversos assuntos: comida, esportes, diálogos pessoais, etc. Uma pequena fração possui, de fato, conteúdo espacial. Um outro problema é a presença recorrente de gírias e expressões informais. Além disso, existe o problema da mobilidade do usuário. Ele pode morar em local, mas postar mensagens sobre outro. Por exemplo, habitantes de Nova York podem postar mensagens sobre o terremoto no Haiti. O usuário pode ainda ter mais de uma localização válida. Ele pode, por exemplo, estar viajando ou se mudar de um local para o outro. Todos esses aspectos dificultam o processo de inferência e alguns deles são abordados pelos autores do trabalho.

Não encontramos estudos anteriores que caracterizassem a classe social dos usuários em redes sociais. O estudo mais próximo que encontramos utiliza dados de redes de telefonia móvel para compreender a segregação e é baseado em dados étnicos e espaciais [Blumenstock & Fratamico, 2013]. Outros estudos, como Mislove et al. [2011]; Pennacchiotti & Popescu [2011], já previram etnia ou raça. Enquanto Pennacchiotti & Popescu [2011] usou uma abordagem mais sofisticada com base na identificação de tópicos e textos, Mislove et al. [2011] basearam sua análise no sobrenome das pessoas. A variedade de raças presentes na sociedade brasileira torna etnia pouco correlacionada com a classe social, como em outras partes do mundo. Assim, uma abordagem baseada em sobrenomes, por exemplo, não seria válida.

## 2.4 Caracterização comportamental de usuários

Um comportamento relevante que devemos identificar e eliminar, tanto quanto possível, é a atuação de usuários *spammers* em nossa formação de amostra. Por usuários *spammers* podemos considerar indivíduos ou organizações com objetivo de disseminar, por exemplo, propagandas, pornografia, mentiras ou vírus de computador, ou seja, atitudes que tendem a provocar ruídos em nossas análises.

No estudo realizado por Benevenuto et al. [2010] houve uma extensa coleta de características referentes aos tweets, links e informações sobre as contas de usuários, cujos *tweets* continham palavras típicas ou URL's referentes a termos *trending topics*. Os usuários foram rotulados manualmente como sendo *spammers* ou não *spammers* e uma posterior avaliação sob a técnica de aprendizagem supervisionada SVM foi produzida. Além disso, foram empregadas as técnicas de ganho de informação e  $\chi^2$  (*Chi Square*) para identificação da importância dos atributos utilizados. Como resultados, podemos

destacar que aproximadamente 70% dos usuários *spammers* e 96% de não-*spammers* foram corretamente classificados.

O trabalho apresentado em Ratkiewicz et al. [2011] realiza uma investigação sobre mensagens políticas postadas no Twitter a fim de encontrar *memes* postados por indivíduos ou organizações com o objetivo de espalhar informações mentirosas e difamatórias, denominados *Astroturf*. De forma geral, a técnica utilizada consiste na avaliação da dinâmica de difusão das mensagens e não propriamente no conteúdo. Desta forma, características como a origem, URL's embutidas e o intervalo entre mensagens sucessivas foram alvo de análise na tentativa de comprovar a hipótese de que estágios iniciais de mensagens exibem padrões comuns e que, portanto, permitem identificar *memes Astroturf*. Por fim, foram apresentados resultados em que *memes* suspeitos eram classificados por meio de um aprendizado supervisionado com base em recursos extraídos da topologia das redes de difusão, análise de sentimento e anotações *crowdsourced*.

Lumezanu et al. [2012] estuda o comportamento de propagandistas no Twitter identificando padrões nos *tweets* que poderiam caracterizar este tipo de comportamento. Propagandistas, de acordo com este trabalho, são usuários que publicam constantemente mensagens com conteúdo tendenciosos. Entre os padrões identificados estão: (i) envio de grandes volumes de *tweets* em curtos períodos de tempo; (ii) *retweeting* ao publicar conteúdo pouco original; (iii) *retweets* publicados rapidamente; e (iv) conluio com outros usuários publicando mensagens duplicadas sobre o mesmo tema simultaneamente.

Em Ghosh et al. [2012] os autores procuraram identificar como ocorre as *link farms*, ou seja, aquisição de seguidores por parte dos usuários *spammers* na rede social Twitter. Após uma coleta de dados, observou-se quais contas foram suspensas pelo Twitter e dentre estas quais postaram URL's presentes em uma lista negra, para identificar os usuários *spammers*. Assim, foi observada a existência de um pequeno grupo de usuários legítimos que, para acumular capital social, acabam seguindo indiscriminadamente os usuários que o seguem, comportamento que é explorado pelos *spammers*.

Após coletar informações sobre cerca de 1,000 usuários do Twitter escolhidos aleatoriamente, tais como suas conexões de seguidos/seguidores e os 100 *tweets* mais recentes, o trabalho descrito em [McCord & Chuah, 2011] analisou o conjunto de dados obtidos para identificar usuários *spammers* usando quatro classificadores: *Random Forest*, *Support Vector Machine* (SVM), *Naive Bayes* e *K-Nearest Neighbor* (KNN). Dentre os atributos identificados estão: (i) distribuição de *tweets* em 24 horas, subdivididos em oito períodos de 3 horas; (ii) número de URL's; (iii) respostas/menções; (iv) uso de palavras chave; (v) *retweets* e (vi) *hashtags*.

Assim, como resultado, o trabalho encontrou como melhor classificador o *Random Forest*, com uma precisão de 95.7%. Os demais algoritmos testado: SVM, *Naive Bayes* e KNN, obtiveram precisão de 93.5%, 91.6% e 92.8% respectivamente.

## 2.5 Análise de sentimentos

Como observado nos trabalhos relacionados, uma grande deficiência recorrente em monitoramento e previsões de eventos é a falta de mecanismos utilizados para distinguir, por exemplo, uma menção a um candidato de um voto, um comentário sobre um filme da intenção de assisti-lo no cinema, entre outros. Assim, a solução proposta neste trabalho é utilizar um método para análise de sentimentos nesta tarefa.

O trabalho realizado por Turney [2002] apresenta um algoritmo de classificação supervisionado para a análise de comentários, por exemplo, na avaliação automotiva ou de filmes. Esta classificação se dá pela média semântica da ocorrência de adjetivos e advérbios de uma frase, considerando como orientação positiva termos com boas associações e negativa quando ocorrem más associações. Para estimar a orientação semântica das frases foi utilizado o algoritmo PMI-IR. A técnica apresentada alcançou uma precisão média de 74%.

Um sistema de análise de sentimentos para mensagens do Twitter, denominado TwiSent, é apresentado em Mukherjee et al. [2012]. Neste trabalho, os autores abordam alguns problemas inerentes da tarefa de identificar sentimentos positivos, negativos e os objetivos dos tweets, tais como: (i) presença de usuários *spammers*; (ii) anomalias no texto como, por exemplo, grafia incorreta; (iii) a especificidade de entidades no contexto do tema pesquisado; e (iv) paradigmas incorporados nos textos.

O sistema apresentado é constituído pelas seguintes partes: (i) coleta dos últimos 200 tweets dos usuários; (ii) detector de polaridade: determinada por uma votação pela maioria dos léxicos encontrados, com base em dicionários de termos em inglês; (iii) filtro de usuários *spammers*; (iv) verificador ortográfico e normalização de texto: foi criada uma lista de abreviaturas e textos ruidosos normalmente encontrados em mensagens do Twitter; (v) manuseio pragmático como, por exemplo, alongamento de palavras e o uso de hashtags e emoticons; (vi) especificidade de entidade para separar a opinião sobre diferentes entidades em uma mesma mensagem em que é utilizado, por exemplo, a seguinte hipótese: palavras estritamente relacionadas se reúnem, portanto se separa entidades por *stop-words*. Desta forma, o sistema com detecção de *spammers* obteve uma precisão de 71.50% frente a 54.45% obtidos quando foi considerada somente a polaridade negativa ou positiva dos tweets. Também foi verificado que alguns artefatos

utilizados como, por exemplo, o corretor ortográfico são importantes ferramentas na análise dos tweets.

Saindo da análise puramente léxica, em Silva et al. [2011] os autores estudaram uma abordagem para análise de sentimentos que leva em consideração ambientes como o Twitter, em que as mensagens são curtas e disponibilizadas em um fluxo de dados constante. Tais limitações exigem que os classificadores operem com recursos limitados, incluindo dados rotulados para formação do modelo de treinamento. Desta forma, a técnica proposta consiste em iniciar o treinamento com um conjunto pequeno de amostras e, por meio de regras de associação, incorporar novas mensagens com o passar do tempo. Além disso, a técnica proposta considera a escolha de conjuntos de treinamento orientados às características qualitativas das mensagens analisadas e também a eliminação de informações irrelevantes e ultrapassadas. Assim, esta estratégia permitiu um ganho de previsão que varia entre de 7% a 58%.

Ambientes de conteúdo textual dinâmico, que são caracterizados por mudança no vocabulário que dificultam a manutenção de uma lista de adjetivos e advérbios analisados como positivos e negativos, também são considerados desafios para os métodos de análise de sentimentos. Calais Guerra et al. [2011] trabalha com a ideia de medir o viés de usuários acerca de um tema usando os endossos realizados, por exemplo, por meio de *retweets*. Como resultado eles observaram que conhecer o viés de apenas 10% dos usuários gera um nível de precisão F1 variando entre 80% a 90% em prever o sentimento do usuário em *tweets*.

A estratégia adotada tem como pressuposto que os detentores de opinião tendem a expressar várias vezes seus sentimentos e de forma consistente, ou seja, não costumam mudar abruptamente de comportamento. Além disso, considera-se que usuários semelhantes compartilham viés semelhante. O problema de prever o viés dos usuários foi modelado como um problema de aprendizagem relacional, em que, utilizou-se um grafo de acordo de opiniões, onde os usuários próximos tendem a ser semelhantes e a quantificação se dá por meio de pesos nas arestas proporcionais ao endosso para um determinado conjunto de utilizadores. Em seguida, foi elaborada uma estratégia para transferência de conhecimento, propagando o viés do usuário para termos associados com o conteúdo. Então, a combinação dos vieses é utilizada para calcular a polaridade global do conteúdo.

O trabalho realizado em Hu et al. [2013] compreende um estudo sobre o problema de análise de sentimentos em cenários não supervisionados com o uso de sinais emocionais. São sinais emocionais utilizados: (i) indicadores de emoção que refletem a polaridade do sentimento de uma mensagem como, por exemplo, os emoticons; e (ii) correlação de emoção capazes de refletir a correlação entre mensagens ou palavras

como, por exemplo, pela teoria da consistência, que sugere que palavras que ocorrem simultaneamente, principalmente em mensagens curtas, possuem a mesma orientação de sentimento. Enfim, são utilizadas informações que relacionam o sentimento de mensagens e palavras para inferir automaticamente rótulos de sentimentos para as mensagens.

Uma comparação abordando oito diferentes métodos para análise de sentimentos é realizada por Gonçalves et al. [2013]. Neste estudo, os autores averiguaram os nível de cobertura, ou seja, fração de mensagens com sentimentos identificados, e a fração dos sentimentos que foram corretamente identificados. Foi verificado que nenhum dos métodos avaliados pode ser considerado o melhor independente da fonte do texto. Após isso, também foi desenvolvido um método competitivo que combina algumas das abordagens existentes.

Tendo em mente que nenhuma técnica estudada é livre de falhas ou considerada a melhor, utilizaremos neste trabalho uma abordagem híbrida, onde um comitê de classificadores decidirá pela maioria dos votos a polaridade da mensagem analisada.

## 2.6 Relevância estatística de dados da API do Twitter

O trabalho realizado por Morstatter et al. [2013] tem como objetivo compreender melhor o impacto causado pelas coletas realizadas por meio da *streaming* API do Twitter, que fornece no máximo 1% dos de todos os tweets produzidos em um determinado momento utilizando um meio de um método de amostragem não documentado.

Assim, os autores utilizam dados coletados com parâmetros referentes ao contexto político ocorrido na Síria no período de 14 de dezembro de 2011 a 10 de janeiro de 2012, e comparam estatisticamente dimensões como as *hashtags*, temas, rede de usuários e geolocalização das mensagens entre a coleta realizada pela API e um conjunto de todos os tweets postados no período, fornecido pelo *Firehouse*, mais custoso de ser obtido.

Verificou-se que nem sempre a amostra obtida pela *streaming* API possui o mesmo nível de cobertura dos dados *Firehouse*, nem mesmo de amostras aleatórias formadas sobre os dados *Firehouse*, indicando alguma tendência na maneira que a *streaming* API fornece os dados. Com isso, fica demonstrado que deve haver uma precaução por parte dos pesquisados com a origem dos dados utilizados, conforme a dimensão que se pretende analisar. Porém, sendo a coleta via API a única disponível gratuitamente, ela é utilizada nesse trabalho.



## Capítulo 3

# PODEReS (Pesquisas de Opinião com Dados Extraídos de Redes Sociais)

O principal objetivo deste trabalho é criar um arcabouço teórico e ferramental que permita a realização de pesquisa de opinião por meio de dados coletados em redes sociais e validar sua eficácia diante dos métodos tradicionalmente utilizados. Diante deste contexto e do conhecido adágio “informação é poder”, elaboramos uma metodologia para a realização de tais avaliações, denominada PODEReS (Pesquisas de Opinião com Dados Extraídos de Redes Sociais).

Conforme esquema apresentado na Figura 3.1, a metodologia empregada consiste em três etapas fundamentais: (i) preparação dos dados, (ii) caracterização das mensagens e dos usuários e (iii) formação de uma amostra representativa da população com análise dos resultados obtidos, em que a segunda etapa representa a maior contribuição deste trabalho.

Se avaliarmos os trabalhos que se propuseram a realizar pesquisas de opinião por meio de dados coletados em redes sociais, normalmente baseados em técnicas de contagem, observamos que nenhum deles levou em consideração pontos importantíssimos em análises dessa natureza: (i) características comportamentais dos usuários como, por exemplo, desconsiderar usuários *spammers* e usuários que postam conteúdos jornalísticos; e (ii) o sentimento das mensagens postadas, uma vez que o texto coletado normalmente não está estruturado.

Além disso, a importância de uma formação de amostra representativa para pesquisas de opinião já foi historicamente comprovada e um ferramental estatístico foi amplamente desenvolvido para analisar características como os níveis de confiança e

margens de erro obtidas com tal processo. Uma amostra representativa é aquela em que os componentes escolhidos possuem características proporcionalmente condizentes o conjunto completo da população como, por exemplo, sexo, idade, localidade, entre outros.

A escolha de uma amostra representativa de dados de redes sociais (assumindo que essa amostra existe, hipótese respaldada por relatórios demográficos apresentados por empresas tais como a Ignite Social Media<sup>1</sup>), que represente com alto grau de semelhança a população real, é um desafio.

Assim, para a formação de uma amostra para pesquisa *online* propomos três etapas: (i) caracterização pessoal de usuários, a fim de que a amostra seja representativa da população; (ii) caracterização comportamental de usuários, de forma a eliminar usuários denominados *spammers* e também aqueles que postam somente conteúdos jornalísticos; e (iii) análise de sentimentos, para determinarmos o significado semântico das mensagens postadas.

Maiores detalhes das etapas de preparação dos dados, caracterização comportamental, análise de sentimentos e análise estatística são mostradas nas seções 3.1, 3.2, 3.3 e 3.4 respectivamente. As etapas para caracterização pessoal, maior contribuição deste trabalho, são mostradas separadamente no capítulo 4.

## 3.1 Preparação dos dados

O Twitter é uma rede social mundial que permite a troca de informações, em tempo real, sobre os mais diversificados assuntos. Nesta ferramenta as mensagens, denominadas *tweets*, possuem um tamanho máximo de 140 caracteres e podem conter URL's (*links* para páginas Web), referências a outros usuários (indicada pelo símbolo @) e *hashtags* com a finalidade de caracterizar um assunto (indicado pelo símbolo #). Outra característica desta ferramenta está na forma em que as relações sociais são desenvolvidas. Diferentemente de outras redes sociais, em que se tem a relação não direcionada de amizade, no Twitter existe a relação de seguidos e seguidores, sem a obrigatoriedade de coexistência entre as ligações.

Este trabalho utilizou duas bases de dados coletadas na rede social Twitter, sendo a primeira referente às eleições municipais brasileiras de 2012, em 6 capitais, e a segunda com mensagens relacionadas a 14 votações do *reality show Big Brother Brasil 13*, denominadas “paredões”.

Neste contexto, após definidas as cidades e “paredões” analisados foram criados

---

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



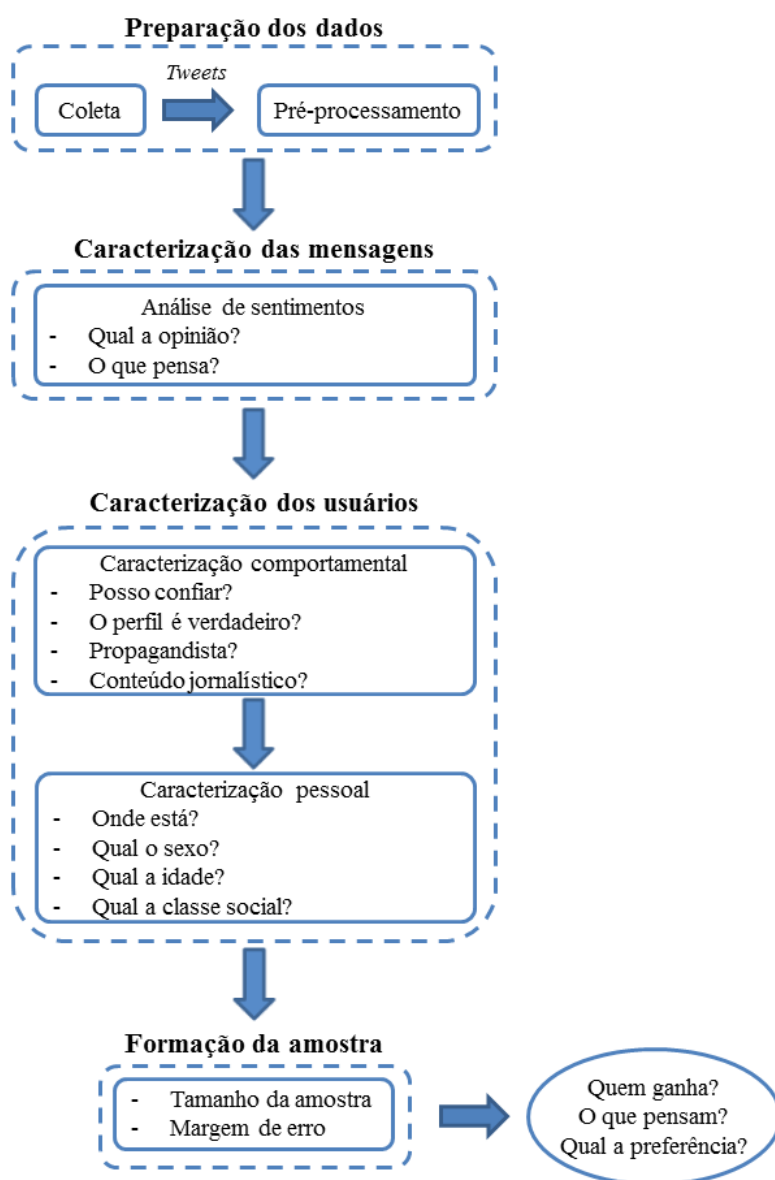


Figura 3.1: Esquema para pesquisa de opinião em redes sociais

conjuntos de termos capazes de identificar cada entidade de interesse. Tais termos constituem principalmente nos nomes ou apelidos dos candidatos. A etapa seguinte constituiu em coletar os *tweets* relacionados a estes termos. Esta coleta foi realizada por meio da API disponibilizada pelo próprio Twitter<sup>2</sup>, em que, no fluxo de mensagens, são filtrados aquelas de interesse.

Algumas etapas da técnica proposta exigirão informações sobre os usuários, tais como faixa etária, sexo e localização geográfica. Quando disponíveis publicamente essas informações são usadas. Caso contrário, são inferidas. Para inferência, coletamos os

<sup>2</sup><http://dev.twitter.com/>

últimos 200 *tweets* postados pelos usuários identificados na primeira fase de coleta e o primeiro nível da rede de amizades, pares de usuários que mutuamente seguem e são seguidos, cultivadas por eles.

O pré-processamento do texto dos *tweets* consistiu na remoção de acentos, pontuações e palavras conhecidas como *stop-words*, que representam, por exemplo, a classe de palavras dos artigos. Em alguns casos, como para verificar a frequência de erros gramaticais na caracterização da idade, foi utilizado o texto completo e sem modificações.

## 3.2 Caracterização comportamental

Como mostrado na Figura 3.1, nesta etapa estamos interessados em responder perguntas como: i) posso confiar no usuário? O perfil é verdadeiro? É um propagandista? É um difusor de conteúdo jornalístico?

Esta análise serve para excluirmos de nossa amostra usuários indesejáveis, ou seja, aqueles em que os *tweets* não representam a opinião de um indivíduo, mas que possuem interesses de outras finalidades. Assim, temos as seguintes definições:

**Def. 1** - Usuários *spammers*: indivíduos ou organizações com interesse de difundir vírus, pornografia ou propagandas.

**Def. 2** - Usuários jornalísticos: indivíduos ou organizações que publicam conteúdos jornalísticos, ou seja, estão interessados apenas na difusão de notícias e não de opinião de indivíduos específicos.

**Def. 3** - Usuários legítimos: conjunto de todos os usuários coletados, excluindo os usuários *spammers* e os usuários jornalísticos.

### 3.2.1 Identificação de usuários *spammers*

Para identificarmos esta categoria de usuários foi utilizada a base de dados desenvolvida por Benevenuto et al. [2010], em que foram manualmente rotulados 355 usuários como *spammers* e 710 como não *spammers*, formando um conjunto total de 1,065 usuários. Tais usuários são descritos por 62 atributos retratando, por exemplo, a média de *tweets* postados com URL e a média de *tweets* postados com *hashtags*.

Inicialmente focamos nosso trabalho em encontrar um bom método para distinguir usuários *spammers* dos não *spammers*. Assim, foram utilizados os seguintes algoritmos: (i) *Support Vector Machine* (SVM), (ii) *Multilayer Perceptron* (MLP), (iii) *k-nearest Neighbor* (KNN) e (iv) *Random Forest* (RF). Escolhemos estes algoritmos

por serem considerados estado da arte na tarefa de classificação ou, no caso do KNN, pelo baixo custo computacional exigido por suas operações.

Todos os métodos utilizados são supervisionados, ou seja, utilizam a informação dos rótulos, previamente identificados, para amostras de treino na tarefa de aprendizagem e para as amostras de teste na etapa de generalização.

Para o SVM utilizaremos um ambiente idêntico ao descrito no trabalho de Benevenuto et al. [2010], ou seja, será utilizado um SVM não linear com kernel *Radial Basis Function* (RBF) a partir da implementação fornecida pelo pacote libSVM. Os parâmetros também serão otimizados conforme descrito por Benevenuto et al. [2010], ou seja, será utilizada a ferramenta *easy*, também provida pelo pacote libSVM e os atributos numéricos serão normalizados.

