GABRIEL DOMINGOS DE ARRUDA

Análise de viés em notícias na língua
portuguesa

São Paulo

2016

GABRIEL DOMINGOS DE ARRUDA

Análise de viés em notícias na língua portuguesa

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de Concentração: Inteligência Artificial

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 02 de dezembro de 2015. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Norton Trevisan Roman

São Paulo

2016

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Arruda, Gabriel Domingos de

Análise de viés em notícias na língua portuguesa / Gabriel Domingos de Arruda ; orientador, Norton Trevisan Roman. – São Paulo, 2016

97 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Versão corrigida

1. Inteligência artificial. 2. Análise de texto - Sistemas computacionais. 3. Notícia - Análise do discurso - Sistemas computacionais. 4. Valores atípicos. I. Roman, Norton Trevisan, orient. II. Título

CDD 22.ed.— 006.3

Dissertação de autoria de Gabriel Domingos de Arruda, sob o título “**Análise de viés em notícias na língua portuguesa**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Sistemas de Informação, aprovada em 02 de dezembro de 2015 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Norton Trevisan Roman

Presidente

Instituição: Universidade de São Paulo

Prof. Dr. Ivandre Paraboni

Instituição: Universidade de São Paulo

Profa. Dra. Ariadne Maria Brito Rizzoni Carvalho

Instituição: Universidade Estadual de Campinas

Agradecimentos

Aos meus pais pelo apoio incondicional a esse projeto, mesmo quando eu estava em dúvida.

Ao meu orientador, Prof. Dr. Norton Trevisan Roman, por todo o suporte nesses anos sempre acompanhado de bom humor.

Aos anotadores, que dispuseram de seu tempo livre para tornar esse projeto possível.

Aos professores e colegas do PPgSI pelo conhecimento compartilhado.

*“The fundamental cause of the trouble is that in the modern world the stupid are cocksure
while the intelligent are full of doubt.”
(Bertrand Russell)*

Resumo

ARRUDA, Gabriel Domingos. **Análise de viés em notícias na língua portuguesa**. 2016. 97 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

O projeto descrito neste documento propõe um modelo para análise de viés em notícias, procurando identificar o viés dos meios de comunicação em relação a entidades políticas. Foram analisados três tipos de viés: o viés de seleção, que avalia o quanto uma entidade é referenciada pelo meio de comunicação; o viés de cobertura, que avalia quanto destaque é destinado a entidade e, por fim, o viés de afirmação, que avalia se estão falando mal ou bem da entidade. Para tal, foi construído um corpus de notícias sistematicamente extraídas de 5 produtores de notícias e classificadas manualmente em relação à polaridade e entidade alvo. Técnicas de análise de sentimentos baseadas em aprendizado de máquina foram validadas utilizando o corpus criado. Criou-se uma metodologia para identificação de viés, utilizando o conceito de *outliers*, a partir de métricas indicadoras. A partir da metodologia proposta, foi analisado o viés em relação aos candidatos ao governo de São Paulo e à presidência a partir do corpus criado, em que se identificou os três tipos de viés em dois produtores de notícias.

Palavras-chaves: Análise de sentimentos. Detecção de *outliers*. Viés.

Abstract

ARRUDA, Gabriel Domingos. **Bias analysis on newswire in portuguese**. 2016. 97 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2015.

The project described here proposes a model to study bias on newswire texts, related to political entities. Three types of bias are analysed: selection bias, which refers to the amount of times an entity is referenced by the media outlet; coverage bias, which assesses the amount of coverage given to an entity and, finally, the assertion bias, which analyses whether the news is a positive or negative report of an entity. To accomplish this, a corpus was systematically built by extracting news from 5 different newswires. These texts were manually classified according to their polarity alignment and associated entity. Sentiment Analysis techniques were applied and evaluated using the corpus. Based on the concept of outliers, a methodology for bias detection was created. Bias was analysed using the proposed methodology on the generated corpus for candidates to the government of the state of São Paulo and to presidency, being identified in two newswires for the three above-defined types.

Keywords: Sentiment Analysis. Outliers detection. Bias.

Lista de figuras

Figura 1 – Árvore sintática da frase “Mary brought a cat.”	25
Figura 2 – Exemplo de sub-árvores sintáticas	25
Figura 3 – Exemplo de grafo de relação entre as palavras	26
Figura 4 – Árvores para representação de <i>frames</i> semânticos	28
Figura 5 – Representação gráfica da árvore de dependências da sentença “ <i>Bell, based in Los Angeles