Utilizaremos implementações disponibilizadas pela ferramenta Weka [Witten & Frank, 2005] nos algoritmos MLP, KNN e *Random Forest*. Para otimizar os parâmetros foi utilizado a estratégia de projeto simples, ou seja, ignoramos a interação entre os fatores, variando um de cada vez e fixando o valor que proporcionou melhor resultado. Os valores finais que foram utilizados são mostrados na Tabela 3.1.

Tabela 3.1: Algoritmos de classificação - parâmetros utilizados

Algoritmo	Parâmetros
SVM	$c = 32.0$ e $g = 0.0078125$
MLP	taxa de aprendizagem = 0.5, camadas escondidas = 1 e momentum = 0.1
KNN	$k = 10$ e métrica de similaridade = $1/\text{distância}$
RF	número de árvores = 20 e profundidade = 10

Em todos os testes foi utilizado o método de validação cruzada. Tal método consiste na divisão do conjunto de dados em  $k$  partes distintas. Para que nosso resultado pudesse ser comparado ao obtido em Benevenuto et al. [2010] o valor de  $k$  foi definido como 5.

Como podemos observar, pelos resultados do processo de classificação apresentados na Tabela 3.2, o classificador que se sobressaiu foi o *Random Forest*, tanto em termos de média F1 quanto em acurácia.

### 3.2.1.1 Importância dos atributos

Na base de dados utilizada cada usuário é representado por um total de 62 atributos. Para avaliarmos a importância de tais atributos e, com isso, tentarmos diminuir a dimensão do problema trabalhado avaliaremos os atributos utilizando os seguintes métodos: (i) Ganho de informação, (ii) *Chi Squared* ( $\chi^2$ ) e (iii) Algoritmo Genético (AG). A quantidade de atributos desejada foi fixada em 10 e 20.

Tabela 3.2: Resultados para classificação de usuários *spammers*

Algoritmos	$f1$	Acurácia (%)
Todos os atributos		
SVM	0.87	85.51
MLP	0.87	87.79
KNN	0.86	86.95
<i>Random Forest (RF)</i>	0.88	88.64
10 atributos selecionados		
RF - Algoritmo genético	0.90	90.23
RF - $\chi^2$	0.87	87.42
RF - Ganho de informação	0.87	87.79
20 atributos selecionados		
RF - Algoritmo genético	0.90	90.33
RF - $\chi^2$	0.87	87.51
RF - Ganho de informação	0.87	87.42

A redução da dimensionalidade dos problemas tem como principais benefícios:

1. Remover atributos irrelevantes, redundantes ou ruidosos;
2. Melhorar o desempenho dos algoritmos de classificação, tanto em termos de tempo de processamento quanto em qualidade das repostas.

Diferentemente dos demais algoritmos de seleção de atributos utilizados, que consideram a importância de cada atributo separadamente, o AG é capaz de considerar as interações entre os diversos atributos. Assim, ao final de um número limite de gerações o melhor indivíduo, ou seja, aquele de melhor *fitness*, tende a encontrar a melhor combinação possível entre os atributos.

Representaremos a *fitness* dos indivíduos pela média harmônica  $F1$  obtida quando os atributos selecionados são submetidos ao algoritmo de classificação *Random Forest*, algoritmo que alcançou melhores resultados para o conjunto completo de atributos. Escolhemos a métrica  $F1$  por ela obter resultados mais representativos em bases com classes desbalanceadas, como em nosso caso.

Para selecionar os melhores parâmetros para o AG, novamente optamos por um projeto simples. Primeiramente foram avaliados a variação do número de indivíduos e geração e, com esse resultado, avaliou-se a variação da probabilidade de cruzamento e, após isso, a probabilidade de mutação. Por ser um método estocástico o algoritmo implementado para o AG foi executado 10 vezes para cada cenário. Os parâmetros finais foram definidos como: (i) 10 atributos: 150 indivíduos, 25 gerações, 0.8 a probabilidade

de cruzamento e 0.1 a probabilidade de mutação; (ii) 20 atributos: 200 indivíduos, 45 gerações, 0.8 probabilidade de cruzamento e 0.05 a probabilidade de mutação.

### 3.2.1.2 Avaliação

Como mostrado na Tabela 3.2, a combinação dos 20 melhores atributos selecionados com o AG classificados pelo algoritmo *Random Forest* alcançou os melhores resultados, tanto em termos de média F1 quanto de acurácia. Para validarmos estatisticamente os resultados alcançados podemos realizar o teste *student t*, ou simplesmente teste t, utilizando a seguinte equação:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \quad (3.1)$$

onde  $\bar{x}$  representa a média F1 dos melhores indivíduos alcançados no final da execução do algoritmo genético,  $\mu_0$  a hipótese alternativa que em nosso caso é o valor médio do F1 obtido pelo baseline SVM, ou seja, 0.9114,  $s$  é o desvio padrão da nossa amostra e  $n$  o tamanho da amostra.

Para a seleção pelo AG de 10 atributos, temos:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.9281 - 0.9114}{\frac{0.0016}{\sqrt{10}}} = 33.01 \quad (3.2)$$

Já para a seleção de 20 atributos, temos:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.9294 - 0.9114}{\frac{0.0008}{\sqrt{10}}} = 71.15 \quad (3.3)$$

Observando o valor tabelado para um nível de confiança de 0.99% e 9 graus de liberdade, que é 2.821, podemos concluir que, em média, os valores obtidos pelos métodos avaliados são diferentes. Em termos práticos, alcançamos melhores resultados com a seleção de atributos pelo algoritmo genético utilizando *Random Forest* do que o método concorrente.

Ao avaliar pelo mesmo teste t a seleção de atributos com o ganho de informação (IG) e  $\chi^2$ , com uma posterior execução pelo algoritmo *Random Forest*, também chegamos a resultados semelhantes, ou seja, os atributos selecionados pelo algoritmo genético alcançou melhores resultados.

Além disso, como mostrado pela Tabela 3.3, os atributos selecionados pelo ganho de informação e  $\chi^2$  foram idênticos, tanto entre os 10 primeiros quanto entre os 20, mudando apenas a ordem de importância. Já o AG conseguiu obter uma maior taxa de diversidade, apenas 3 idênticos nos 10 melhores selecionados, são eles: (4) fração

de *tweets* com url's, (62) idade da conta do usuário e (30) número médio de url's em cada *tweet*. Entre os 20 selecionados 8 foram idênticos, que são os mesmos dos citados para 10 atributos adicionados aos seguintes: (22) número médio de menções por *tweet*, (25) número máximo de menções por *tweet*, (10) número médio de url's por número de palavras em cada *tweet*, (53) tempo máximo entre as mensagens e (2) fração de *tweets* respondidos.

Observando apenas os atributos selecionados com o algoritmo genético podemos ver que 50% dos atributos escolhidos nos 10 melhores estão presentes entre os 20 melhores, são eles: (4) fração de *tweets* com url's, (3) fração de *tweets* com as palavras de spam, (62) idade da conta do usuário, (30) número médio de url's em cada *tweet* e (25) número máximo de menções por *tweet*.

Uma descrição completa sobre o significado de cada numeração mostrada na Tabela 3.3 está disponível em: <http://homepages.dcc.ufmg.br/~fabricio/spammerscollection.html>.

Tabela 3.3: Comparação atributos selecionados

Posição	<i>Chi squared</i>	Ganho de informação	AG (10)	AG (20)
1	4	4	4	22
2	62	30	3	59
3	30	62	62	26
4	1	1	37	30
5	2	2	39	25
6	48	48	30	56
7	47	42	25	17
8	42	43	34	33
9	43	47	32	3
10	6	6	55	24
11	22	10	-	62
12	10	37	-	52
13	45	49	-	15
14	18	45	-	36
15	49	22	-	10
16	37	18	-	4
17	25	25	-	7
18	46	16	-	53
19	53	53	-	60
20	16	46	-	2

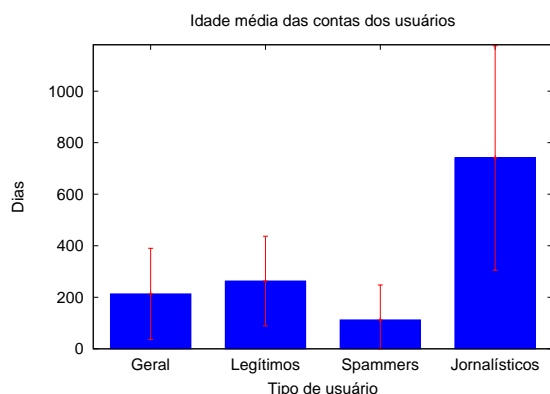


Figura 3.2: Idade média das contas dos usuários

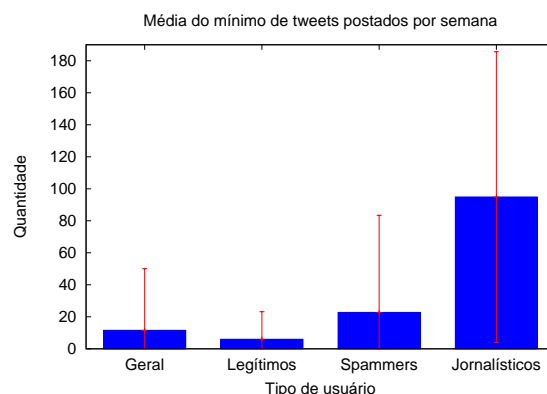


Figura 3.3: Média do mínimo de *tweets* postados por semana

### 3.2.2 Identificação de usuários jornalísticos

A base de dados para a segunda categoria de usuários, jornalísticos, foi formada por 70 usuários coletados que possuíam a substring “notícia” em seus nomes da conta do Twitter. Estes usuários foram descritos pelos mesmos 62 atributos definidos para a classe *spammer*. Assim, a base de dados utilizada passa a possuir 710 usuários rotulados como não *spammers*, 355 como *spammers* e 70 jornalísticos, formando uma base geral com 1,135 usuários.

Para verificarmos se os atributos utilizados para descrever a base inicialmente são suficientes para distinguir os usuários anteriores dos jornalísticos verificamos as médias, com respectivos desvios padrão, de alguns deles, conforme listado abaixo:

1. Idade média da conta dos usuários: Como mostrado no gráfico da Figura 3.2 usuários legítimos possuem, em média, contas aproximadamente duas vezes mais antigas que usuários *spammers* e duas vezes mais recentes que os usuários jornalísticos, que possuem contas muito mais consolidadas.
2. Número médio do mínimo de *tweets* por semana: contas de usuários *spammers* postam um número significativamente maior de *tweets* por semana quando comparados à média geral e aos usuários legítimos, mas substancialmente menor quando comparados com os usuários jornalísticos, como mostrado no gráfico da Figura 3.3.
3. Fração média de *tweets* respondidos: como pode ser observado pelo gráfico da Figura 3.4, os usuários *spammers* e jornalísticos possuem uma relação interpessoal muito menos intensa que os usuários legítimos.

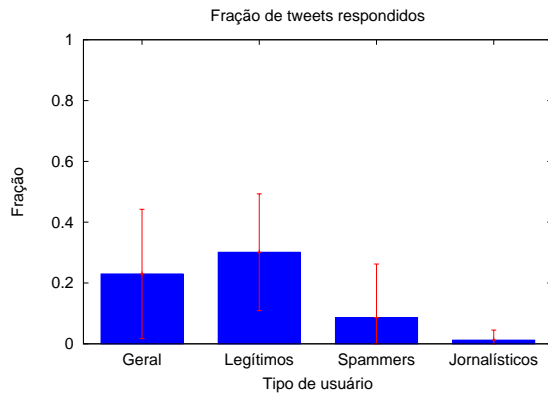


Figura 3.4: Fração média de *tweets* respondidos

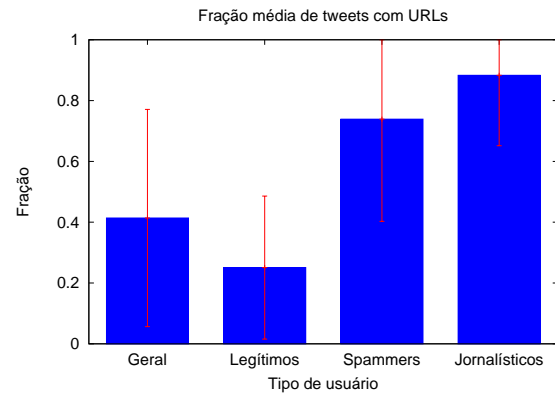


Figura 3.5: Fração média de *tweets* com URLs

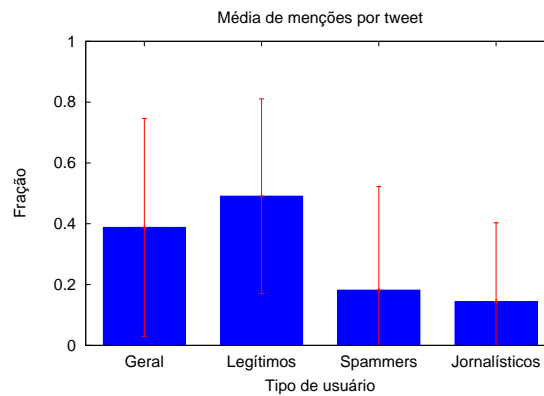


Figura 3.6: Média de menções por *tweet*

4. Fração média de *tweets* com URLs: como mostrado no gráfico da Figura 3.5 os usuários *spammers* e jornalísticos tendem a divulgar um número bastante superior de URLs a cada *tweet* postado.
5. Número médio de menções por *tweet*: usuários legítimos fazem uma quantidade significativamente maior de menção a outros usuários que os usuários *spammers* e jornalísticos, como mostrado na Figura 3.6.

Assim, podemos verificar que os atributos utilizados para descrever os usuários contrastam bem as diferentes classes analisadas e que, apesar de pequena, a amostra de usuários jornalísticos é suficiente para captar padrões de comportamentos desta classe de indivíduos.



### 3.2.3 Método resultante

Como mencionado anteriormente o melhor método identificado para separar as classes de *spammers* e não *spammers*, melhor taxa de acerto e média F1, foi o *Random Forest*. Para reduzirmos a dimensionalidade do problema e com isso tentarmos melhorar a taxa de acerto dos algoritmos o algoritmo genético se sobressaiu aos concorrentes analisados, obtendo atributos mais diversificados. Os atributos selecionados foram:

- Número mínimo/máximo/médio de menções por *tweet*
- Número mínimo/médio de *tweets* postados por semana
- Número médio de caracteres numéricos por *tweet*
- Número máximo/médio de URLs em cada *tweet*
- Número mínimo de *tweets* postados em um dia
- Número máximo/mediano de caracteres por *tweet*
- Fração de *tweets* com palavras *spam*
- Idade da conta
- Tempo mínimo/máximo entre mensagens
- Número mínimo de palavras por *tweet*
- Número médio de URL por número de palavras em cada *tweet*
- Fração de *tweets* com URLs
- Número mediano de *hashtags* por número de palavras em cada *tweet*
- Fração de *tweets* respondidos

Os resultados obtidos são apresentados no gráfico da Figura 3.7 e mostra não haver sobreposição dos intervalos de confiança para o melhor conjunto atributos selecionados/*Random Forest*, logo este é estatisticamente o melhor resultado, com 99% de confiança. Observe que a escala do eixo y varia entre 0.84 a 0.9.

Como a relação dos termos necessários para analisar o atributo “fração de *tweets* com palavras *spam*”, ou seja, palavras que constavam no *trending topics* do Twitter no período da coleta, não estava disponível foi necessário substituí-lo. O atributo escolhido para isto foi “número de seguidores por seguidos”.

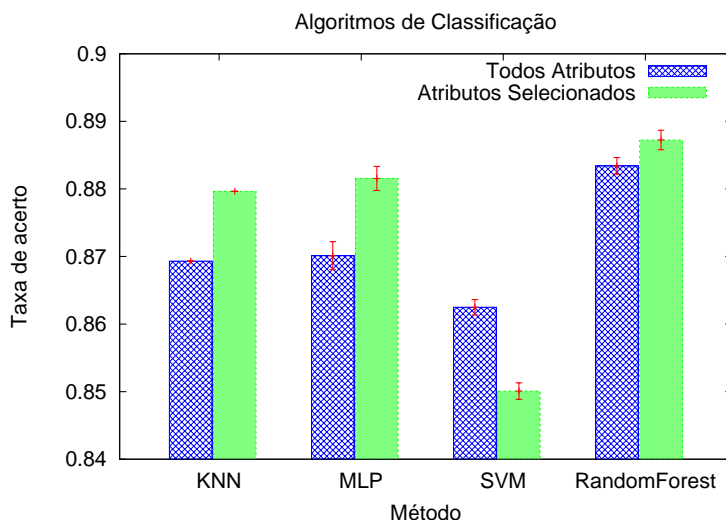


Figura 3.7: Algoritmos para análise comportamental

Os usuários da categoria jornalísticos, que possuíam a substring “noticia” em seus nomes da conta do Twitter, foram descritos nos mesmos atributos definidos para as classes anteriores. A execução do método com o algoritmo *Random Forest*, utilizando os atributos selecionados, no conjunto rotulado contendo usuários *spammers*, jornalísticos e legítimos resultou em uma taxa média de acerto de 88.77% e média F1 de 0.88.

Observando os usuários que foram rotulados posteriormente como jornalísticos observamos que o classificador se comportou bem, com erros concentrados em considerar usuários que postam muito e que possuem contas antigas como jornalísticos, quando nem sempre eram.

Assim, a etapa de análise comportamental consistirá no algoritmo *Random Forest* treinado com a base rotulada contendo usuários *spammers*, jornalístico e legítimos, descritos pelos 20 atributos selecionado pelo AG.

### 3.3 Análise de sentimentos

Como mostrado na Figura 3.1, nesta etapa estamos interessados em responder perguntas como: i) qual a opinião do indivíduo? O que o indivíduo pensa? Para tanto, será realizada uma análise de sentimentos nos *tweets* coletados.

Foi observado na literatura, conforme mostrado na seção 2.5, que existe uma grande gama de métodos para realização de análises desta natureza. Vimos também, na comparação realizada por Gonçalves et al. [2013], que existe uma grande dependência da técnica utilizada como a fonte do texto verificado, não sendo encontrado nenhuma

que ganhasse em todos os cenários. Logo, considerando as inúmeras aplicações práticas possíveis para aplicar tais técnicas, este ainda é um campo fértil para novos estudos.

Entre os métodos mais utilizados para a realização de análise de sentimentos em mensagens de *micro-blogs* estão os seguintes:

- Dicionários léxicos: conjunto de palavras previamente rotuladas como positivas, negativas ou neutras, em que, conforme as ocorrências indicam as polaridades das mensagens analisadas.
- Algoritmos supervisionados com mensagens classificadas manualmente.
- Algoritmos supervisionados com rótulos baseados em indicadores de emoção, como os *emoticons*.

Neste trabalho, analisamos mensagens que se referem apenas a uma entidade, ou seja, único candidato ou participante do programa. Além disso, utilizaremos uma abordagem baseada em um comitê de classificação, onde os *tweets* serão rotulados conforme a polaridade indicada pela maioria de três classificadores: léxico, supervisionado com rotulação manual e supervisionado com rotulação automática. Cada classificador atribuirá +1, caso a mensagem seja positiva, ou -1, caso seja negativa. Se a soma for positiva a mensagem é considerada positiva, se a soma for negativa a mensagem é considerada negativa e em caso de soma zero a mensagem é descartada das análises. Mais detalhes sobre tais classificadores são apresentados nas subseções a seguir.

### 3.3.1 Dicionário Léxico

Como mencionado anteriormente, o dicionário léxico é formado por um conjunto de palavras rotuladas como representantes de sentimentos positivos, negativos ou neutros. No idioma inglês existem diversos dicionários bem consolidados para tal tarefa como o MPQA *Opinion Corpus*<sup>3</sup> e o *General Inquirer*<sup>4</sup>, mas o mesmo não ocorre em português, idioma alvo deste trabalho.

Para o português encontramos na literatura dicionários, por exemplo, constituídos a partir da extração de termos em *blogs* com comentários sobre componentes de veículos [Ribeiro Jr. et al., 2012], entre outros. Uma prática comum que também é realizada em trabalhos deste tipo é expandir o dicionário inicialmente desenvolvido por dicionários de sinônimos de palavras.

---

<sup>3</sup>[http://www.cs.pitt.edu/mpqa/opinionfinder\\_1.html](http://www.cs.pitt.edu/mpqa/opinionfinder_1.html)

<sup>4</sup><http://www.wjh.harvard.edu/~inquirer/>

Neste trabalho rotulamos um conjunto inicial de palavras específicas para cada cenário com sentimentos positivos ou negativos. Este conjunto foi definido observando os *tweets* coletados em cada contexto. Para as eleições verificou-se que normalmente são usados termos para designar a conduta política do candidato ou a vontade do eleitor. São palavras como “ladrão” e “corrupto” para negativos e “apoio” e “voto” para positivos. Já para o BBB 13, verificamos a existência de termos mais pessoais para designar os participantes como “barraqueira” e “mentirosa” como negativos e “engraçada” e “bonita” para positivos.

Após isso, expandimos tais conjuntos pelo dicionário de sinônimos Dicio<sup>5</sup> e removemos duplicatas, ou seja, termos que aparecem nas duas listas finais. Um resumo quantitativo dos termos rotulados é apresentado na Tabela 3.4.

Tabela 3.4: Dicionários Léxico - número de palavras

	Eleições	BBB 13
Positivo	194	251
Negativo	292	175

### 3.3.2 Algoritmo supervisionado - Rotulação manual

Algoritmos supervisionados utilizam informações obtidas por meio de conteúdos rotulados, previamente identificados, para amostras de treino na tarefa de aprendizagem e, então, generalizam as observações feitas para classificar novos dados.

Nesta etapa, utilizamos como treinamento um conjunto de 200 *tweets* distintos escolhidos aleatoriamente e rotulados manualmente como positivo ou negativo em cada cenário analisado. Quatro classificadores serão avaliados: i) SVM; ii) *Random Forest*; iii) *Naive Bayes*; e iv) *Naive Bayes* Multinomial.

Os resultados obtidos, considerando um processo com validação cruzada de 5 *folds*, são apresentados na Tabela 3.5. Considerando que a melhor taxa de acerto em todas as bases foi obtida pelo algoritmo *Naive Bayes* este será utilizado nesta etapa da análise de sentimentos dos *tweets*.

### 3.3.3 Algoritmo supervisionado - Rotulação automática

Conforme mostrado por Hu et al. [2013] indicadores de emoção sugerem que o sentimento contido em uma mensagem é consistente com os sinais presentes nela. Neste

---

<sup>5</sup><http://www.dicio.com.br>

Tabela 3.5: Rotulação manual - Acurácia (%)

	Eleições	BBB13
SVM	62.5	51
<i>Random Forest</i>	75.5	57.5
<i>Naive Bayes</i>	80.5	62
<i>Naive Bayes</i> Multinomial	78	60.5

trabalho utilizamos a lista de *emoticons* mostrados na Tabela 3.6 para indicar a polaridade de um *tweet* como positivo ou negativo.

Com isso, escolhemos para cada cenário analisado 1,000 *tweets* distintos de forma aleatória para cada classe, resultando em 2,000 *tweets*, e prosseguimos com a análise dos algoritmos de classificação apresentados na subseção 3.3.2.

Os textos de cada mensagem passaram por um pré-processamento, em que, foram retiradas palavras *stop-words*, pontuações, *emoticons* e termos com frequência menor que três no conjunto, resultando em um total de 1,062 diferentes palavras no cenário das eleições e 916 no BBB 13.

Tabela 3.6: Lista de *Emoticons*

Positivo	:)	:-)
Negativo	:(	:-(

Os resultados obtidos, considerando um processo com validação cruzada de 5 *folds*, é apresentado na Tabela 3.7. Considerando que a melhor taxa de acerto em todas as bases foi obtida pelo algoritmo *Naive Bayes* Multinomial este será utilizado nesta etapa da análise de sentimentos dos *tweets*.

Tabela 3.7: Rotulação automática dos tweets selecionados com *emoticons* - Acurácia (%)

	Eleições	BBB13
SVM	65.85	60.65
<i>Random Forest</i>	69.95	65.45
<i>Naive Bayes</i>	67.8	64.15
<i>Naive Bayes</i> Multinomial	74.8	69.15

### 3.3.4 Concordância entre os algoritmos de análise de sentimentos

Para verificarmos a concordância entre os três métodos utilizados no comitê de classificação, utilizamos os tweets rotulados manualmente para os cenários das Eleições e BBB e procedemos pela verificação da compatibilidade entre os rótulos atribuídos.

Foram comparados quatro cenários para cada uma das bases utilizadas: (i) concordância entre os três algoritmos: Léxico (L), Rotulação Manual (RM) e Rotulação automática (RA); (ii) concordância entre L e RM; (iii) concordância entre L e RA; e (iv) concordância entre RM e RA.

Considerando que nem todas as mensagens avaliadas possuem termos de sentimento catalogados pelo dicionário léxico iremos desconsiderar desta análise as mensagens classificadas como neutras.

O algoritmo RM é treinado com as bases de tweets rotuladas manualmente. Para não testarmos este algoritmo com a mesma base de treinamento realizamos um processo de validação cruzada com cinco partições, onde tweets testados não faziam parte do conjunto de treinamento.

Os resultados alcançados são mostrados nos gráficos das Figuras 3.8, 3.9, 3.10 e 3.11, para o cenário das eleições, e Figuras 3.12, 3.13, 3.14 e 3.15, para o cenário do BBB. Como podemos observar, a comparação entre os rótulos atribuídos pelos três métodos foram 34% diferentes entre os tweets da base Eleições e 75% diferentes na base do BBB, ou seja, existe um nível diferente na complexidade de atribuir sentimentos para as mensagens desses dois contextos.

Os métodos que mais concordaram foram L e RM na base das Eleições, alcançando 82%, e L e RA na base do BBB, alcançando 51%. Os métodos com maior discordância foram RM e RA na base das Eleições, divergindo em 30%, e na base do BBB houve um empate entre L e RM com a combinação RM e RA, divergindo em 51%.

Estes resultados refletem a grande complexidade em atribuir sentimentos para as mensagens dos dois contextos analisados. Motivos para tanto são a forte presença de ironias e piadas, em ambos os cenários. Um exemplo para este tipo de comportamento é a seguinte mensagem constantemente reproduzida durante o período de coleta das eleições: “Amor sim, Russomanno não”. Um léxico que catalogasse o termo “Amor” como sendo de sentimento positivo estaria automaticamente errando o rótulo atribuído para a mensagem.

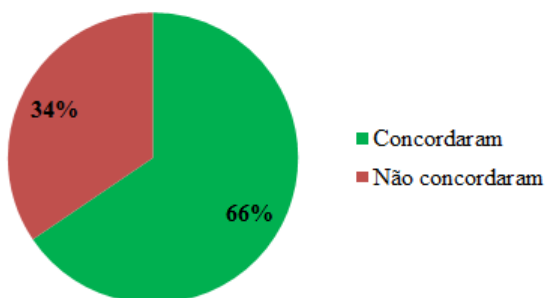
**Eleições - Concordância (L, RM e RA)**

Figura 3.8: Eleições - concordância entre L, RM e RA

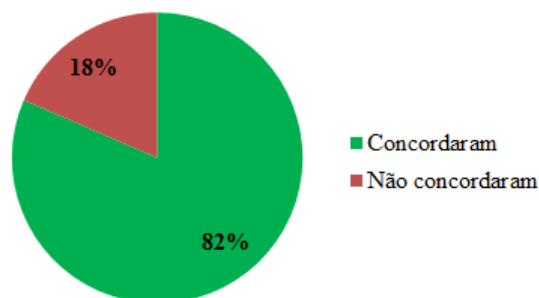
**Eleições - Concordância (L e RM)**

Figura 3.9: Eleições - concordância entre L e RM

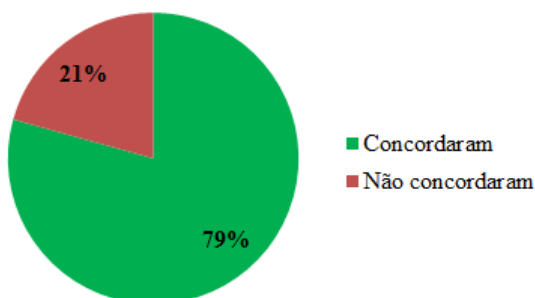
**Eleições - Concordância (L e RA)**

Figura 3.10: Eleições - concordância entre L e RA

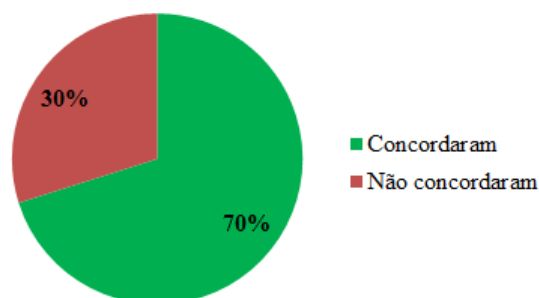
**Eleições - Concordância (RM e RA)**

Figura 3.11: Eleições - concordância entre RM e RA

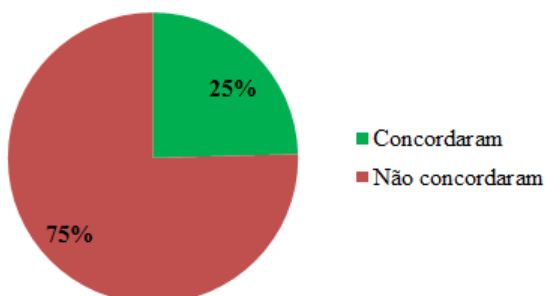
**BBB - Concordância (L, RM e RA)**

Figura 3.12: BBB - concordância entre L, RM e RA

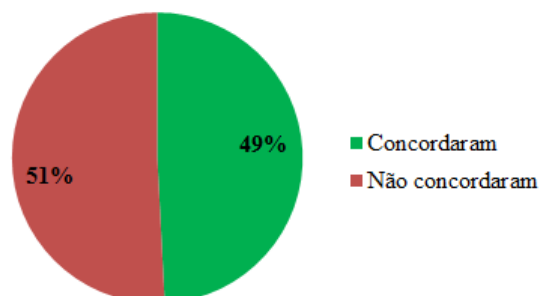
**BBB - Concordância (L e RM)**

Figura 3.13: BBB - concordância entre L e RM

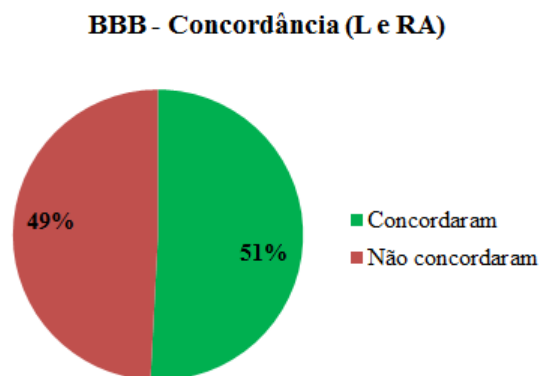


Figura 3.14: BBB - concordância entre L e RA

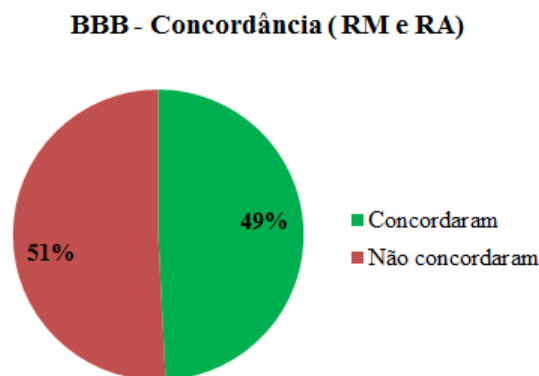


Figura 3.15: BBB - concordância entre RM e RA

### 3.3.5 Avaliação do método

Para avaliarmos os métodos selecionados para a análise de sentimentos utilizamos 200 tweets distintos para cada cenário, BBB e Eleições, escolhidos aleatoriamente e rotulados manualmente como positivo ou negativo. Assim, avaliamos a cobertura, ou seja, porcentagem de tweets não classificados como neutros, e a acurácia provida por cada método. Os resultados obtidos são mostrados na Tabela 3.8.

Como podemos observar o dicionário léxico construído foi capaz de prover uma boa acurácia. No entanto, tal método possui uma baixa cobertura: apenas 43% das mensagens no contexto das Eleições e 31% do BBB foram classificadas como positivas ou negativas.

O método capaz de prover a melhor acurácia foi o classificador supervisionado com mensagens de treinamento rotuladas manualmente, mas devido a pouca quantidade de tweets rotulados optamos por utilizar o Comitê de classificação em nosso trabalho, por ser capaz de alcançar uma boa acurácia e uma cobertura razoável.

Tabela 3.8: Análise de sentimentos - avaliação do método (Métricas de acurácia e cobertura em porcentagem (%))

Método	Eleições		BBB	
	Acurácia	Cobertura	Acurácia	Cobertura
Dicionário Léxico (L)	74.7	43	70.5	31
Supervisionado manualmente (RM)	80.5	100	62.0	100
Supervisionado automaticamente (RA)	63.5	100	49.5	100
Comitê	80.3	79	59.1	64



## 3.4 Tamanho da amostra

Pesquisas com objetivos de verificar atitudes, valores e opiniões da população sem entrevistar todos os indivíduos, ou seja, com formação de amostra, são historicamente estudadas e uma estrutura para tal atividade é bem definida em diversos trabalhos encontrados na literatura.

Além da representatividade dos indivíduos, com proporções condizentes com o conjunto completo da população, um dos principais aspectos que devem ser observados para esta atividade é o tamanho da amostra que será utilizada. Este tamanho irá interferir diretamente em dimensões como a margem de erro obtida no processo.

Margem de erro é a diferença máxima provável entre a medida do estimador observado na amostra e o verdadeiro valor encontrado na população [Bracarense, 2009]. Por exemplo, se aceitarmos uma margem de erro de 5% em um resultado estimado em 60% valor real poderá variar em entre 55 a 65%.

Assim, a margem de erro, tamanho da amostra e confiança obtida nas pesquisas de opinião, em que o tamanho da população é infinito ou desconhecido, podem ser definidos pela fórmula apresentada na equação 3.4, em que  $n$  representa o tamanho da amostra,  $z$  o escore aproximado da distribuição normal para a probabilidade de confiança desejada,  $d$  o desvio permitido do resultado obtido para o real e  $\sigma^2$  o valor da variância de uma pesquisa anterior.

$$n = \left( \frac{z \times \sigma^2}{d} \right)^2 \quad (3.4)$$

Neste trabalho será utilizado uma confiança de 95.5% e  $\sigma^2$  correspondente a 0.5, já que o valor da variância de uma pesquisa anterior é desconhecida. Em consonância com Arkin & Colton [1950]; Gil [2010] os valores utilizados, tamanho da amostra para cada margem de erro, são mostrados na Tabela 3.9.

Tabela 3.9: Tamanho da amostra (95.5% de confiança)

Margem de erro (%)	Tamanho da amostra
1	10000
2	2500
3	1111
4	625
5	400
6	278
7	204
8	157
9	124
10	100
11	83
12	70
13	60
14	51
15	45

# Capítulo 4

## Caracterização pessoal

Para formação de amostras representativas da população analisaremos as seguintes características pessoais nos usuários identificados em nossas coletas: sexo, idade e classe social.

Também foi realizado um estudo para inferir a localização geográfica dos indivíduos, trabalho publicado em [Rodrigues et al., 2013]. Parte deste trabalho é apresentado no Apêndice A. Este atributo não foi analisado nos cenários avaliados pois, no caso do BBB, não existia informações sobre a distribuição geográfica da população real. Para as Eleições, consideramos que o indivíduo pertencia ao mesmo município do candidato ao qual ele se referiu nas mensagens coletadas.

Os métodos analisados para o processo de caracterização foram baseados nos textos publicados pelos usuários e em atributos não textuais extraídos dos tweets, estratégia similar a adotada por diversos trabalhos encontrados na literatura [Nguyen et al., 2013; Benevenuto et al., 2010; Mahmud et al., 2012; Cheng et al., 2010]. As etapas se subdividiram basicamente nas seguintes:

1. Coleta de tweets: foram coletados, utilizando a API do Twitter, os 200 tweets mais recentemente publicados pelos usuários identificados. Usamos os últimos 200, pois é um limite que a API do Twitter nos permite coletar de forma mais simplificada. De forma geral, esses usuários possuem duas características principais: publicaram tweets públicos e em Português. Para identificarmos os usuários coletados utilizamos, normalmente, duas abordagens distintas: (i) usuários que postaram mensagens relacionadas com as Eleições 2012 ou com o programa BBB 13 e (ii) usuários que utilizaram um termo comum no idioma Português, de forma semelhante à abordagem seguida em Nguyen et al. [2013]. O termo escolhido para orientar o processo de rastreamento foi *coisa*. De acordo com um documento pu-

blicado pela Academia Brasileira de Letras<sup>1</sup>, coisa é o substantivo mais frequente usado no Português Brasileiro. Observe que primeiro coletamos os tweets e deles encontramos os usuários para, então, recuperar seus últimos 200 tweets.

2. Pré-processamento das mensagens: foram retiradas acentuações gráficas, pontuações, termos *stop words*, caracteres não ASCII e considerados todos os caracteres em minúsculo.
3. Verificação da importância dos termos utilizados: avaliamos os termos encontrados utilizando as métricas de Ganho de Informação ou pelo *Chi Squared*( $\chi^2$ ).
4. Classificação de usuários utilizando algoritmos supervisionados: resultados com quatro classificadores diferentes são apresentados: *Naive Bayes Multinomial*, *Naive Bayes*, SVM e *Random Forest*. Todos os resultados foram obtidos usando as versões Weka destes classificadores [Witten & Frank, 2005]. O *Naive Bayes Multinomial* (NBM) foi escolhido por ser extremamente rápido e apresentar bons resultados com texto e o SVM e *Random Forest* por estarem entre os classificadores estado da arte. Em todos os classificadores, exceto o SVM, foram utilizados os parâmetros padrão. Os parâmetros para o SVM foram otimizados usando o ferramenta *easy*, que realiza uma pesquisa de grade na escolha dos valores. Os experimentos foram realizados utilizando um procedimento de validação cruzada de cinco partições.
5. Avaliação dos resultados: os resultados serão avaliados conforme valores encontrados para acurácia e média F1, métrica apropriada para avaliação em conjuntos desbalanceados.

Especificidades sobre cada caracterização serão discutidas nas seções a seguir.

## 4.1 Sexo

Como mostrado na Figura 3.1, dentre as características pessoais uma que queremos identificar para as formações de nossas amostras é o sexo do usuário, ou seja, feminino ou masculino. Para tanto, o método proposto se baseia em duas etapas, são elas:

- Dicionário de nomes: conjunto de nomes normalmente utilizados para designar pessoas do sexo masculino ou feminino.

---

<sup>1</sup><http://www.academia.org.br/>

- Classificação por algoritmos supervisionados: um classificador é treinado a partir de mensagens de usuários previamente rotulados como sendo de sexo masculino ou feminino.

O método proposto consistirá na utilização das duas etapas mencionadas em sequência, ou seja, ele verificará a existência do nome do usuário no dicionário e quando o nome não estiver catalogado, rotulará o indivíduo pelo algoritmo de classificação supervisionado.

### 4.1.1 Dicionário de nomes

Para a primeira etapa, inicialmente utilizaremos uma coleta anteriormente realizada de nomes de usuários e seus respectivos sexos pela rede social Facebook<sup>2</sup>. Após isso, foi utilizado o catálogo de nomes para bebês contido no sítio BebeAtual<sup>3</sup> para ampliar o dicionário corrente. Cada nome foi identificado com as informações do sexo (masculino, feminino ou unissex) e a frequência em que o nome foi encontrado para cada um deles.

O dicionário gerado foi então pré-processado, conforme já descrito anteriormente, descartados duplicatas, nomes com menos de três caracteres e desconsiderados a informação de sobrenomes.

Como resultado foi criado um dicionário com 21,378 nomes femininos, masculinos e unissex. Após uma filtragem por apenas nomes masculinos e femininos, considerando também os nomes unissex caso a frequência em algum dos sexos fosse dez vezes superior a do outro, obtivemos um total de 20,801 nomes sendo 11,671 femininos e 9,130 masculinos.

### 4.1.2 Classificadores

Na segunda etapa, consideramos que o dicionário criado fornecia uma base confiável pela qual poderíamos classificar os usuários. Desta forma, coletamos os últimos *tweets* de usuários que postaram mensagens nos seguintes contextos: Eleições municipais de 2012, BBB 13 e usuários que seguiam os perfis *Mens Health Brasil*<sup>4</sup> e *Womens Health Brasil*<sup>5</sup>. Sendo as duas últimas fontes de usuários escolhidas por apresentarem contextos tipicamente masculinos e femininos, respectivamente.

Dentre os conjuntos de usuários coletados foram selecionados somente aqueles em que o nome foi identificado pelo dicionário de nomes como sendo do sexo masculino

---

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

<sup>3</sup><http://bebeatual.com/>

<sup>4</sup>[https://twitter.com/menshealth\\_br](https://twitter.com/menshealth_br)

<sup>5</sup><https://twitter.com/WomensHealthBR>

ou feminino, para os quais foi possível coletar no mínimo 50 *tweets*. Tais *tweets* foram pré-processamentos. Após isso, para cada usuário foi formado um vetor para identificar a presença das palavras mencionadas, com frequência mínima de três.

Os resultados alcançados pela avaliação de quatro classificadores são mostrados nas Tabelas 4.1, 4.2, 4.3, 4.4, 4.5 e 4.6. Podemos observar pelos F1 encontrados, que o SVM e o *Random Forest* obtiveram resultados muito inferiores em classes altamente desbalanceadas como é o caso das bases *Mens Health* e *Womens Health*. Já o *Naive Bayes* normalmente perde para o *Naive Bayes* Multinomial. Assim, o algoritmo *Naive Bayes* Multinomial será utilizado daqui em diante.

Tabela 4.1: Sexo *Womens Health* - F1 utilizando todos os termos

Usuários	Homem (836)	Mulher (2,971)	Média
SVM	Nenhum homem	0.877	0.438
<i>Random Forest</i>	0.171	0.870	0.520
<i>Naive Bayes</i>	0.469	0.776	0.622
<i>Naive Bayes</i> Multinomial	0.492	0.819	0.655

Tabela 4.2: Sexo *Mens Health* - F1 utilizando todos os termos

Usuários	Homem (9,513)	Mulher (2,180)	Média
SVM	0.897	0.004	0.450
<i>Random Forest</i>	0.888	0.210	0.549
<i>Naive Bayes</i>	0.775	0.422	0.598
<i>Naive Bayes</i> Multinomial	0.833	0.441	0.637

Tabela 4.3: Sexo Revistas (*Womens* e *Mens Health*) - F1 utilizando todos os termos

Usuários	Homem (10,349)	Mulher (5,151)	Média
SVM	0.863	0.634	0.748
<i>Random Forest</i>	0.802	0.536	0.669
<i>Naive Bayes</i>	0.771	0.652	0.711
<i>Naive Bayes</i> Multinomial	0.794	0.660	0.727

Para tentarmos melhorar os resultados obtidos executamos os algoritmos Ganho de Informação e  $\chi^2$ , e verificamos o melhor número de atributos, palavras digitadas, capazes de separar as duas classes de sexo dos usuários, para quantidades variando de 10 em 10. Para tanto, utilizamos a base completa de usuários, 23,100 indivíduos, e os *tweets* tratados de forma análoga ao já descrito anteriormente.

Tabela 4.4: Sexo BBB - F1 utilizando todos os termos

Usuários	Homem (518)	Mulher (788)	Média
SVM	0.446	0.79	0.617
<i>Random Forest</i>	0.412	0.744	0.578
<i>Naive Bayes</i>	0.601	0.763	0.682
<i>Naive Bayes</i> Multinomial	0.590	0.772	0.681

Tabela 4.5: Sexo Eleições - F1 utilizando todos os termos

Usuários	Homem (3,759)	Mulher (2,535)	Média
SVM	0.832	0.716	0.774
<i>Random Forest</i>	0.724	0.592	0.658
<i>Naive Bayes</i>	0.708	0.656	0.682
<i>Naive Bayes</i> Multinomial	0.727	0.654	0.690

Tabela 4.6: Sexo Base Completa - F1 utilizando todos os termos

Usuários	Homem (14,626)	Mulher (8,474)	Média
SVM	0.856	0.694	0.775
<i>Random Forest</i>	0.776	0.581	0.678
<i>Naive Bayes</i>	0.751	0.652	0.701
<i>Naive Bayes</i> Multinomial	0.754	0.653	0.703

Como mostrado pela Tabela 4.7, 20 pode ser considerado uma quantidade pequena de termos capaz de obter boa média F1 e ao mesmo tempo proporciona uma alta cobertura entre os usuários, dentre as palavras selecionadas 97.25% dos usuários utilizou no mínimo uma. As palavras são listadas abaixo:

- obrigada, obrigado, cansada, lindo, amei, adorei, adoro, amiga, amo, lt (abreviação para *Last tweet*), gol, futebol, delicia, linda, @hugogloss, lindos, libertadores, campeao, jogador e palmeiras.

Como podemos observar, os termos selecionados, em geral, apresentam sufixos masculinos (“o”) ou femininos (“a”). Além disso, termos tipicamente atribuídos a universos distintos de gêneros também foram selecionados, como palavras relacionados ao futebol.

### 4.1.3 Avaliação do método

Como mencionado anteriormente, o método proposto consiste execução em sequência de duas fases: verificação do nome no dicionário e, se necessário, classificação por um

Tabela 4.7: Sexo Base Completa - seleção de atributos

# atributos	Usuários que utilizaram os termos (%)	F1		
		Homem (14,626)	Mulher (8,474)	Média
10	93.39	0.836	0.593	0.714
20	97.25	0.812	0.651	0.732
30	98.83	0.796	0.655	0.725
40	99.36	0.785	0.647	0.716
50	99.77	0.789	0.655	0.722
60	99.77	0.781	0.650	0.715
70	99.82	0.781	0.653	0.717
80	99.88	0.780	0.654	0.717
90	99.88	0.779	0.655	0.717
100	99.88	0.774	0.655	0.714

algoritmo supervisionado. Para avaliar a eficácia de tal método isolamos, dentre os usuários coletados na base das Eleições e do BBB, aqueles que indicavam *links* para *blogs* na rede social Blogspot<sup>6</sup>, com informação do sexo do usuário.

Verificando somente os usuários cujos nomes estão presentes no dicionário, encontramos os resultados apresentados na Tabela 4.8, que retrata a corretude do dicionário desenvolvido.

Tabela 4.8: Acerto do dicionário

Base	Número de usuários analisados	Acerto (%)
Eleições	4,344	98.11
BBB13	884	97.73

Posteriormente, avaliamos o acerto obtido pelo dicionário, algoritmo supervisionado e algoritmo supervisionado com uso do dicionário. Na avaliação do método onde o dicionário é avaliado separadamente atribuímos o sexo “masculino” para todos os nomes não catalogados. Os resultados são mostrados na Tabela 4.9, em que na base coletada dos blogs BBB foram analisados 1,293 usuários e na base das eleições foram analisados 5,875 usuários. Assim, consideramos que o conjunto *Naive Bayes* Multinomial (20 atributos) + dicionário é o mais apropriado para a classificação do sexo.

---

<sup>6</sup><http://www.blogger.com/>



Tabela 4.9: Acerto das variações do método (%)

	BBB13	Eleições
Dicionário	74.40	87.76
<i>Naive Bayes</i> Multinomial (20 atributos)	78.11	80.14
<i>Naive Bayes</i> Multinomial (20 atributos) + dicionário	90.02	92.54

## 4.2 Idade

Como mostrado na Figura 3.1, uma característica que queremos identificar nos usuários analisados, para formação de nossas amostras, é a idade. Diversos trabalhos já abordaram este assunto e verificaram que tanto a forma de linguagem dos indivíduos, como a maior recorrência no uso de preposições ou artigo, quanto as relações interpessoais, como a criação de vínculos de amizades, são características capazes de diferenciar indivíduos de diferentes faixas etárias.

Assim, inicialmente mostraremos a estratégia utilizada para formação da nossa base de dados, identificação e avaliação do poder discriminativo de atributos textuais e não textuais e, por fim, um estudo sobre o poder preditivo de alguns classificadores estado da arte nesta tarefa.

### 4.2.1 Base de dados

A base de dados utilizada nesta etapa foi construída a partir de dois processos distintos, são eles:

1. Coleta em blogs:

Usuários identificados nas coletas BBB 13, Eleições 2012 ou que usaram a palavra “coisa” em suas mensagens, cujos *tweets* continham indicação para um blog da rede Blogspot. Assim, foi realizado um *crawler* sobre tais blogs e extraída as idades dos usuários pelo campo descrição com a expressão “X anos”, onde X representa um valor entre 1 e 99. Para garantir que a expressão continha informação da idade do indivíduo e não, por exemplo, de tempo de experiência profissional, cada texto foi analisado e filtrado manualmente.

2. Coleta manual:

Foi utilizada a ferramenta de busca do Twitter<sup>7</sup> e por meio de pesquisas pelos termos “tenho X anos” ou “fiz X anos”, onde X representa um valor entre 10 e

<sup>7</sup><https://twitter.com/search-home>

99, analisou-se manualmente se a mensagem tratava de uma experiência pessoal e se a fotografia apresentada no perfil era condizente com a idade mencionada, identificando, desta forma, usuários com suas respectivas idades.

Além disso, para a classe de usuários com idade superior a 45 anos, foram realizadas pesquisas pelos termos “sou aposentado”, “aposentado”, “aposentadoria”, “netinhos” ou “cobap” (Confederação dos Aposentados e Pensionistas do Brasil) e por seguidores dos usuários de *screen names* “terceira\_idade”, “blogda3idadesp”, “aTerceiraIdade” e “nucleo3idade”. Tais indivíduos foram adicionados à base quando as fotografias apresentadas no perfil evidenciavam uma pessoa com aparência condizente a esta faixa etária.

Como resultado, obteve-se uma base com 1,709 usuários rotulados. Categorizando estes indivíduos em cinco faixas etárias, de forma compatível com as faixas etárias divulgadas pelo TSE para o perfil do eleitorado, sendo 290 com idade menor que 16 anos, 806 entre 16 a 24, 284 entre 25 a 34, 158 entre 35 a 44 e 171 acima de 45. Finalmente, foram coletados os últimos *tweets* de cada usuário identificado.

## 4.2.2 Características textuais e não textuais

Tendo uma base de dados com usuários rotulados, podemos identificar as características que são capazes de distinguir pessoas de diferentes faixas etárias, um passo para a construção de um método automático de inferência da idade. Analisaremos basicamente dois grupos de características: textuais e não textuais.

### 4.2.2.1 Atributos textuais

Para extrair características textuais, utilizamos os 200 tweets mais recentes postados pelos usuários. Inicialmente, consideramos os atributos mais discriminativos em todo o texto. Em uma segunda fase, com base em nossa intuição, observação das palavras mais frequentemente utilizadas em cada idade e seguindo exemplos encontrados em trabalhos relacionados, identificamos um conjunto de termos característicos de cada faixa etária, retratando por exemplo, eventos cotidianos, lugares e expressões típicas de cada etapa da vida.

Nesta segunda fase, para cada característica considerada relevante um conjunto de termos foi manualmente definido e suas frequências médias e medianas foram avaliadas, conforme listado abaixo:

- Filmes até 16 anos: crepusculo, edward, bella, jacob, harry potter, hermione, voldemort, dumbledore.

- Ídolos até 16 anos: fresno, nx zero, restart, fiuk, avril lavigne, justin bieber, hanna montana, miley cyrus, #justinbieber, luan, luan santana, justin, taylor, demi, demi lovato, selenia gomes, mile cyrus, onedirection, liam, nial, zayn.
- Televisão até 16 anos: rebelde, rbd, high school musical, hsm, glee, malhao.
- Bar: bar, buteco, boteco, botequim, butequim, barzin.
- Bebidas: bebi, vou beber, estou bebendo, to bebendo, beberei, encher a cara, enche a cara, estou bebado, to bebado, estou tonto, to tonto, tomar uma, tomar todas.
- Faculdade: faculdade, graduacao, universidade, facul.
- Maquiagem: blush, makeup, delineador, po compacto, sombra, rimel, batom, demaquilante, gloss, lapis de olho, lapis de boca.
- Abreviações: sdv, mds, amr, agr, favor, favoor, bgs, ngm, facul, sdds, flw, pfvr, awn, scrr.
- Diversão noturna: boate, balada, discoteca.
- Política: partidos, cassado, mandato, #eleicao, @brazilnocorrupt, petista, senadora, governo, politica, dilma, pt, psdb, pmdb, justica, partido, stf, presidente, campanha, ministro, lula, aecio, corrupto, corrupcao.
- Relacionamento: apaixonada, xonada, xonei, namorado, namorada, amando.
- Religião: deus, papa, santo, cristo, santa, jesus, morte, francisco.
- Escola: escola, colegio, aula de matematica, aula de fisica, aula de redacao, aula de portugues, aula de geografia, aula de historia, recreio, atv de.
- Casamento: festa de casamento, festa de noivado, despedida de solteiro (a), casamento, cha de panela, noivado, meu esposo (o), meu marido.
- Termos frequentes mais de 45 anos: louco, macon, maconaria, the voice, #thevoicebrasil, miss, #missbrasil, redacao, leitura, plantao, coluna, olhos, asilo, gps, cartal, portal, @veja, inss, carta, editora, previdencia, lazer, cardapio, senadora, aposentado, aposentados, aposentadoria, previdenciario, neto, netos, netinhos, terceira idade.

Além destes termos, para cada faixa etária considerada, definimos um conjunto de gírias e internetês normalmente utilizadas. Como dicionário inicial destas palavras coletamos termos do blog Dicionariopopular<sup>8</sup>, um catálogo de gírias e expressões populares, e de Linguadedoido<sup>9</sup>, um catálogo de internetês. Tais catálogos também passaram pelo processo de pré-processamento.

Para atribuímos cada palavra dos dicionários às faixas etárias, seja ela gíria ou internetês, calculamos o *tf-idf* (*term frequency-inverse document frequency*) de cada termo nos documentos constituídos pela concatenação dos textos de todos usuários de cada faixa definida. Foram considerados os 20 termos com maior valor para esta métrica, que se diferenciavam dos 20 melhores termos das demais faixas etárias. Os termos selecionados na divisão em três faixas etárias são mostrados abaixo:

#### Gírias:

- ate 25 anos: brusa, mauricinho, cachola, la longe, painho, fica na sua, krau, nao se toca, dimais, minha loira, chorado, kenga, bredo, mijao, sangue frio, bafafa, atazana, sai do meu pe, cocozinho e mais alem;
- 25 a 45 anos: pantera, picuma, cantarolar, ficar de molho, cumadi, cafe pequeno, verdinha, rala rala, bater um rango, istamu, colala, pipocada, aos trancos e barrancos, proceis, pe da vida, malinar, paranho, zoreia, desmantelo e acalentar;
- mais 45 anos: espocar, manda bala, cacarejam, pincelada, fomento, acabar em pizza, vapt, trancoso, vapt vupt, buchudo, tempestade em copo dagua, de varga, dar um piti, correno, cacoete, nem relógio trabalha de graça, apeia, nem que a vaca tussa, sistemático e de vento em popa.

#### Internetês:

- ate 25 anos: chegay, mels, morenin, beim, pokin, fikei, pufavo, longi, qrendo, obgda, fossi, moziin, xego, pkna, pusha, brink s, paliacada, esqeci, xorar e baxin;
- 25 a 45 anos: xover, xeid, fikr, veix, flei, voley, anaum, pulus, dexha, futs, bilhetin, maix, noix, cntg, amgo, nads, dmais, qlqr, hoji e umilia;
- mais 45 anos: deshar, marka, pueira, seol, decha, naum, huahuahua, kero, mlke, kbca, nunk, meldels, novis, anju, xeru, cursin, boua, rasho, okay e noes.

Outros atributos textuais analisados foram:

<sup>8</sup><http://dicionariopopular.blogspot.com.br/>

<sup>9</sup><http://linguadedoido.blogspot.com.br/2008/07/dicionrio-de-internets.html>

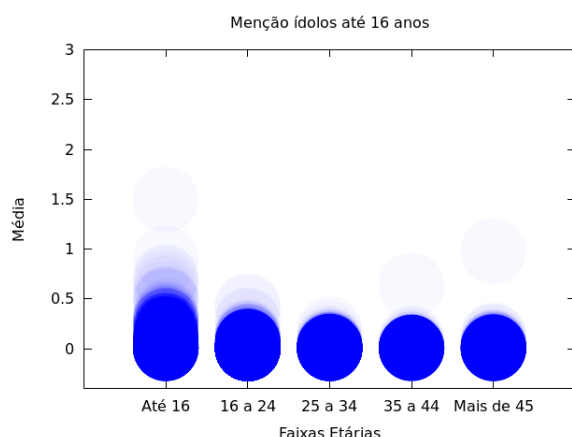


Figura 4.1: Média de menção a ídolos até 16 anos por usuário

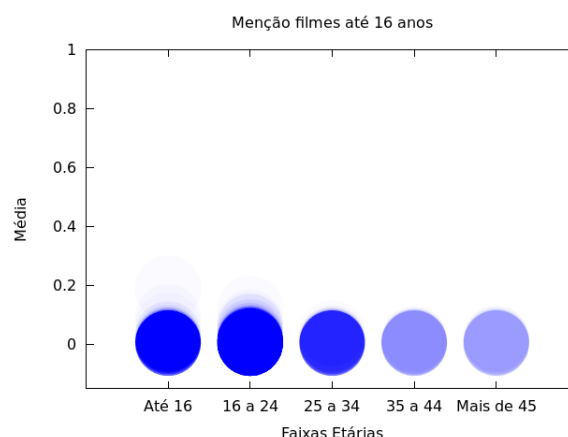


Figura 4.2: Média de menção a filmes até 16 anos por usuário

- Média e mediana de termos positivos: conjunto de palavras manualmente selecionadas e expandidas pelo dicionário de sinônimos Dicio<sup>10</sup>. São termos como: bom, admiro e adoro.
- Média e mediana de termos negativos: conjunto de palavras criado de forma análoga às positivas. São termos como: péssimo, raiva e desprezível.
- Média e mediana de erros ortográficos;
- Média e mediana do uso de preposições;
- Média e mediana do uso de artigos;
- Média e mediana do uso de emoticons.

Para exemplificar o poder discriminativo dos conjuntos de termos criados são mostrados nos gráficos das Figuras 4.1, 4.2, 4.3 e 4.4, as médias por usuário de menções a termos dos conjuntos “ídolos até 16 anos”, “filmes até 16 anos”, “diversão noturna” e “termos frequentes mais de 45 anos”, respectivamente. Estes gráficos foram construídos de forma a retratar por cada circunferência a média de menções que o usuário utilizou em termos dos conjuntos em seus tweets mais recentes. Foram desconsiderados usuários que não mencionaram os termos dos conjuntos, ou seja, média igual a 0. Desta forma, quanto mais escuro o ponto mais usuários possuem o mesmo comportamento.

Como podemos observar, os conjuntos selecionados retratam o comportamento de usuários em diferentes etapas da vida, com maior ocorrência dos três primeiros conjuntos entre os mais jovens e do último entre os mais velhos.

<sup>10</sup><http://www.dicio.com.br>

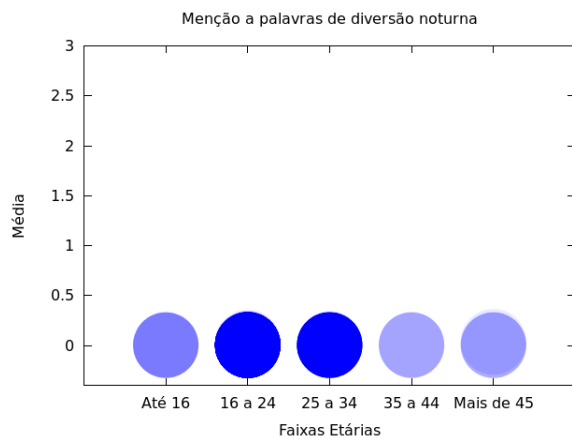


Figura 4.3: Média de menção a palavras de diversão noturna por usuário

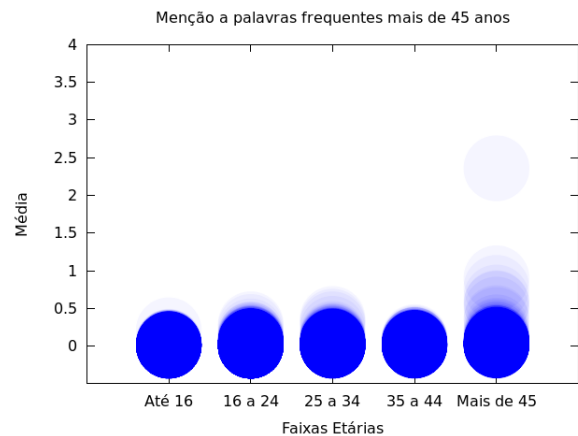


Figura 4.4: Média de menção a palavras frequentes mais de 45 anos por usuário

#### 4.2.2.2 Atributos não textuais

Complementando os atributos textuais apresentados na subseção anterior, também extraímos e analisamos os seguintes atributos não textuais:

- Sexo: inferido conforme método apresentado na seção 4.1;
- Média e mediana de hiperlinks: URL's identificadas;
- Média e mediana de hashtags: indicadas pelo símbolo “#”;
- Média e mediana do número de caracteres nas palavras;
- Média e mediana do número de caracteres nos tweets;
- Média e mediana da frequência de tweets por dia;
- Média e mediana da frequência de tweets por semana;
- Média e mediana da frequência de tweets por mês;
- Média e mediana do número de palavras por tweet;
- Divulgação de coordenadas geográficas nos tweets publicados;
- Preenchimento de cidade válida: considerando cidades brasileiras;
- Número de seguidores;
- Número de seguidos.

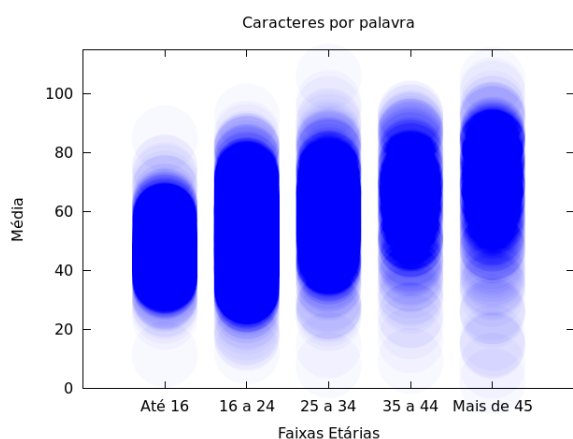


Figura 4.5: Média do tamanho das palavras por usuário

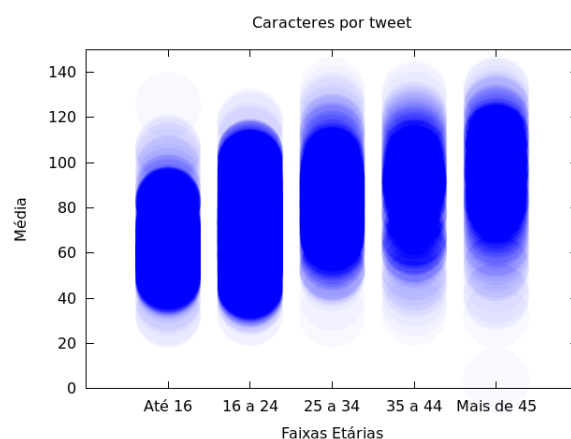


Figura 4.6: Média do tamanho dos tweets por usuário

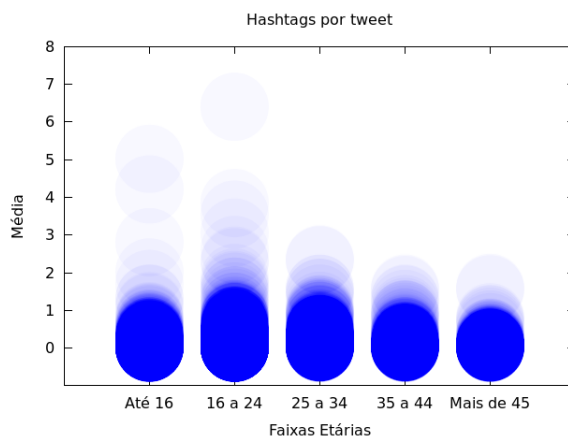


Figura 4.7: Média do número de hashtags por usuário

Seguindo o mesmo procedimento já explicado para os gráficos de atributos textuais o comportamento dos usuários em alguns dos atributos não textuais são mostrados nos gráficos das Figuras 4.5, 4.6 e 4.7, para os tamanhos médios de palavras, tweets e uso de hashtags, respectivamente. Podemos observar que pessoas mais velhas tendem a escrever palavras e tweets com mais caracteres, e que o uso de hashtags é maior entre os usuários mais jovens.

### 4.2.3 Classificação

Depois de analisar os recursos textuais e não textuais que podem distinguir os usuários de diferentes faixas etárias, esta seção utiliza esses recursos para criar modelos de classificação capazes de discriminar os usuários. Os experimentos foram realizados

Tabela 4.10: Principais termos para divisões de idade em 3 e 5 classes, conforme medida de Ganho de Informação

3 Classes	5 Classes
lt	lt ( <i>Last tweet</i> )
governo	governo
uu	uu (Expressão para comemoração)
dormir	politica
politica	dormir
sdds	agr (Agora)
ne	amo
vou	sdds (Saudades)
haha	pt (Partido dos trabalhadores)
tava	mds (Meu Deus)

considerando os usuários divididos em 3 e 5 faixas de idade, sendo a de 3 com as idades menor que 25, entre 25 e 45 e maior que 45 e a de 5 com as idades menor que 16, 16 a 24, 25 a 34, 35 a 44 e maior de 45.

Diferentes combinações dos recursos foram utilizadas para descrever os dados, incluindo o texto completo (em que foram observadas a presença de 5,135 termos na divisão por 3 classes e 2,859 em 5 classes, formado pelas palavras com frequência maior que 5 e Ganho de Informação maior que 0), atributos textuais selecionados, atributos não textuais selecionados e uma combinação dos atributos textuais e não textuais selecionados.

Os 10 termos mais relevantes, conforme medida de Ganho de Informação computada com o texto completo, são mostrados na Tabela 4.10, em que podemos observar uma grande presença de palavras relacionadas ao mundo político, abreviações e internetês.

Como podemos observar pela Tabela 4.11, o conjunto de classificador/atributos que conseguiu alcançar melhores resultados nos dois recortes de faixas etárias foi o *Naive Bayes Multinomial* (NBM) utilizando o texto completo postado pelos usuários. No entanto, os atributos textuais e não textuais selecionados, dado a significativa diminuição de termos avaliados, também obtiveram bons resultados quando utilizado em conjunto com o algoritmo SVM como, por exemplo, perdendo os textuais selecionados em apenas 0.1 de f1 na divisão em três classes. Isto retrata que se definidos mais cuidadosamente os termos melhores resultados poderão ser alcançados.



Tabela 4.11: Resultados obtidos utilizando o texto completo, atributos textuais selecionados, não textuais selecionados e todos selecionados (textuais com não textuais) - métrica de acurácia mostrada em porcentagem (%)

3 Classes								
	Texto completo		Textuais		Não textuais		Todos selecionados	
	<i>f1</i>	Acurácia	<i>f1</i>	Acurácia	<i>f1</i>	Acurácia	<i>f1</i>	Acurácia
<i>NBM</i>	0.81	81.51	0.51	64.42	0.55	64.31	0.55	64.36
<i>Naive Bayes</i>	0.78	77.12	0.68	66.88	0.63	67.41	0.71	71.50
SVM	0.70	74.84	0.74	75.48	0.67	70.63	0.76	76.18
<i>Random Forest</i>	0.66	70.51	0.71	72.79	0.66	68.46	0.71	73.26
5 Classes								
	Texto completo		Textuais		Não textuais		Todos selecionados	
	<i>f1</i>	Acurácia	<i>f1</i>	Acurácia	<i>f1</i>	Acurácia	<i>f1</i>	Acurácia
<i>NBM</i>	0.66	66.06	0.32	47.75	0.36	41.49	0.36	42.19
<i>Naive Bayes</i>	0.61	61.20	0.51	49.74	0.42	44.88	0.52	52.95
SVM	0.43	53.66	0.57	60.15	0.38	50.15	0.58	60.39
<i>Random Forest</i>	0.46	51.02	0.52	55.18	0.46	49.33	0.53	55.00

### 4.3 Classe social

Como mostrado na Figura 3.1, uma característica que queremos identificar nos usuários analisados, para formação de nossas amostras, é a classe social na qual ele se enquadra, característica ainda não explorada em outros trabalhos.

Na teoria sociológica, o conceito de classe social é muitas vezes contestado [Crompton, 2008] e já foi amplamente discutido por sociólogos, como Weber e Bourdieu [Bourdieu, 1987]. Em particular, a definição de classe social pode ser subjetiva ou objetiva [Schiffman, 2007]. Dada uma definição simplista, “classes são categorias sociais que compartilham atributos subjetivamente marcantes usados para classificar pessoas dentro de categorias em um sistema de estratificação econômica” [Wright, 2003]. Estes atributos subjetivos podem ser definidos pela ocupação, educação, estilo de vida, níveis de renda, ou uma combinação entre eles. Em uma definição mais objetiva as classes se referem a “como as pessoas estão localizadas na desigualdade da distribuição dos bens materiais” [Wright, 2003]. Neste caso, a renda e riqueza são determinantes.

Levando em consideração esta definição objetiva podemos supor que pessoas dentro de uma determinada classe social tendem a ter estilos de vida semelhantes em virtude de seus níveis de renda e gostos em comum [Solomon, 2010; Katz-Gerro, 1999; Settle et al., 1979]. Neste contexto, este trabalho associa estilo de vida a riqueza dos bairros que as pessoas costumam visitar para rotular automaticamente usuários brasileiros em diferentes classes sociais.

Mais especificamente, inicialmente foi coletada uma lista de locais (ou *venues*) de uma das mais populares redes sociais baseadas em localização - *Foursquare*, e etiquetamos cada local de acordo com a riqueza do bairro em que ele está localizado. A rede *Foursquare* trabalha com 9 diferentes categorias para os locais, são elas: arte e lazer, faculdade e educação, alimentação, café e outros lugares, vida noturna, residências, parques e locais abertos, lojas e serviços e viagens e transportes. Consideramos todas estas categorias, como prova relevante para inferir a classe social de um usuário e, assim, exploramos a riqueza do bairro de todos os lugares visitados por ele em um determinado período de tempo.

A riqueza de um bairro foi definida de acordo com a renda per capita dos moradores, dado obtido no censo de 2010 realizado pelo Instituto Brasileiro de Geografia e Estatística [IBGE, 2010]. De acordo com a frequência de visitas a diferentes bairros e sua respectiva riqueza, nós rotulamos os usuários em classes diferentes. Experimentamos cenários com duas, três e quatro divisões de classe social. Em seguida, exploramos diferentes recursos textuais que podem distinguir os usuários em classes diferentes, incluindo hábitos de consumo, entretenimento, meios de transporte, entre outros. Estas características são, em uma terceira fase, usadas para gerar um modelo de classificação. Dependendo do número de classes, o classificador obtém taxas de precisão variando de 57,09% a 73,74%.

### 4.3.1 Base de dados

Esta seção mostra como a base de dados com usuários rotulados de acordo com a classe social foi criada. Assim, são detalhadas duas fases do processo: (i) seleção de usuários e (ii) bairros de usuários rotulados.

#### 4.3.1.1 Seleção de usuários

Para construir a nossa base de dados foram selecionados usuários do Twitter, com três características principais: (i) brasileiros, (ii) com os tweets públicos, e (iii) pelo menos uma interação *Foursquare* em uma das 23 capitais brasileiras pré-selecionadas nos últimos 200 tweets. Usamos os 200 tweets mais recentes, pois é um limite que a API do Twitter nos permite coletar de forma simplificada. Interações *Foursquare* incluem *check-ins* (o usuário diz aos amigos que está em um determinado lugar), *tips* (mensagens com dicas e opiniões de usuários sobre um determinado lugar) e *mayorships* (título dado para o usuário mais frequente em um determinado local nos últimos 60 dias). Cada interação *Foursquare* tem uma localização associada (latitude e longitude), que pode ser mapeada em um lugar específico (por exemplo, o bairro de uma cidade).

A interação do usuário em um determinado local é uma evidência de que o usuário visitou aquele lugar particular.

Como a maioria dos usuários do Twitter não publicam mensagens do *Foursquare* em suas contas (nossa coleta de dados mostra que, considerando todo o território brasileiro, 30% dos usuários brasileiros postaram interações do *Foursquare* no Twitter), o rastreamento de dados via a API do Twitter, dadas todas as suas restrições, é lento. Por isso, foram utilizadas duas estratégias distintas: (i) o uso de um termo comum no idioma Português para coletar os tweets, de forma semelhante à abordagem seguida em Nguyen et al. [2013], (ii) a utilização de uma caixa delimitadora na API do Twitter, considerando as coordenadas das 23 cidades previamente definidas.

O termo escolhido para orientar o processo de rastreamento foi *coisa*. De acordo com um documento publicado pela Academia Brasileira de Letras<sup>11</sup>, coisa é o substantivo mais frequente usado no Português Brasileiro. Observe que primeiro coletamos os tweets e deles encontramos os usuários para, então, recuperar seus últimos 200 tweets. Além dos tweets, para usuários com contas no *Foursquare*, também foram coletadas suas *tips* e *mayorships*. Check-ins só foram obtidos a partir de mensagens no Twitter, dadas as restrições de acesso por parte da API do *Foursquare*.

Desta forma, foram coletados 426,001 usuários que utilizam o termo *coisa* de 8 setembro - 2 outubro de 2013, mas apenas 7,135 tinham interações *Foursquare*. Usando a caixa delimitadora, foram identificados 107,413 usuários, com apenas 8% tendo interações *Foursquare*. Como não houve uma superposição entre os usuários coletados com ambas as estratégias, o conjunto de dados final foi composto por 15,435 usuários.

#### 4.3.1.2 Bairros de usuários rotulados

Como mencionado anteriormente, assumimos que a riqueza dos bairros que um usuário visita é um bom indicador de sua classe social. Por isso, antes de rotular os usuários de acordo com uma classe social, devemos rotular os bairros visitados por eles. Para esse fim, foram selecionadas as 27 capitais do Brasil e seus bairros extraído do Censo Brasileiro de 2010, juntamente com a renda mensal média per capita de cada bairro [IBGE, 2010] ou sub-região. Esta informação não estava disponível para quatro cidades, que foram desconsideradas. Das 23 cidades consideradas, São Paulo é a mais populosa, com 12 milhões de habitantes e 96 bairros. A menos populosa é Boa Vista, com 300 mil pessoas e 55 bairros. A cidade com o menor número de bairros, 22, é Salvador. As cidades selecionadas estão distribuídas entre as cinco principais regiões do país.

Após obter o rendimento médio mensal per capita de cada região queremos

---

<sup>11</sup><http://www.academia.org.br/>

Tabela 4.12: FGV - Definição de classe social de acordo com o rendimento mensal familiar

Classe	Renda mensal
A	Acima de R\$9.745,00
B	R\$7.475,00 a R\$9.745,00
C	R\$1.734 a R\$7.475,00
D	R\$1.085,00 a R\$1.734,00
E	Abaixo de R\$1.085,00

associá-lo a uma classe social. Experimentamos três mapeamentos diferentes, com duas (alta/baixa), três (alta/média/baixa) e quatro classes sociais (alta/média-alta/média-baixa/baixa). Os limiares que distinguem diferentes classes foram definidos de acordo com a classificação fornecida pela FGV<sup>12</sup>. Encontramos também duas outras classificações oficiais do governo, fornecidas pela Secretária Assuntos Estratégicos (SAE) e pela Associação Brasileira de Empresas e Pesquisas (ABEP), mas que são constantemente criticadas por não refletir a realidade. Este não é o caso com a divisão de classes da FGV, apresentados na Tabela 4.12.

Os números mostrados na Tabela 4.12 referem-se a renda familiar, ao passo que os números extraídos do censo são per capita. Como também divulgado pelo IBGE as famílias brasileiras têm, em média, 3.2 membros [IBGE, 2010], convertemos esses números, terminando com as três divisões de classes sociais diferentes apresentadas na Tabela 4.13. Com base nestes limites, atribuímos uma classe para cada bairro das cidades selecionadas, seguindo as três estratégias de divisão de classes mostradas anteriormente.

A partir dos 2,134 bairros considerados, verificou-se interações *Foursquare* em 1,415, o que corresponde a 66%. Durante o processamento marcação dos bairros com três classes, 287 destas regiões foram consideradas de classe baixa, 937 de médias e 181 de alta. Quando a divisão foi realizada considerando quatro classes, as regiões média-baixa e média-alta foram divididas em 711 e 236 localidades, respectivamente. Estes números mostram o fato de a classe média ser maior do que as outras duas. Ao usar apenas duas classes, 998 bairros foram considerados de classe baixa e 181 de classe alta.

O processo de rotulação dos bairros foi análogo ao de rotulação de usuários. Para cada usuário, foram mapeadas cada uma de suas interação *Foursquare* para um bairro usando a API do Bing Maps<sup>13</sup>, que converte a latitude e longitude para in-

<sup>12</sup>Fundação Getúlio Vargas - <http://portal.fgv.br/en>

<sup>13</sup><http://www.microsoft.com/maps/>

Tabela 4.13: Número de usuários do Twitter encontrado em cada classe

2 Classes		
	Rendimento	# Usuários
Alta	Acima R\$1.438	9,578
Baixa	Abaixo R\$1.438	5,242
Total		14,820
3 Classes		
	Rendimento	# Usuários
Alta	Acima de R\$ 2335,95	5,299
Média	R\$ 541,88 a R\$ 2335,94	8,295
Baixa	Abaixo de R\$541,87	1,055
Total		14,649
4 Classes		
	Rendimento	# Usuários
Alta	Acima de R\$ 2335,95	5,916
Média Alta	R\$ 1438,91 a R\$ 2335,95	2,786
Média Baixa	R\$ 541,87 a R\$ 1438,91	4,387
Baixa	Abaixo de R\$ 541,87	1,149
Total		14,238

formações como a cidade e o bairro. Se o local envolvido na interação está em um bairro classificado como classe alta, de acordo com dados da Tabela 4.13, a interação foi automaticamente considerada como classe alta. Isto foi feito para todos os lugares que aparecem em interações *Foursquare*. Observe que esse mapeamento é aproximado, pois poderíamos ter estabelecimentos de classe média em áreas de classe baixa ou alta e vice-versa.

Para cada usuário, somou-se o número de lugares que ele visitou em cada uma das classes sociais consideradas, sendo a classe com o maior número de lugares visitados como o rótulo de usuário. Por exemplo, se um usuário tinha 10 interações *Foursquare* sendo 8 classificadas como classe média e outras 2 como classe alta, ele foi rotulado como classe média.

A distribuição de classes obtida está disponível na Tabela 4.13. Observe que o número de usuários rotulados é menor do que o original (15,435). Para três classes, por exemplo, 14,649 usuários foram rotulados. Isso aconteceu porque 786 usuários foram classificados como indefinidos. É atribuída a classe indefinido para um usuário quando há um empate em seu número de visitas a lugares de diferentes classes, por exemplo, 5 interações em classe alta e 5 interações em bairros de classe média. Observe também que quando se utiliza duas classes há mais dados disponíveis sobre a classe alta do que para a classe baixa. Ao utilizar três ou quatro classes, por sua vez, a classe média

Tabela 4.14: Número médio de interações *Foursquare* considerando a classe atribuída

		Classe atribuída			
		Classe Alta	Classe Média	Classe Baixa	Indefinida
Interações	Classe Alta	10	2	0	2
	Classe Média	2	12	0	2
	Classe Baixa	0	0	8	0

concentra a maioria dos usuários, seguido pela classe alta e baixa.

#### 4.3.1.3 O processo de rotulação é eficaz?

É difícil mensurar a eficácia do nosso processo de rotulação, uma vez que as verdadeiras classes sociais dos usuários são desconhecidas, ou seja, não existe um padrão ouro para rotulação. No entanto, ao analisar os usuários marcados com cada classe, encontramos evidências de que a nossa rotulação pode ser razoável.

Começamos por mostrar, na Tabela 4.14, a mediana do número de interações por classe, quando consideramos três classes sociais. Assumimos, durante o processo de rotulação, que os usuários deveriam ir a lugares em regiões de renda similar. Isto é corroborado pelos dados da Tabela 4.14. Mesmo os usuários indefinidos, ou seja, aqueles que apresentaram igualdades no número de interações em lugares pertencentes a diferentes classes, têm sido sempre em lugares de classe mais próxima, por exemplo, alta e média, mas nunca alta e baixa. A média do número de interações é de 8, 10 e 12 para as classes baixa, média e alta, mas este número pode ser tão baixo quanto 1 ou tão alto como 518. Independentemente da classe social, o número médio de interações por usuário é de 29 e a mediana 13.

Tabela 4.15: Média da distância máxima (km) entre quaisquer dois lugares visitados pelos usuários, considerando a classe atribuída

Interação	Baixa	Média	Alta
<i>Check-ins</i>	74.4	160.9	219.8
<i>Tips</i>	375.5	552.6	1,001.5
<i>Mayorships</i>	163.7	254.8	591.4
Total	183.3	296.3	553.9

Outro dado interessante é apresentado na Tabela 4.15. Para cada usuário, calculamos a distância máxima entre quaisquer dois lugares visitados por ele e relatamos as médias de acordo com diferentes tipos de interações *Foursquare*. Como observado, as distâncias máximas podem ser classificadas como mais baixas para classe baixa e

mais altas para a classe alta. Ou seja, os usuários da classe baixa caminham distâncias menores do que os das classes mais altas. Esses números podem refletir viagens, por exemplo, que mostram que as pessoas de classe alta são mais propensas a viajar. Em relação ao número de *majorships*, um comportamento semelhante é observado. Lembre-se que um usuário recebe uma *majorship* se ele tem o maior número de *check-ins* em um local nos últimos 60 dias. Embora os prazos não sejam considerados na abordagem atual, *majorships* em lugares mais distantes podem sugerir que pessoas de classe alta viajam mais do que das classes média ou classe baixa.

Uma das limitações da abordagem de rotulação proposta é que o usuário deve ter contas tanto do *Foursquare* quanto no Twitter. Se em vez de interações *Foursquare* tivéssemos rotulado usuários de acordo com as coordenadas GPS associado a todos os seus tweets disponíveis e atribuído uma classe pela riqueza do bairro mais frequentemente visitado, o resultado seria diferente?

Para responder a essa questão, mapeamos as coordenadas de todos os tweets de todos os usuários para os bairros, de forma similar ao que fizemos para as interações *Foursquare*. Note que os *check-ins* estão presentes nos tweets e, portanto, eles são um subconjunto desse novo conjunto de evidências. Dos 15,435 usuários na base de dados, 11,903 teriam o mesmo rótulo quando classificados com interações *Foursquare* ou com as coordenadas dos tweets, e 3,532 não. Isso representa a concordância de 77%, contando com os usuários indefinidos. Se excluirmos os usuários indefinidos da base de dados, a concordância sobe de 77% para 91,05%. Isto significa que a rotulação com coordenadas tweets diminui o número de usuários indefinidos.

Como explicado anteriormente, não existem provas capazes de demonstrar que uma abordagem é melhor sobre a outra ou de explicar por que usar os tweets diminui desacordo. Uma hipótese é que, em geral, teríamos mais evidências com os tweets que é geralmente maior do que o número de interações *Foursquare* (em média, temos 194 tweets por usuário e 29 interações *Foursquare*). Além disso experimentação com este segundo processo de etiquetagem é deixado para trabalho futuro.

### 4.3.2 Características textuais

Tendo uma base de dados com usuários rotulados queremos identificar características capazes de distinguir pessoas de diferentes classes sociais, um passo para a construção de um método automático de inferência de classe social. Uma fonte potencial de evidência é o texto que os usuários escrevem. É bem conhecido que o texto pode distinguir pessoas de diferentes sexo, idades ou regiões [Nguyen et al., 2013], mas não temos conhecimento de qualquer estudo prévio sobre classes sociais.

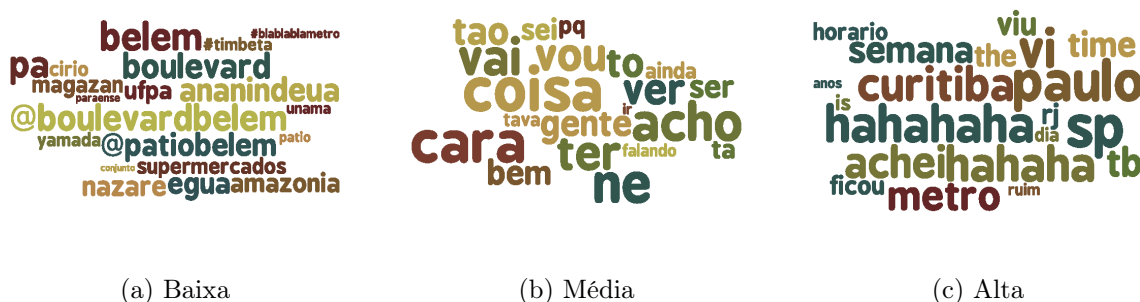


Figura 4.8: Termos discriminativos para três classes sociais

Para isso, analisamos o texto dos 200 tweets mais recentes de cada usuário. Começamos por realizar alguns pré-processamentos simples no texto, como a remoção de termos *stop words* e todos os outros com frequência menor ou igual a 5. Todos os experimentos foram realizados considerando unigramas. Assim, consideramos apenas a presença ou ausência de um termo no conjunto de tweets postados por um usuário e utilizamos a estatística de Ganho de Informação [Witten & Frank, 2005] para classificar esses termos de acordo com seu poder de discriminar os usuários em diferentes classes sociais. Todos os termos com ganho de informação igual a 0 foram removidos a partir do conjunto de dados. Após este processo, obtivemos 21,961, 11,826 e 9,362 termos para as divisões em duas, três e quatro classes, respectivamente. Observe que o maior número de termos para a divisão com menos classes é devido ao maior número de usuários e, conseqüentemente, de tweets (ver Tabela 4.13).

Ao classificar os termos de acordo com o Ganho de Informação para todas as três divisões de classe, observou-se um fato interessante. O termo melhor classificado foi a palavra *egua*, que é uma expressão regional comumente usado no norte e nordeste do Brasil. Descendo a lista de termos, observamos que muitos dos top-termos também foram compostos por expressões regionais. As nuvens de *tags*, ilustradas na Figura 4.8, mostram os termos mais representativos quando considerada uma divisão em três classes. Dado o ranking obtido pelo Ganho de Informação  $R$  e a base de dados dos usuários  $D$ , descrito pela presença ou ausência de termos, foram associados os termos em  $R$  com a classe mais frequente em  $D$ . Os 20 termos mais relevantes classificados de acordo com este processo são listados nas nuvens de *tags*.

Na Figura 4.8 observamos nomes de cidades e lugares do norte do país, como paraense (pessoa nascida no Pará), Belém (capital do estado do Pará), Amazônia e égua, a expressão regional. O vocabulário da classe média (Figura 4.8 (b)) traz *face*, abreviação para o Facebook. Para a classe alta (Figura 4.8(c)), há uma forte presença



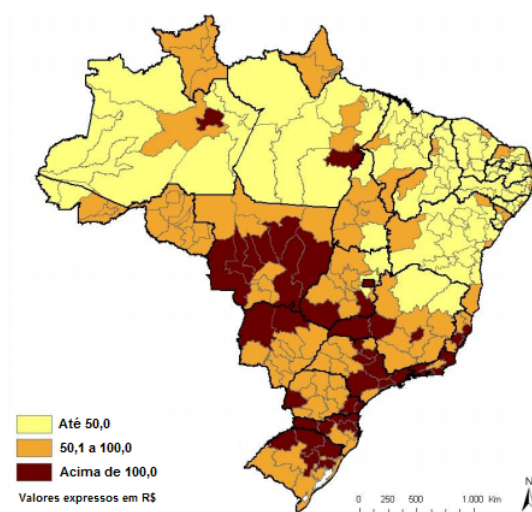


Figura 4.9: Distribuição do Produto Interno Bruto brasileiro (PIB) per capita em diferentes regiões do país

de *hahaha* e novamente o nome de cidades do sul e sudeste, como Curitiba, São Paulo (sp) e Rio de Janeiro (rj).

Do ponto de vista econômico, os estados do sul e sudeste do Brasil são conhecidos por serem mais ricos do que os do Norte e Nordeste. Estas diferenças regionais de renda fazem o regionalismo muito relevante ao distinguir as classes. Assim, analisamos a distribuição das classes sociais dos usuários para cada região separadamente. Mostramos uma análise em profundidade da relação entre as regiões e classes sociais, considerando o conjunto de dados com três classes, um comportamento semelhante é encontrado para duas e quatro classes.

A Tabela 4.16 apresenta a distribuição de classe na base de dados por região, em que a linha denominada “Diverso” corresponde aos usuários que tiveram interações *Foursquare* em locais de diferentes regiões, mas sempre na mesma classe social. Comparando a distribuição de classe em diferentes regiões com a distribuição global do conjunto de dados (linha rotulada como total), observa-se que o número de usuários nas classes baixa e média no norte são muito próximos (573 e 635, respectivamente), fato que não acontece em nenhuma outra região. Cerca de metade dos usuários de classe baixa em nosso conjunto de dados são do norte do Brasil, e junto com os usuários do nordeste, correspondem a 85% das pessoas de baixa classe na base de dados. Em contrapartida, no sudeste, o número de pessoas nas classes média e alta é muito similar, acima de 2,000, enquanto que apenas 36 usuários foram classificados como classe baixa.

O que alguns poderiam ver como um viés no processo de coleta, na verdade, reflete muito bem a realidade do país, onde pessoas do sudeste têm uma renda muito

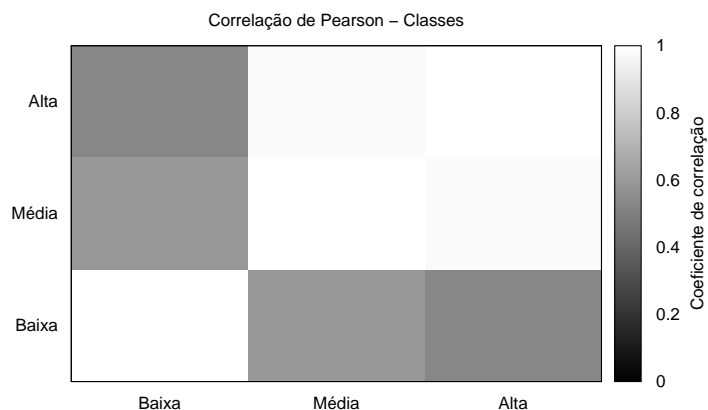


Figura 4.10: Correlação de Pearson do vocabulário das classes utilizando 100 termos

maior do que as do norte e nordeste. A Figura 4.9 mostra a distribuição do Produto Interno Bruto (PIB) do Brasil em todas as regiões. Quanto mais escura a região mais rica ela é. Observe que o norte e nordeste são em sua maioria de cor clara, enquanto o sul, sudeste e centro-oeste têm a maioria das regiões mais escuras, com a riqueza do centro-oeste vindo principalmente de atividades rurais.

Tabela 4.16: Distribuição dos usuários em classes sociais conforme sua região

Região	Classe			Total
	Baixa	Média	Alta	
Norte	573	635	45	1,253
Nordeste	322	2,293	511	3,126
Centro-oeste	1	280	184	465
Sul	25	1,269	854	2,148
Sudeste	36	2,373	2,254	4,663
Diverso	98	1,445	1,451	2,994
Total	1,055	8,295	5,299	14,649

Voltando aos vocabulários, calculamos a correlação de Pearson para os 100 principais termos obtidos pelo Ganho de Informação, como descrito para as nuvens de *tags*. A Figura 4.10 mostra a correlação entre os vocabulários, onde regiões mais escuras representam correlações mais baixas. Observe que há uma forte correlação entre os vocabulários das classes média e alta e, como desejado, a classe mais baixa tem fraca correlação com a classe alta.

Até agora, exploramos e identificamos termos relevantes e discriminativos em

mensagens postadas pelos usuários. Desta forma, com base no conhecimento dos sociólogos e do que eles dizem serem bons indícios para distinguir classes sociais, definimos quatro conjuntos de indicadores, são eles: (i) lazer, (ii) consumo, (iii) educação, e (iv) transporte. Para cada um desses indicadores, um conjunto de termos foi manualmente definido, conforme listado abaixo:

- Lazer (73 termos): lazer, passeio, cinema, filme, clube, praia, rio, sitio, cachoeira, museu, estadio, arena, boate, beile, danca, dançar, balada, teatro, show, shows, bar, barzinho, boteco, pub, musical, viagem, viajar, viajei, turismo, zoologico, parque, shopping, trilha, piscina, churrasco, esporte, jogo, jogar, aeroclube, hipismo, esgrima, golfe, tenis, futebol, volei, basquete, mma, ufc, natacao, nadar, judo, lutar, ciclismo, pedalar, bale, ballet, caminhada, corrida, treino, treinar, time, boliche, novela, jornal, revista, livro, literatura, serie, seriado, desenho, fotografia, jardinagem e arte.
- Consumo (116 termos): roupa, vuitton, cartier, chopard, tiffany, pucci, chloe, diesel, colcci, daslu, lacoste, cea, riachuelo, renner, pernambucanas, chanel, polo, jaguar, bmw, mercedes, audi, porsche, toyota, kia, fiat, vw, volkswagen, hyundai, chevrolet, ford, honda, kawasaki, ducati, kasinski, suzuki, yamaha, sephora, avon, natura, jequiti, boticario, maquiagem, perfume, cosmetico, creme, locao, bronzeador, shampoo, xampu, condicionador, hidratante, esmalte, esfoliante, estetica, cabeleireiro, esteticista, academia, massagem, lingerie, paisagismo, celular, notebook, tablet, tv, televisao, gps, mp3, iphone, ipad, adega, convenio, casa, apartamento, condominio, favela, comunidade, morro, mansao, cobertura, apart, barracao, hotel, kichenette, loft, predio, aluguel, sushi, salmon, caviar, lagosta, escargot, pizza, lasanha, estrogonofe, sanduiche, vinho, cerveja, cachaca, refrigerante, espumante, champagne, whisky, vodka, rodizio, churrascaria, pizzaria, soverteria, hamburgueria, lanchonete, restaurante, buffet, bistro, steakhouse, panificadora, padaria e supermercado.
- Educação (14 termos): prova, materia, disciplina, enem, vestibular, intercambio, escola, estudei, faculdade, universidade, estudar, concurso, fies e educacional.
- Transporte: (14 termos): carro, moto, aviao, helicoptero, trem, metro, onibus, brt, taxi, lotacao, aeroporto, bicicleta, transito e gasolina.

As categorias lazer e consumo podem, no futuro, ser divididas em outras subcategorias, como música, esporte, etc, mas, neste momento, consideramos conforme descrito. Dos 14,649 usuários utilizados para classificação em 3 classes, 99.9% têm pelo

Tabela 4.17: Principais termos para lazer e consumo conforme medida de Ganho de Informação

Lazer	Consumo
time	<i>steakhouse</i>
rio	hotel
bar	convenio
<i>pub</i>	panificadora
balada	supermercado
praia	bistro
teatro	ipad
tenis	iphone
museu	sushi
<i>show</i>	cerveja
estadio	restaurante

menos uma menção das palavras que definem a categoria lazer, 88% para consumo, 65% para educação e 70% para transporte. Assim, podemos concluir que as pessoas falam sobre assuntos relacionados com a classe social.

A próxima pergunta é se esses termos podem discriminar classes. Assim, foi calculado o Ganho de Informação para cada uma das categorias criadas manualmente. Mostramos os resultados obtidos pelas duas categorias mais promissoras: lazer e consumo. Os termos obtidos no topo do ranking são listados na Tabela 4.17. Para consumo, apenas 20% das palavras tinham valores de ganho de informação maior que 0. Para o lazer, este número aumentou para 50%. Na sociedade brasileira palavras como bistrô e sushi estão normalmente relacionados com a classe alta. Já futebol, bar e cerveja podem ser consideradas universais, mas mesmo assim eles foram indicados como tendo um alto poder de discriminação.

### 4.3.3 Classificação

Depois de analisar os recursos textuais que podem distinguir os usuários de diferentes classes, esta seção utiliza esses recursos para criar modelos de classificação capazes de discriminar os usuários de diferentes classes. Os experimentos foram realizados considerando os usuários divididos em 2, 3 e 4 classes, conforme o número de usuários descritos na Tabela 4.13. Diferentes combinações dos recursos foram utilizadas para descrever os dados, incluindo atributos textuais (21,961, 11,826 e 9,362 termos para as classes 2, 3 e 4), categorias textuais isoladas (lazer, consumo, educação e transporte - veja a seção anterior para mais detalhes) e a possível combinação entre as categorias.

Os parâmetros para o SVM foram otimizados usando o ferramenta *easy*, que

Tabela 4.18: Resultados obtidos para 2, 3 e 4 classes sociais, utilizando os atributos textuais na base de dados balanceada e original

	Base de dados original					
	2 Classes		3 Classes		4 Classes	
	$f_1$	Acurácia (%)	$f_1$	Acurácia (%)	$f_1$	Acurácia (%)
<i>Naive Bayes Multinomial</i>	0.74	73.74	0.69	68.88	0.57	57.09
SVM	0.75	76.18	0.69	69.48	0.51	57.26
<i>Random Forest</i>	0.65	65.65	0.54	56.41	0.43	45.11
	Base de dados balanceada					
<i>Naive Bayes Multinomial</i>	0.76	76.11	0.73	73.08	0.59	58.59
SVM	0.73	73.40	0.71	70.93	0.53	53.48
<i>Random Forest</i>	0.60	61.17	0.53	53.49	0.39	39.52

Tabela 4.19: Resultados para a classificação de classe social usando NBM em duas regiões

		2 Classes		3 Classes		4 Classes	
		$f_1$	Acurácia (%)	$f_1$	Acurácia (%)	$f_1$	Acurácia (%)
Nordeste	Original	0.82	82.36	0.73	72.62	0.60	60.11
	Balanceada	0.82	82.50	0.73	72.88	0.59	59.61
Sudeste	Original	0.81	80.10	0.89	89.74	0.69	69.81
	Balanceada	0.87	87.55	0.93	93.52	0.81	80.92

realiza uma pesquisa de grade na escolha dos valores. Os valores encontrados para  $c$  e  $\gamma$  foram de 2.0 e 0.000488. Todos os experimentos foram realizados utilizando um procedimento de validação cruzada de cinco partições.

Como a classe baixa tem muito menos exemplos do que as demais os resultados dos experimentos são sempre relatados de duas formas, sendo a primeira com a base de dados original e a segunda com uma versão balanceada, gerada de acordo com o número de exemplos na classe mais baixa.

Os primeiros resultados, utilizando todos os atributos textuais (após o pré-processamento descrito na seção anterior) para descrever os usuários, com precisão e  $f_1$  (média harmônica de precisão e recuperação) são apresentados na Tabela 4.18. Tal como esperado, quanto mais baixo o número de classes, melhor o desempenho da classificação. Com duas classes, obtemos um  $f_1$  de 0.75. Este valor diminui para 0.69 quando três classes são consideradas e 0.57 para quatro classes. Observe que entre os classificadores NBM e SVM a diferença de resultados nas bases de dados original e balanceada não é significativa, sendo NBM com três classes uma exceção.

Para ambas as versões da base de dados, não há nenhuma significância estatística nos valores de  $f_1$  obtidos por NBM e SVM e, em todos os casos, os resultados do

*Random Forest* são estatisticamente piores do que os de NBM e SVM. Como os resultados de eficácia são comparáveis e o NBM é mais eficiente, os seguintes experimentos relatados nesta dissertação usam apenas o NBM para gerar os modelos de classificação.

Ao analisar os dados dos textos para distinguir as classe sociais nós observamos uma alta influência de termos regionais na classificação. Como as regiões geográficas do Brasil tem muitas outras diferenças, além de renda, foi realizado um experimento que dividiu o conjunto de dados original em partes menores, cada uma delas contendo os usuários de uma região específica. Isso resultou em conjuntos de dados muito desequilibrado. Relatamos os resultados para duas regiões que tiveram amostras suficientes: Nordeste e Sudeste. Os resultados são apresentados na Tabela 4.19.

Observe que, nesta abordagem, os resultados de  $f1$  para o nordeste são 0.82, 0.73 e 0.6 para 2, 3 e 4 classes em cenários desbalanceados, o que é muito semelhante aos resultados obtidos no cenário balanceado. Para o sudeste os resultados de precisão e  $f1$  são surpreendentemente altos. Ao classificar os usuários em três classes obteve-se os melhores valores de  $f1$  (0.89), em seguida, ao usar duas classes (0.81). Utilizando quatro classes os valores diminuem, com um  $f1$  de 0.69.

Considerando três divisões de classes, o conjunto de dados balanceado obtém valores muito elevados de  $f1$ , atingindo 0.93. A distribuição de classe para a região sudeste pode ser encontrada na Tabela 4.16. Como a região tem apenas 36 usuários na classe baixa, a base de dados balanceada com 124 usuários tornou mais fácil a tarefa para o classificador.

Em um segundo conjunto de experimentos, foram utilizadas as categorias previamente definidas de termos para classificar os usuários, como mostrado na Tabela 4.20. Inicialmente, cada categoria foi utilizada isoladamente e, em seguida, combinadas em um único conjunto de dados (L + C + E + T). Observe que estes conjuntos de dados tem um número muito menor de atributos. Lazer é definido por 73 termos, Consumo por 116 e Educação e Transporte por 14 cada. Assim, L + C + E + T tem 217 atributos. A base de dados com menor dimensionalidade, ao utilizar todos os atributos, foi obtida na divisão de quatro classes, com 9,362 termos. Assim, a maior base de dados nos experimentos com categorias de texto utiliza 2% dos atributos da base de dados original.

Observando os resultados com dois valores de classes sociais, obtivemos  $f1$  de 0.65, enquanto o  $f1$  obtido com todos os atributos foi de 0.74. Perdemos 0.1 em  $f1$ , mas reduziu o conjunto de dados em 98%. Note que, quando isoladas, as categorias de consumo e lazer são as mais promissoras. Movendo para as divisões de três e quatro classe, o  $f1$  diminui de 0.69 e 0.57 quando se utiliza todos os termos a 0.55 e 0.41, respectivamente, quando se utiliza o conjunto combinado de categorias pré-definidas.

Tabela 4.20: Resultados obtidos para 2, 3 e 4 classes sociais utilizando diferentes atributos

	Base de dados original					
	2 Classes		3 Classes		4 Classes	
	$f1$	Acurácia (%)	$f1$	Acurácia (%)	$f1$	Acurácia (%)
Todos	0.74	73.74	0.69	68.88	0.57	57.09
Lazer	0.54	64.63	0.49	53.28	0.33	42.03
Consumo	0.59	65.33	0.51	54.54	0.36	43.69
Educação	0.51	64.6	0.40	49.81	0.25	41.52
Transporte	0.51	64.61	0.47	52.6	0.28	41.41
L+C+E+T	0.65	66.1	0.55	57.3	0.41	45.85
	Base de dados balanceada					
Todos	0.76	76.11	0.73	73.08	0.59	58.59
Lazer	0.56	56.36	0.43	43.73	0.32	32.74
Consumo	0.57	57.55	0.47	47.68	0.31	32.84
Educação	0.49	51.35	0.39	41.57	0.26	29.80
Transporte	0.54	55.52	0.39	40.82	0.28	31.13
L+C+E+T	0.62	61.79	0.53	52.64	0.37	38.17

São também apresentados os resultados para uma versão balanceada do conjunto de dados, mas eles são estatisticamente piores do que aqueles obtidos com os conjuntos de dados originais.

Estes resultados mostram que a definição de categorias de acordo com os hábitos dos usuários de forma bem estudada é muito útil para discriminar as classes sociais, mesmo no ambiente *online*. Novos conjuntos de palavras cuidadosamente criadas podem melhorar os resultados obtidos até agora, ou pelo menos reduzir drasticamente o número de termos usados na primeira tentativa de caracterizar a classe social apresentada neste trabalho.





## Capítulo 5

# Avaliação do arcabouço

A última etapa do arcabouço proposto consiste na formação de amostras a partir do conjunto total de dados coletados, considerando as características pessoais e comportamentais dos usuários, aliado a caracterização do sentimento expresso nas mensagens.

Para validarmos nossos resultados, iremos avaliar estatisticamente as amostras obtidas e comparar os acertos encontrados diante dos resultados reais e de métodos tradicionalmente utilizados em pesquisas de opinião, como enquetes publicadas por grandes sítios da Web e resultados divulgados por institutos de pesquisas renomados.

A contabilidade dos votos atribuídos a cada entidade analisada (candidato ou participante do programa) será computada conforme as equações 5.1, para votos de eliminação, e 5.2, para votos de aprovação. Observe que em ambos os cenários levamos em consideração tanto votos positivos quanto os negativos, mas dependendo do objetivo um deles é distribuído igualmente entre as demais entidades não citadas. Por exemplo, se estamos diante de um cenário de eleição política e um usuário falou mal de apenas um candidato durante o período analisado seu voto é distribuído igualmente entre os candidatos não citados. Esta estratégia foi utilizada para aumentar a quantidade de votos computados e, conseqüentemente, permitir a formação de mais amostras.

$$Votos\ candidato_i = \frac{negativo_i + \sum_{j=1, i \neq j}^N \left(\frac{positivo_j}{N-1}\right)}{\sum_{j=1}^N positivo_j + \sum_{j=1}^N negativo_j} \quad (5.1)$$

$$Votos\ candidato_i = \frac{positivo_i + \sum_{j=1, i \neq j}^N \left(\frac{negativo_j}{N-1}\right)}{\sum_{j=1}^N positivo_j + \sum_{j=1}^N negativo_j} \quad (5.2)$$

Além disso, o conjunto de votos será dado pela média obtida após a formação de 100 amostras no conjunto de usuários válidos encontrados em cada contexto. Cada amostra será composta pelo maior número de usuários possíveis, conforme Tabela 3.9,

até uma margem de erro máxima de 15 pontos percentuais. Apesar do limite da margem de erro ser grande, se comparada às obtidas pelos institutos tradicionais de pesquisa (normalmente entre 2 e 4 pontos percentuais), é um valor necessário para obtenção de amostras em cenários onde caracterizamos os usuários em um maior nível de detalhes.

Os resultados serão comparados pelo acerto do candidato vencedor ( $p@1$ ) e, no caso das eleições, também para o acerto dos dois primeiros colocados, que disputarão o segundo turno.

Serão apresentados resultados considerando dois tipos de contabilização de votos: “menção” dos usuários, onde serão permitidos votos múltiplos oriundos de um único indivíduo e “usuários únicos”, em que será computado apenas um voto para cada usuário, que se refere a apenas um candidato durante o período analisado. Um usuário é equivalente a uma conta/perfil no Twitter.

A subdivisão de usuários em cada amostra, conforme caracterização pessoal, será feita considerando as dimensões analisadas como eventos independentes. Portanto, ela será dada pela interseção dos conjuntos, com a multiplicação das probabilidades de ocorrência de cada grupo. Por exemplo, se queremos uma amostra formada por 60% masculina e 10% com idade entre 20 e 30 anos resultará em 6% de usuários masculinos com idade entre 20 e 30 anos.

Especificidades sobre cada contexto avaliado e os resultados obtidos são mostradas nas subseções a seguir, que considerarão dois eventos: BBB 13 e Eleições 2012.

## 5.1 *Big Brother* Brasil (BBB)

O *Big Brother* Brasil (BBB) é um *reality show* apresentado na televisão brasileira. Neste programa, os participantes são confinados em uma casa, sem contato com o mundo externo, e submetidos a votações públicas que determinam sua eliminação, denominados “paredões”, ou a vitória no jogo, na etapa final.

Os meios para votações disponibilizados para os telespectadores do programa são ligações telefônicas, mensagens de texto curtas do tipo SMS (*Short Message Service*) e o sítio oficial do programa<sup>1</sup>. Não existe a premissa de voto único, ou seja, é permitido que uma mesma pessoa emita sua opinião inúmeras vezes.

Desta forma, os votos para “paredões” serão computados conforme a equação 5.1 e para a final conforme a equação 5.2. Não haverá distinção do tipo da etapa de votação na apresentação dos resultados, ou seja, o número de acertos e erros será a acumulada obtida durante as duas etapas do programa.

---

<sup>1</sup><http://www.globo.com/bbb>

Para avaliarmos o potencial de previsão dos tweets foram coletadas 16,537,070 mensagens, publicadas por 2,909,861 usuários distintos, entre os dias 28 de dezembro de 2012 e 12 de junho de 2013. Dentre os usuários identificados no período considerado, em média cada um postou 5.68 tweets sobre o assunto, com uma mediana igual a 1. O usuário mais frequente postou 11,548 mensagens e o menos frequente 1.

O perfil dos telespectadores do programa, utilizados para formação de amostra nas etapas de caracterização pessoal de sexo, idade e classe social, foram obtidos em reportagem publicada pelo jornal Folha de São Paulo<sup>2</sup> e ilustra que:

- Sexo: feminino (61%) e masculino (39%);
- Idade: 4 a 11 anos (8%), 12 a 17 anos (10%), 18 a 24 anos (12%), 25 a 34 anos (19%), 35 a 49 anos (24%) e maior de 50 (27%).
- Classe social: classes DE (12%), classe C (53%) e classes AB (35%).

Em nosso trabalho, utilizaremos três divisões de classe: DE como baixa, C como média e AB como alta. As subdivisões nas faixas etárias também considerarão três faixas: menor que 25 anos, 25 a 49 anos e acima de 49 anos.

Restringimos a caracterização dos usuários a somente aqueles identificados no período de cada votação. Assim, foi possível coletar os últimos 200 tweets de 530,953 contas. Após os processos de caracterização encontramos o seguinte perfil para os usuários que postaram mensagens sobre o BBB:

- Sexo: feminino (63.42%) e masculino (36.58%);
- Idade: até 25 (84.60%), 25 a 45 (13.43%) e maior de 45 (1.96%).
- Classe social: baixa (6.08%), média (73.70%) e alta (20.22%).

A distribuição de sexo encontrada foi bem próxima do perfil dos telespectadores. No entanto, características como a faixa etária e a classe social dos indivíduos foram bem diferentes, refletindo uma característica normalmente encontrada no Twitter de usuários mais jovens e com classe social concentrada entre média e alta, como já observado nas seções 4.2 e 4.3.

---

<sup>2</sup><http://www1.folha.uol.com.br/fsp/ilustrada/140076-bbb-tem-plateia-feminina-mais-velha-e-da-classe-c.shtml>

Tabela 5.1: Resultados BBB - p@1 (Sen: análise de sentimento; SJ: remoção de usuários *spammers* e jornalísticos; Sex: sexo; ID: idade; CS: classe social)

Característica					Acerto / Votações com amostras	
Conjunto completo de mensagens						
Sen	SJ	Sex	ID	CS	Menção	Usuário Único
					5/14	-
X					9/14	-
Amostras						
					5/14	4/14
	X				4/14	4/14
X					9/14	9/14
		X			5/14	4/14
			X		4/14	3/14
				X	4/14	4/14
X	X				9/14	10/14
	X	X			4/14	4/14
	X		X		4/14	3/14
	X			X	4/14	4/14
X		X			9/14	11/14
X			X		9/14	10/14
X				X	11/14	11/14
		X	X		4/14	3/14
		X		X	4/14	4/14
			X	X	1/4	0/2
X	X	X			9/14	10/14
X	X		X		9/14	10/14
X	X			X	11/14	11/14
X		X	X		10/14	10/14
X		X		X	9/14	10/14
		X	X	X	1/4	0/1
	X	X	X		4/14	3/14
	X	X		X	4/14	4/14
	X		X	X	1/4	0/2
X			X	X	2/3	1/1
X	X	X	X		9/14	11/14
X	X	X		X	9/14	10/14
X		X	X	X	2/2	1/1
	X	X	X	X	1/3	0/1
X	X		X	X	1/2	1/1
X	X	X	X	X	1/1	1/1
Institutos Tradicionais						
Enquete UOL					10/14	-

Os resultados são apresentados na Tabela 5.1, para 64 cenários divididos em 32 onde a menção é avaliada e 32 para quando utilizamos a estratégia de identificar usuários únicos. O período de análise de cada votação, seja ela paredão ou final, foi delimitado pelos dias em que a votação estava aberta para os telespectadores, normalmente 2 dias.

Como ponto de comparação serão utilizados os resultados de enquetes realizadas pelo portal Uol<sup>3</sup>, também mostrados na Tabela 5.1. Podemos observar que nesses resultados houve uma taxa de acerto de aproximadamente 71%. Os erros obtidos nestas enquetes, em comparação com o resultado real, aconteceram nos “paredões” 2, 4, 6, 8 do programa e, entre eles, três foram motivados por empates entre os candidatos com maior quantidade de votos da pesquisa e o outro por prever a eliminação da candidata Kamilla, o que acabou não acontecendo.

Como podemos observar, os melhores resultados encontrados, 11 acertos, foram superiores ao obtido pela enquete do UOL, 10 acertos, e consideravelmente melhor que a análise de menção simples, 5 acertos. Os três erros cometidos, normalmente, ocorreram na quinta, sexta e oitava votações, em que os métodos previram a eliminação dos candidatos Elieser ou Kamilla, o que acabou não acontecendo. De forma geral, observamos que estes dois candidatos mantiveram durante todo o programa um alto nível de rejeição, o que pode ter causado nosso erro.

Dentre as etapas realizadas a análise de sentimentos foi aquela capaz de melhorar mais significativamente os resultados. Porém, uma melhoria na técnica, de forma a identificar com maior confiança mensagens negativas, poderá proporcionar melhores resultados ao arcabouço.

A abordagem de identificar usuários únicos foi relevante principalmente nos casos em que a análise de sentimento aliada a pelo menos mais uma caracterização estava presente no método, o que foi capaz de proporcionar um acerto de até 2 votações a mais.

Considerando somente os cenários onde menção e usuários únicos conseguiram formar as mesmas quantidades de amostras verificamos que o seguinte padrão no método vencedor: nenhuma caracterização (menção 1/usuários únicos 0), uma caracterização (menção 2/usuários únicos 0), duas caracterizacoes (menção 2/usuários únicos 3), três caracterizações (menção 1/usuários únicos 3), quatro caracterizações (menção 0/usuários únicos 4). Com isso, podemos concluir que, o processo de caracterização dos usuários não influencia a acurácia do método de menção, mas é determinante para proporcionar bons resultados quando avaliamos usuários únicos, fazendo com que esta

---

<sup>3</sup><http://www.uol.com.br/>

técnica se sobressaia da concorrente.

Não foi possível formar amostras com margem de erro máxima de 15 pontos em diversas votações, dificuldade que aumentou a cada novo nível de caracterização inserido no método. Assim, apesar da proporção de acerto ser maior quando usamos todas as caracterizações, 100%, não podemos afirmar que ela foi realmente decisiva no processo.

A combinação de características no método que mais atrapalhou a formação de amostras foi classe social aliada a idade dos indivíduos. Observando as distribuições encontradas no perfil real dos telespectadores em comparação ao perfil dos usuários coletados no Twitter já poderíamos prever esta dificuldade, por exemplo, nossas amostras deveriam ser formadas por 70% de usuários acima de 25 anos enquanto a distribuição na distribuição encontrada havia aproximadamente 15%.

## 5.2 Eleições Municipais 2012

Nesta etapa, as eleições municipais ocorridas no ano de 2012 foram avaliadas de acordo com a intenção de votos dos eleitores. Para avaliarmos o potencial de previsão do Twitter foram coletados 1,295,418 tweets, publicadas por 222,412 usuários distintos, entre os dias 18 de julho de 2012 e 17 de outubro de 2012. Dentre os usuários identificados no período considerado, em média cada um postou 5.8 tweets sobre o assunto, com uma mediana de 1. O usuário mais frequente postou 5,072 mensagens e o menos frequente 1.

Assim, foram escolhidas 6 capitais de estado para serem analisadas, com base no total de tweets coletados durante o período de análise e na importância econômica das regiões, são elas: Belo Horizonte (MG), Curitiba (PR), Porto Alegre (RS), Rio de Janeiro (RJ), Salvador (BA) e São Paulo (SP). A partir deste momento representaremos os nomes das cidades pelas siglas de seus respectivos estados. Nossa análise irá focar na previsão do resultado para o acerto do primeiro lugar (p@1), vencedor das eleições, e para os dois primeiros lugares, candidatos que disputam o segundo turno.

Como conhecido, os votos eleitorais para prefeitos são computados unicamente, ou seja, só é permitido que um eleitor escolha um candidato. O perfil dos eleitores, mostrado na Tabela 5.2, utilizados para formação de nossa amostras, foram obtidos diretamente do sítio do Tribunal Superior Eleitoral (TSE<sup>4</sup>), para o sexo e a idade, e de publicações da Fundação Getúlio Vargas (FGV<sup>5</sup>), para a classe social. Neste trabalho, os valores desconhecidos para o sexo serão distribuídos igualmente entre masculino e

---

<sup>4</sup><http://www.tse.jus.br/>

<sup>5</sup><http://portal.fgv.br/>

feminino, as faixas etárias serão novamente agrupadas em três: menor de 25, 25 a 44 e maior de 45 e a classe social em baixa (DE), média (C) e alta (AB).

Tabela 5.2: Perfil do eleitorado em 2012

Característica		MG	PR	RS	RJ	BA	SP
Sexo (%)	Masc.	46.034	45.826	45.375	45.327	45.724	46.169
	Fem.	53.891	54.174	54.625	54.531	54.2	53.626
	Desconhecido	0.075	0	0	0.14	0.076	0.205
Idade (%)	16	0.223	0.219	0.210	0.152	0.250	0.248
	17	0.479	0.745	0.435	0.400	0.637	0.579
	18 a 20	5.113	5.815	4.497	4.792	5.386	5.293
	21 a 24	8.081	8.715	7.498	7.391	8.299	8.196
	25 a 34	23.263	23.333	21.502	20.359	26.822	23.095
	35 a 44	19.687	20.754	17.855	18.231	20.979	20.286
	45 a 59	25.070	26.557	26.419	25.939	23.526	24.887
	60 a 69	9.562	9.891	11.386	11.047	7.861	9.357
	70 a 79	5.276	3.286	6.323	6.425	3.761	4.890
Maior de 79	3.246	0.685	3.875	5.265	2.480	3.16	
Classe (%)	E	14.27	3.64	14.41	19.09	21.99	13.41
	D	12.23	9.6	9.91	11.87	14.05	10.3
	C	52.54	58.2	49	48.84	48.29	54.63
	AB	20.97	28.56	26.68	20.2	15.67	21.66

O período de análise das votações foi considerado como a semana que antecedeu o pleito, ou seja, entre os dias 30 de setembro de 2012 e 6 de outubro de 2012, mas para evitar possíveis flutuações anormais causadas, por exemplo, por repercussão dos últimos debates eleitorais, consideramos as últimas 10 semanas com peso linearmente crescente para os votos computados, ou seja, a semana 1 com peso 1, a semana 2 com peso 2, sucessivamente.

Antes de definirmos este período de uma semana avaliamos diversos outros cenários, dentre eles: 1 ou 2 dias, onde pouquíssimas amostras foram formadas, e o período completo de coleta, onde acontecimentos momentâneos como, por exemplo, debates eleitorais, interferiam significativamente nos resultados alcançados.

Após coletarmos os últimos tweets dos usuários que postaram mensagens para as cidades avaliadas, realizamos os processos de caracterização discutidos anteriormente. O perfil dos usuários que postaram mensagens sobre cada cidade é mostrado na Tabela 5.3.

As características identificadas foram bem diferentes das encontradas para a base do BBB, com uma participação significativamente maior de homens, pessoas mais velhas e de classes sociais mais elevadas. Comparando com o perfil real dos eleitores

também observamos grandes diferenças, principalmente no sexo e classe social dos indivíduos.

Tabela 5.3: Perfil dos usuários do Twitter - Eleições 2012

Característica		MG	PR	RS	RJ	BA	SP
Sexo (%)	Masc.	75,31	73,57	75,90	69,08	77,61	74,58
	Fem.	24,69	26,42	24,10	30,92	22,39	25,41
Idade (%)	Até 25	28,19	42,37	27,77	48,80	16,50	35,28
	25 a 45	29,23	30,08	40,87	31,96	36,90	41,21
	Maior de 45	42,58	27,54	31,36	19,23	46,60	23,50
Classe (%)	Baixa	2,03	2,24	2,43	2,58	11,62	3,02
	Média	25,31	35,39	32,79	40,94	18,31	27,73
	Alta	72,66	62,36	64,78	56,48	70,07	69,24

Os resultados obtidos são apresentados na Tabela 5.4, para a avaliação do vencedor, e na Tabela 5.5, quando a intenção é descobrir os dois participantes que irão concorrer no segundo turno. Para cada avaliação foram criados 64 cenários, divididos em 32 onde analisamos menções e 32 para análise de usuários únicos. Quadrados de cor verde representam acertos, pretos a não formação de amostra e, conseqüentemente, brancos os erros.

Como ponto de comparação serão utilizados resultados de pesquisas de opinião realizadas pelos renomados institutos: Instituto Brasileiro de Opinião e Estatística (IBOPE<sup>6</sup>) e DataFolha<sup>7</sup>, também mostrados na Tabela 5.4, para o acerto do vencedor, e na Tabela 5.5, para o acerto dos dois primeiros colocados. Podemos observar que dentre as 6 cidades escolhidas para análise o Data Folha realizou pesquisas em somente 5 delas.

A cidade mais crítica para o IBOPE foi a capital de SP, em que não foi possível acertar nem o primeiro nem os dois primeiros colocados, pois até o terceiro lugar foi divulgado empate entre os candidatos. Outras cidades em que a pesquisa também falhou foram BA, para o primeiro colocado, e PR, para o segundo turno. O DataFolha acertou todos os vencedores do primeiro turno, nas cidades analisadas. Para o segundo turno errou em SP e PR.

Comparando nossos resultados com os cenários mais complicados para os institutos tradicionais só foi possível a correta previsão para o vencedor de SP quando utilizada análise de sentimentos das mensagens, mas para os dois primeiros colocados nenhuma estratégia acertou. A previsão para BA superou o IBOPE, lembrando que o

<sup>6</sup><http://www.ibope.com.br/>

<sup>7</sup><http://datafolha.folha.uol.com.br/>



DataFolha não avaliou esta cidade, sendo correta em praticamente todos os cenários avaliados, desde a simples menção aos níveis mais detalhados de caracterização.

Se por um lado a estratégia do usuário único foi determinante para encontrar o vencedor do PR ela também prejudicou a correta avaliação para a capital de RS, tanto para o vencedor quanto para os dois candidatos escolhidos para o segundo turno.

Considerando somente os cenários onde menção e usuários únicos conseguiram formar as mesmas quantidades de amostras verificamos que o seguinte padrão no método vencedor: i) p@1: nenhuma caracterização (menção 0/usuários únicos 1), uma caracterização (menção 1/usuários únicos 2), duas caracterizacoes (menção 1/usuários únicos 4), três caracterizações (menção 0/usuários únicos 2); e ii) segundo turno: nenhuma caracterização (menção 1/usuários únicos 0), uma caracterização (menção 2/usuários únicos 2), duas caracterizacoes (menção 1/usuários únicos 6), três caracterizações (menção 2/usuários únicos 4), quatro caracterizações (menção 0/usuários únicos 1). Com isso, podemos concluir que, a identificação de usuários únicos melhora significativamente os acertos e que esta técnica é capaz de alcançar melhores resultados a medida que o nível de caracterização é elevado.

Muitas amostras não puderam ser formadas, com margem de erro máxima de 15 pontos percentuais, mas observamos que nos poucos casos com maiores níveis de caracterização houve melhora de resultado, por exemplo, para encontrar o vencedor do RJ. Assim, de forma geral, podemos avaliar que o arcabouço apresentado é promissor em pesquisas eleitorais, sendo no mínimo tão bom quanto a verificação de menções e superando em alguns casos a técnica concorrente.

### 5.3 Considerações finais

Após avaliarmos o arcabouço proposto sob diversas perspectivas e distintos âmbitos de aplicações, BBB e Eleições, podemos fazer as seguintes considerações:

1. Existe uma grande diferença no perfil do público em cada evento analisado com destaque para o sexo masculino e indivíduos mais velhos para as Eleições;
2. O total de votos computado para cada candidato, formado pela média obtida após a formação de 100 amostras distintas, é capaz de refletir um comportamento análogo ao observado com a análise do conjunto completo de mensagens.
3. A análise de sentimentos foi a caracterização com maior impacto positivo sobre os resultados obtidos;

4. Quanto maior o nível de caracterização mais relevante é identificar usuários únicos;
5. A técnica de identificar usuários únicos ganhou na maioria das comparações realizadas com a menção, em cenários onde a quantidade de amostras obtidas por cada processo foi igual;
6. Os resultados obtidos com maiores níveis de caracterização são superiores, nos dois casos avaliados, em comparação com a técnica concorrente de contagem de menção sem caracterização;
7. Assim como os métodos tradicionais possuem suas dificuldades, que culminam no erro de previsão, nosso método também possui as suas, sendo a principal delas formar amostras representativas, com alto grau de caracterização, com baixas margens de erro. O que motiva este comportamento é que apesar de serem coletados muitos indivíduos, o perfil encontrado nos usuários do Twitter é muito concentrado em usuários jovens e de classes sociais mais elevadas; e
8. Pesquisas em redes sociais podem ser uma alternativa as pesquisas tradicionais, principalmente se levarmos em conta os custos necessário para realização do método concorrente. Mas também podem ser encaradas como uma técnica complementar, ou seja, uma nova fonte de informações.

Tabela 5.4: Resultados Eleições - p@1 (Sen: análise de sentimento; SJ: remoção de usuários *spammers* e jornalísticos; Sex: sexo; ID: idade; CS: classe social) - Quadrados de cor verde representam acertos, pretos a não formação de amostra e, consequentemente, brancos os erros.

Característica					Resultado											
Sen	SJ	Sex	ID	CS	Menção						Usuário Único					
					MG	PR	RS	RJ	BA	SP	MG	PR	RS	RJ	BA	SP
Conjunto completo de mensagens																
											-	-	-	-	-	-
X											-	-	-	-	-	-
Amostras																
X																
	X															
		X														
			X													
				X												
X	X															
	X	X														
	X		X													
	X			X												
X		X														
X			X													
X				X												
		X	X													
		X		X												
			X	X												
X	X	X														
X	X		X													
X	X			X												
X		X	X													
		X	X	X												
	X	X		X												
X	X		X	X												
X	X	X		X												
X		X	X	X												
	X	X	X	X												
X	X		X	X												
X	X	X	X	X												
Institutos Tradicionais																
IBOPE					-	-	-	-	-	-						
DataFolha					-	-	-	-	-	-						

Tabela 5.5: Resultados Eleições - Segundo turno (Sen: análise de sentimento; SJ: remoção de usuários *spammers* e jornalísticos; Sex: sexo; ID: idade; CS: classe social) - Quadrados de cor verde representam acertos, pretos a não formação de amostra e, consequentemente, brancos os erros.

Característica					Resultado											
Sen	SJ	Sex	ID	CS	Menção						Usuário Único					
					MG	PR	RS	RJ	BA	SP	MG	PR	RS	RJ	BA	SP
Conjunto completo de mensagens																
											-	-	-	-	-	-
X											-	-	-	-	-	-
Amostras																
X																
	X															
		X														
			X													
				X												
X	X															
	X	X														
	X		X													
	X			X												
X		X														
X			X													
X				X												
		X	X													
		X		X												
			X	X												
X	X	X														
X	X		X													
X	X			X												
X		X	X													
X		X		X												
		X	X	X												
	X	X		X												
	X		X	X												
X			X	X												
X	X	X	X													
X	X	X		X												
X		X	X	X												
	X	X	X	X												
X	X		X	X												
X	X	X	X	X												
Institutos Tradicionais																
IBOPE					-	-	-	-	-	-						
DataFolha					-	-	-	-	-	-						

## Capítulo 6

# Conclusões e Trabalhos Futuros

Nesta dissertação apresentamos um arcabouço para realização de pesquisas em redes sociais. O arcabouço proposto é constituído das seguintes etapas: (i) coleta e preparação dos dados; (ii) caracterização das mensagens e dos usuários; e (iii) formação de uma amostra representativa da população, com análise dos resultados. A segunda etapa representa a nossa maior contribuição.

Para caracterizar as mensagens, como não foi encontrado na literatura um algoritmo para análise de sentimentos em cenários diversos, capaz de prover bons resultados, foi proposto um método baseado em um comitê de classificação composto por um classificador léxico, um supervisionado treinado com tweets rotulados manualmente e um supervisionado com tweets rotulados automaticamente, por meio de sinais emocionais encontrados em *emoticons*.

A caracterização do indivíduo se subdividiu em duas etapas: a primeira comportamental, em que objetivamos excluir das nossas amostras usuários *spammers* e jornalísticos, e a segunda pessoal, com caracterização do sexo, faixa etária e classe social dos indivíduos.

Para a caracterização comportamental propomos um método onde um classificador *Random Forest* é treinado com um conjunto de usuários rotulados como não *spammers*, *spammers* e jornalísticos, baseado no trabalho de Benevenuto et al. [2010]. Com isso, obtivemos uma acurácia de 88.77% e média F1 de 0.88.

A etapa de caracterização do sexo é realizada em duas fases. Inicialmente é verificado se o nome do usuário consta em um dicionário contendo 20,801 nomes previamente rotulados como masculinos ou femininos e, caso contrário, um algoritmo *Naive Bayes* é utilizado. Esta técnica alcançou um acerto de aproximadamente 90%.

Para classificarmos os usuários conforme suas faixas etárias definimos os seguintes valores de interesse: menor de 25 anos, entre 25 e 45 anos e maior de 45 anos. O método

utilizado consiste na classificação pelo algoritmo *Naive Bayes Multinomial*. Esta fase alcançou uma acurácia de 81.51% e média F1 de 0.81.

A classe social dos indivíduos foi inferida novamente pelo classificador *Naive Bayes Multinomial*. Para rotular os usuários da base de treinamento, coletamos indivíduos com contas simultaneamente nas redes sociais Twitter e Foursquare. Desta forma, conforme o nível de riqueza onde a maioria das interações de cada pessoa era realizada um rótulo era atribuído como classe social baixa, média ou alta. Este processo alcançou uma acurácia de 73% e média F1 de 0.73.

Assim, analisamos bases com mensagens coletadas acerca de 14 votações do *reality show Big Brother Brasil 13 (BBB)* e das eleições municipais de 2012 em 6 capitais brasileiras. Como resultados verificamos que o arcabouço proposto é capaz de melhorar os resultados obtidos por trabalhos anteriores, baseados principalmente em técnicas simples de contagem, e é competitivo frente a métodos adotados por institutos tradicionais de pesquisa e enquetes realizadas por grandes sítios, principalmente se levarmos em consideração a diferença de custos entre uma pesquisa *on-line* de uma presencial.

## 6.1 Trabalhos Futuros

Como trabalhos futuros, sugerimos algumas melhorias no processo de classificação, o que poderia proporcionar uma maior acurácia ao arcabouço.

As caracterizações de sexo e idade já foram bem estudadas, mas considerando como pioneiro o trabalho realizado para a inferência de classes sociais ainda há muito a ser estudado. Por exemplo: (i) analisar como o uso das categorias do Foursquare poderiam melhorar o processo de rotulagem; (ii) verificar o ganho que pode ser obtido com uso de atributos não textuais no processo de classificação; (iii) incluir novas categorias de termos e verificar seu impacto; (iv) avaliar os possíveis contrastes causados por rotular os usuários somente com as informações geográficas presentes nos tweets, sem uso da rede Foursquare; (v) variar o tempo em que as interações com a rede Foursquare serão aceitas no processo de rotulação; (vi) avaliar a existência de diferenças na importância relativa entre as diversas interações com o Foursquare no processo de rotulação e atribuir pesos a elas; e (vii) analisar o processo com mensagens postadas em outros países;

Na caracterização comportamental, apesar de já alcançarmos bons resultados, o aumento da base de usuários jornalísticos poderá melhorar a etapa, distinguindo de forma mais precisa, por exemplo, usuários legítimos com contas antigas e que postam muito dos usuários jornalísticos.

A análise de sentimentos foi a caracterização identificada como a mais relevante no arcabouço, considerando o aumento de acertos proporcionado. Verificamos que a melhoria desta etapa, principalmente no que diz respeito a correta verificação de mensagens com sentimentos negativos, poderá melhorar os resultados das amostras.

Também estudamos um processo de caracterização da localização dos indivíduo. O processo apresentado trabalha com dois tipos de informações, o texto e a rede de amizades, para realizar a inferência. Não utilizamos este método para avaliar os cenários analisados, pois não havia informações oficiais sobre a distribuição geográfica dos indivíduos, caso do *Big Brother*, e para as eleições consideramos que se um usuário fala sobre o pleito de uma cidade, logo ele pertence a ela. Futuramente, este arcabouço poderá ser expandido com esta funcionalidade. Um bom ambiente para teste será as eleições nacionais que ocorrerão em outubro de 2014.

Um problema encontrado na técnica apresentada é a dificuldade de construir amostras que sejam representativas da população, com margem de erro máxima de 15 pontos percentuais. Assim, também propomos avaliar o arcabouço em um cenário mais amplo, que exista um número superior de usuários postando mensagens sobre o assunto, o que ajudaria a formar amostras mais significativas, em termos de margens de erros obtidas.

As etapas que constituem o arcabouço estão sendo incorporadas às ferramentas do Observatório da Web<sup>1</sup>.

---

<sup>1</sup><http://observatorio.inweb.org.br/>





# Referências Bibliográficas

- Arkin, H. & Colton, R. (1950). *Tables for Statisticians*. College outline series. Barnes & Noble.
- Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. Em *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pp. 492--499, Washington, DC, USA. IEEE Computer Society.
- Benevenuto, F.; Magno, G.; Rodrigues, T. & Almeida, V. (2010). Detecting spammers on twitter. Em *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.
- Blumenstock, J. & Fratamico, L. (2013). Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data. Em *Proc. of the 4th Annual Symposium on Computing for Development*, ACM DEV-4 '13, pp. 11:1--11:10.
- Bourdieu, P. (1987). What makes a social class? on the theoretical and practical existence of groups. *Berkeley Journal of Sociology*, 32:1--18.
- Bracarense, P. A. (2009). *Estatística Aplicada às Ciências Sociais*. Iesde Brasil Sa. ISBN 9788538709596.
- Calais Guerra, P. H.; Veloso, A.; Meira, Jr., W. & Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. Em *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pp. 150--158, New York, NY, USA. ACM.
- Cheng, Z.; Caverlee, J. & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. Em *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 759--768. ACM.

- Crompton, R. (2008). *Class and Stratification*. MPG Books.
- Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*:1695.
- Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. Em *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, pp. 115--122, New York, NY, USA. ACM.
- Davis Jr., C. A.; Pappa, G. L.; de Oliveira, D. R. R. & de L. Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735--751. ISSN 1467-9671.
- Folha, D. (2012). Data folha - dúvidas. [http://datafolha.folha.uol.com.br/duvida/-duvidas\\_pesquisa.php](http://datafolha.folha.uol.com.br/duvida/-duvidas_pesquisa.php). Access date: 23 novembro 2012.
- G1 (2012a). Facebook alcança 1 bilhão de usuários ativos mensais. <http://g1.globo.com/tecnologia/noticia/2012/10/facebook-atinge-1-bilhao-de-usuarios-ativos-mensais.html>. Access date: 25 novembro 2012.
- G1 (2012b). LinkedIn anuncia 10 milhões de usuários no brasil. <http://g1.globo.com/tecnologia/noticia/2012/10/linkedin-anuncia-10-milhoes-de-usuarios-no-brasil.html>. Access date: 25 novembro 2012.
- G1 (2012c). Twitter chega a 500 milhões de usuários, diz estudo. <http://g1.globo.com/tecnologia/noticia/2012/07/twitter-chega-500-milhoes-de-usuarios-diz-estudo.html>. Access date: 25 novembro 2012.
- Gayo-Avello, D. (2012). "I wanted to predict elections with twitter and all i got was this lousy paper" – A balanced survey on election prediction using twitter data.
- Gayo-Avello, D.; Metaxas, P. & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 490--493.
- Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398--409.
- Ghosh, S.; Viswanath, B.; Kooti, F.; Sharma, N. K.; Korlam, G.; Benevenuto, F.; Ganguly, N. & Gummadi, K. P. (2012). Understanding and combating link farming in the twitter social network. Em *Proceedings of the 21st international conference on World Wide Web, WWW '12*, pp. 61--70, New York, NY, USA. ACM.

- Gil, A. (2010). *Como elaborar projetos de pesquisa*. Atlas. ISBN 9788522458233.
- Gomide, J.; Veloso, A.; Meira, Jr., W.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. Em *Proceedings of the 3rd International Web Science Conference, WebSci '11*, pp. 3:1--3:8, New York, NY, USA. ACM.
- Gonçalves, P.; Araújo, M.; Benevenuto, F. & Cha, M. (2013). Comparing and combining sentiment analysis methods. Em *Proceedings of the First ACM Conference on Online Social Networks, COSN '13*, pp. 27--38, New York, NY, USA. ACM.
- Goswami, S.; Sarkar, S. & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 214--217.
- Hu, X.; Tang, J.; Gao, H. & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. Em *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13*, pp. 607--618, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- IBGE (2010). Brazilian Census. [http://www.ibge.gov.br/english/estatistica/-populacao/censo2010/default\\_resultados\\_universo.shtm](http://www.ibge.gov.br/english/estatistica/-populacao/censo2010/default_resultados_universo.shtm). Access date: 30 janeiro 2014.
- IBOPE (2012a). Metodologia. <http://www.eleicoes.ibope.com.br/Paginas/-Metodologia.aspx>. Access date: 23 novembro 2012.
- IBOPE (2012b). Metodologia de pesquisa. [http://ibope.com.br/calandraWeb/-BDarquivos/sobre\\_pesquisas/metodologia\\_pesquisa.html](http://ibope.com.br/calandraWeb/-BDarquivos/sobre_pesquisas/metodologia_pesquisa.html). Access date: 23 novembro 2012.
- Institute, O. I. (2012). PObama wins the election! (on twitter). <http://www.zerogeography.net/2012/11/obama-wins-election-on-twitter.html>. Access date: 25 novembro 2012.
- Jungherr, A.; Jürgens, P. & Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with twitter: What 140 characters reveal about political sentiment". *Soc. Sci. Comput. Rev.*, 30(2):229--234. ISSN 0894-4393.

- Katz-Gerro, T. (1999). Cultural consumption and social stratification: leisure activities, musical tastes, and social location. *Sociological Perspectives*, pp. 627--646.
- Lin, F. & Cohen, W. W. (2010). Semi-supervised classification of network data using very few labels. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 0:192--199.
- Lumezanu, C.; Feamster, N. & Klein, H. (2012). #bias: Measuring the tweeting behavior of propagandists. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 210--217.
- Macskassy, S. A. & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935--983.
- Mahmud, J.; Nichols, J. & Drews, C. (2012). Where is this tweet from? inferring home locations of twitter users. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 511--514.
- Manning, C. D.; Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- McCord, M. & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. Em *Proceedings of the 8th international conference on Autonomic and trusted computing, ATC'11*, pp. 175--186, Berlin, Heidelberg. Springer-Verlag.
- Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P. & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 554--557.
- Morstatter, F.; Pfeffer, J.; Liu, H. & Carley, K. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 400--408.
- Mukherjee, S.; Malu, A.; A.R., B. & Bhattacharyya, P. (2012). Twisent: a multistage system for analyzing sentiment in twitter. Em *Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM '12*, pp. 2531--2534, New York, NY, USA. ACM.
- Nguyen, D.; Gravel, R.; Trieschnigg, D. & Meder, T. (2013). "How old do you think i am?": A study of language and age in twitter. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, ICWSM 2013, pp. 439--448.

- Peersman, C.; Daelemans, W. & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. Em *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pp. 37--44, New York, NY, USA. ACM.
- Pennacchiotti, M. & Popescu, A.-M. (2011). Democrats, republicans and starbucks aficionados: user classification in twitter. Em *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pp. 430--438. ACM.
- PPP (2012). Our methodology. <http://www.publicpolicypolling.com/aboutPPP/about-us.html>. Access date: 23 novembro 2012.
- Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A. & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. Em *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pp. 249--252, New York, NY, USA. ACM.
- Ribeiro Jr., S.; Jr., Z.; Meira Jr., W. & Pappa, G. L. (2012). Positive or negative? using blogs to assess vehicles features. *ENIA 2012 - Brazilian Conference on Intelligent System*.
- Rodrigues, E.; Assuncao, R.; Pappa, G.; Miranda Filho, R. & Meira Jr., W. (2013). Uncovering the location of twitter users. Em *Brazilian Conference on Intelligent Systems*, BRACIS-13, pp. 237--241.
- Schiffman, L. (2007). *Consumer behavior*. Pearson Prentice Hall, Upper Saddle River, NJ, 9. ed. edição. ISBN 0131869604.
- Schler, J.; Koppel, M.; Argamon, S. & Pennebaker, J. (2006). Effects of age and gender on blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 199--205.
- Settle, R. B.; Alreck, P. L. & Belch, M. A. (1979). Social class determinants of leisure activity. volume 6.
- Silva, I. S.; Gomide, J.; Veloso, A.; Meira, Jr., W. & Ferreira, R. (2011). Effective sentiment stream analysis with self-augmenting training and demand-driven projection. Em *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pp. 475--484, New York, NY, USA. ACM.

- Solomon, M. (2010). *Consumer behavior: buying, having, and being*. Prentice Hall. ISBN 9780136110927.
- Terra (2012). Google plus tem mais de 400 milhões de usuários; 100 mi ativos. <http://tecnologia.terra.com.br/noticias/0,,OI6160879-EI12884,00-Google+Plus+tem+mais+de+milhoes+de+usuarios+mi+ativos.html>. Access date: 25 novembro 2012.
- Times, T. N. Y. (2012). President map. <http://elections.nytimes.com/2012/-results/president>. Access date: 25 novembro 2012.
- Tumasjan, A.; Sprenger, T.; Sandner, P. & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Em *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 178--185.
- Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Em *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pp. 417--424, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2nd edição. ISBN 0-12-088407-0.
- Wright, E. O. (2003). *Encyclopedia of Social Theory*, capítulo Social Class. Sage Publications.
- Wu, F. Y. (1982). The potts model. *Reviews of Modern Physics*, 54:235--268.
- Zhang, H. (2004). The Optimality of Naive Bayes. Em Barr, V. & Markov, Z., editores, *FLAIRS Conference*. AAAI Press.

# Apêndice A

## Localização

No Twitter a informação sobre a localização do usuário pode ser informada de três maneiras distintas, cada uma delas com graus diferentes de precisão e acurácia. O usuário pode informar em seu perfil o local onde mora. Como esse campo pode ser preenchido livremente, um grande volume de localizações inválidas, como “Marte”, ou com baixa precisão, como “Brasil”, são informadas pelos usuários. A segunda maneira consiste em obter a localização geográfica a partir do endereço de IP da máquina. Esse tipo de georeferenciamento não é muito confiável e deve ser atualizado continuamente. No Brasil, por exemplo, esse serviço localiza corretamente 72% dos IP’s dentro de um raio de 40 quilômetros. A terceira forma é obtida a partir das coordenadas do GPS de aparelhos celulares. Esse terceiro tipo é o que tem maior precisão e confiabilidade, porém, como está restrito aos casos em que o usuário posta a mensagem de um aparelho celular com GPS e ainda permite que essa informação seja divulgada, esse tipo de informação geográfica está presente apenas em uma pequena fração dos *tweets*. Em alguns países como o Brasil, essa proporção não passa de 1%.

Apesar da informação geográfica não ser explícita em grande parte dos casos, alguns aspectos sobre o comportamento do usuário podem nos dar dicas sobre sua localização. Por exemplo, o conjunto de *tweets* publicados por um usuário podem nos fornecer informação sobre onde ele reside. Alguns trabalhos têm sido desenvolvidos nesse sentido. [Cheng et al., 2010] estimam a localização do usuário identificando palavras que caracterizam determinadas localizações, com por exemplo o termo “rockets” que está associado à cidade de Houston. Os autores definem que esse tipo de palavra deve ter uma alta frequência em um determinado ponto do espaço e essa frequência deve cair rapidamente quando nos afastamos desse ponto. Mahmud et al. [2012] inferem a localização do usuário, também com base no texto, utilizando algoritmos de classificação. Esses últimos são capazes de inferir a localização para diferentes níveis

de granularidade: cidade, estado e zona temporal.

Além do texto, as relações de seguidor/seguido entre os usuários também podem nos trazer informação geográfica. Sabe-se que, principalmente em países onde a língua falada não é o inglês, as relações de amizade no *Twitter* tendem a refletir a proximidade geográfica entre usuários. Tendo isso em vista, a rede de amizades pode ser usada como fonte de informação para o processo de inferência. [Davis Jr. et al., 2011] propõem um método de estimação segundo o qual localização de um usuário será aquela mais frequente entre seus amigos. Ao determinar as relações de amizade consideram que dois usuários são amigos apenas se eles se seguem mutuamente. Isso evita que páginas institucionais ou perfis de celebridades atrapalhem o processo de inferência. Ao longo deste trabalho consideraremos essa mesma definição de amizade entre os usuários. Alguns dos problemas encontrados por esses autores se referem ao reduzido número de amigos que alguns usuários possuem, o que dificulta bastante o processo de inferência. Além disso, usuários com muitos amigos também são fonte de erros, visto que tais relações de amizades, muito provavelmente, não refletem a proximidade geográfica.

A nossa principal contribuição nesta etapa é apresentar um método que seja capaz de incorporar os dois tipos de informação, o texto e a rede de amizades, para fazer inferência sobre a localização do usuário.

## A.1 Metodologia

Seja  $\theta_i$  a localização do usuário  $i$ , onde estamos interessados na localização a nível de cidade. Chamaremos de  $\boldsymbol{\theta}_{-i}$  o vetor com as localizações de todos usuários, com exceção do  $i$ -ésimo. O vetor  $\mathbf{w}_i$  será formado por todas as palavras dos *tweets* postados pelos usuários nos últimos tempos. O nosso objetivo é, para um usuário  $i$  cuja localização é desconhecida, encontrar o valor mais provável de  $\theta_i$ . Para encontrarmos esse valor, precisamos da distribuição de probabilidade de todo o vetor  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_N)$ , onde  $N$  é o número total de usuários analisados. Essa distribuição será obtida através de um amostrador de *Gibbs* [Gelfand & Smith, 1990]. Vamos gerar os valores dos  $\theta_i$  desconhecidos a partir da seguinte distribuição condicional:

$$P(\theta_i | \boldsymbol{\theta}_{-i}, \mathbf{w}_i).$$

Sabe-se que a amostra do vetor de  $\theta$ 's gerados dessa maneira terá uma distribuição que aproxima a conjunta  $P(\boldsymbol{\theta})$ .

Para encontrarmos a expressão dessa distribuição condicional, podemos fazer al-



gumas simplificações. Sabemos que

$$P(\theta_i | \boldsymbol{\theta}_{-i}, \mathbf{w}_i) \propto P(\theta_i | \boldsymbol{\theta}_{-i}) P(\mathbf{w}_i | \theta_i, \boldsymbol{\theta}_{-i}).$$

Vamos analisar primeiramente o termo  $P(\theta_i | \boldsymbol{\theta}_{-i})$ . É razoável supor que as localizações dos usuários no grafo seguem um Campo de Markov. Dessa maneira  $P(\theta_i | \boldsymbol{\theta}_{-i})$  se simplifica em

$$P(\theta_i | \boldsymbol{\theta}_{-i}) = P(\theta_i | \boldsymbol{\theta}_{\partial i})$$

onde o vetor  $\boldsymbol{\theta}_{\partial i}$  contém a localização de todos os vizinhos de  $i$ . Essa é uma suposição razoável, visto que se fornecemos a informação sobre a localização dos amigos de um usuário, toda a informação contida no resto da rede é dispensável.

Vamos supor ainda que a distribuição das localizações dos usuários pode ser modelada por um campo markoviano denominado Modelo de Potts [Wu, 1982]. Esse modelo é uma generalização de um modelo mais simples, o Modelo de Ising. No modelo de Ising cada local pode pertencer a duas classes, o que seria equivalente a termos apenas duas localizações. Já no modelo de Potts podemos ter um número arbitrário de classes ou de localizações. De acordo com esse modelo, probabilidade de um local pertencer a uma determinada classe será uma função crescente do número de vizinhos desse local pertencentes à essa classe, ou seja

$$P(\theta_i | \boldsymbol{\theta}_{\partial i}) \propto \exp \left( \beta \sum_{j: j \in \partial i} \sigma_{ij} \right)$$

onde  $\sigma_{ij}$  é uma função que recebe valor 1 se  $i$  e  $j$  pertencem à mesma classe e zero, caso contrário. O parâmetro  $\beta$  é conhecido como a temperatura do modelo e mede o grau de interação entre os locais. Para  $\beta > 0$  temos um modelo atrativo, ou seja, locais vizinhos tenderão pertencer à mesma classe.

Vejamos agora como podemos simplificar o termo  $P(\mathbf{w}_i | \theta_i, \boldsymbol{\theta}_{-i})$ . Primeiramente, observe que se queremos prever o texto de um usuário, dado que sabemos sua localização, a informação geográfica sobre seus amigos é desnecessária. Dessa maneira, essa probabilidade se simplifica a

$$P(\mathbf{w}_i | \theta_i, \boldsymbol{\theta}_{-i}) = P(\mathbf{w}_i | \theta_i).$$

Para encontrarmos o valor de  $P(\mathbf{w}_i | \theta_i)$  utilizaremos o método *Naive Bayes*. Vamos

considerar que as palavras postadas pelo usuário são independentes entre si, ou seja

$$P(\mathbf{w}_i|\theta_i) = \prod_j P(w_{ij}|\theta_i)$$

onde  $w_{ij}$  denota a  $j$ -ésima palavra publicada pelo  $i$ -ésimo usuário. Essa suposição é aparentemente pouco razoável, porém método *Naive Bayes* têm mostrados ótimos resultados apesar de sua simplicidade [Zhang, 2004]. Cada uma das probabilidade  $P(w_{ij}|\theta_i)$  pode ser estimada como a proporção de vezes que a palavra  $w_{ij}$  aparece dentre todas as palavras publicadas por usuários que residem na localização  $\theta_i$ .

Temos então que a probabilidade condicional  $P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i)$  fica na forma

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i) \propto P(\theta_i|\boldsymbol{\theta}_{\partial i}) \prod_j P(w_{ij}|\theta_i).$$

E os valores de  $\theta_i$  para aqueles usuários cuja localização é desconhecida podem ser atualizados através do amostrador de *Gibbs*.

## A.2 Avaliação do método

A fim de verificar a adequação do método proposto consideramos um conjunto de 8,477 usuários do *Twitter* residentes nas cidades de Belo Horizonte, Rio de Janeiro e São Paulo. O número total de usuários coletados dessas três cidades são respectivamente 1,402, 4,061 e 3,014. Consideramos a princípio apenas essas três localidades, pois em cidades muito pequenas a informação geográfica disponível não é suficiente para fazermos inferência.

Esses 8,477 usuários formam um grafo composto por 140,715 arestas, o qual é apresentado na Figura A.1. Notamos que existe uma componente fortemente conectada nesse grafo e vários usuários isolados, com menos de dois amigos. Sobre o conteúdo foram coletados os 200 *tweets* mais recentes postadas pelos usuários. A fim de validar o método, 70% dos usuários foram selecionados aleatoriamente para compor o conjunto de treinamento (5,933 usuários), e os restantes, 30% do conjunto, para validação (2,544 usuários), onde 16.54% são de Belo Horizonte, 47.91% de São Paulo e 35.55% do Rio de Janeiro.

Os resultados foram comparados com os outros dois métodos, o primeiro baseado no grafo de amizade e o segundo sobre o conteúdo dos *tweets*. O primeiro *baseline* foi o MRW, método proposto por [Lin & Cohen, 2010]. Implementamos este algoritmo usando o *Igraph* do pacote R [Csardi & Nepusz, 2006] e fixamos a probabilidade de

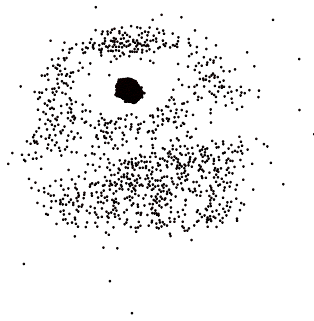


Figura A.1: Grafo de amizades na base com 8,477 usuários do Twitter

Tabela A.1: Localização: precisão dos métodos (%)

Método	Belo Horizonte	São Paulo	Rio	Total
MRW	23.14	55	39.78	50.27
Naive Bayes	42.15	63.7	55.99	60.46
IDA (tf)	38.02	82.95	59.26	73.98
IDA (tf-idf)	33.88	84.01	59.26	74.79

teleportação em 0.5. O segundo *baseline* foi o classificador Naive Bayes, usando apenas o conteúdo de *tweets* para inferir a localização do usuário. O vetor de termos dado aos algoritmos foi obtido excluindo palavras *stopping words* e termos com frequência inferior a três. Os últimos foram excluídos porque são prováveis erros de digitação ou palavras inexistentes. Este processo resultou em um conjunto com 5,557,173 palavras.

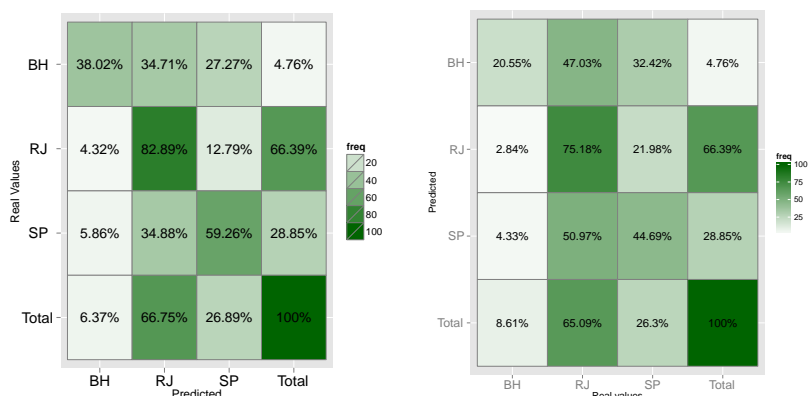
A Tabela A.1 mostra os resultados obtidos pelos três métodos considerando os 2,544 usuários no conjunto de validação. Observe que a MRW previu corretamente a localização de 50.27% dos usuários, enquanto Naive Bayes previu corretamente 60.46%. Assim, a informação a partir do conteúdo de mensagens é mais importante para prever a localização do que as ligações de amizade no grafo por si só. Aqui temos de levar em conta que é mais fácil para MRW para prever a localização dos lugares com maior número de usuários.

Os resultados obtidos pelo método proposto são relatados nas duas últimas linhas da Tabela A.1 (IDA - abordagem de dados - integrados), e a localização do  $i$ -th usuário é estimado pela a maior probabilidade entre todas as localizações possíveis. Por enquanto, vamos nos focar nos resultados de IDA (tf), que considera a frequência

das palavras nos *tweets*. Em geral, o método infere correctamente a localização de 73.98% dos usuários. Para os usuários que vivem em Belo Horizonte, São Paulo e Rio de Janeiro a proporção de inferências corretas foram, respectivamente, 38.02%, 82.95% e 59.26%. Combinando os dois tipos de dados melhorou a precisão global, mas para Belo Horizonte, a precisão diminuiu de 42.15 para 38.02. Isto pode ser devido a variância da amostra, não refletindo qualquer aspecto intrínseco do problema. Na verdade, a diferença entre as taxas de sucesso dos dois métodos não é estatisticamente significativas. Isto pode ser visto através da construção de um intervalo de confiança de 95% para a taxa obtida pelo método de Bayes Naive, que é dado por [33.21, 50.79], cobrindo a nova taxa de sucesso de 38.02%

Como as classes estão desbalanceadas a Figura A.2 apresenta duas matrizes de confusão sendo uma para medida do recall e outra para a precisão. A matriz de recall mostra a probabilidade de o usuário ser classificado como sendo da cidade de  $i$ , já que ele é de fato da cidade de  $i$ . Por exemplo, um usuário do Rio de Janeiro tem uma probabilidade de 82.89% para ser classificado como sendo do Rio de Janeiro; 4.32% de chance de ser classificado como sendo de Belo Horizonte e 12.79% de chance de ser classificado como de São Paulo. A segunda matriz mostra os valores de precisão, ou seja, a probabilidade de um usuário ser da cidade de  $i$ , uma vez que foi classificado como sendo daquela cidade. Por exemplo, se um usuário é classificado como sendo de Rio de Janeiro, ele tem 82.46% de chance de realmente ser de lá, 2.47% a ser de Belo Horizonte e 15.07% de chance de ser de São Paulo. Observe que, para São Paulo, a maioria dos erros consistem em classificar o usuário como sendo do Rio de Janeiro. Isso pode ocorrer devido a uma maior interação entre os usuários dessas duas cidades.

Figura A.2: Matrizes de confusão alcançadas com a aplicação da metodologia proposta. A primeira matriz mostra as medidas de recall e a segunda a medida da precisão



A desvantagem do pré-processamento realizado nos *tweets* é que ele usa um

grande conjunto de termos, a maioria dos quais provavelmente não ajudam a diferenciar uma cidade da outra. Assim, mudamos a nossa maneira de atribuir pesos aos termos substituindo a frequência pelos  $tf - idf$  (*term frequency-inverse document frequency*) Manning et al. [2008]. O valor do  $tf - idf$  é alto quando o termo é raro em toda a base, mas muito comum no documento em análise. Vamos denotar por  $D$  o conjunto completo de documentos (no nosso caso o conjunto de *tweets*),  $d$  um documento específico (*tweet*) e  $t$ , um termo específico. O  $tf - idf$  é composto por duas partes. O primeiro é a frequência do termo, dada por

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$

onde  $f(t, d)$  é a frequência do termo  $t$  em um documento  $d$ . A segunda parte é a frequência do documento inversa, definida como

$$idf(t, D) = \log \frac{|D|}{\{d \in D : t \in d\}}$$

onde  $|D|$  é o número total de documentos e  $\{d \in D : t \in d\}$  é o número de documentos em que o termo  $t$  ocorre. O  $tf - idf$  é dado por uma combinação destes dois termos, definidos como

$$tfidf(t, d) = tf(t, d) \times idf(t, D).$$

A Figura A.3 apresenta os valores do recall e da precisão alcançada usando  $tf - idf$ . Observe que os resultados obtidos são muito semelhantes aos que temos, sem qualquer fator de ponderação. Assim, talvez uma seleção mais inteligente de atributos relevantes deve ser realizada antes de estimar as probabilidades. Deixamos isso para trabalhos futuros.

Esta etapa do trabalho foi publicada em [Rodrigues et al., 2013].

Figura A.3: Matrizes de confusão dos resultados alcançados através da aplicação da metodologia proposta e correção pelo fator Tf-Ifd. A primeira matriz mostra as medidas de recall e a segunda a medida da precisão

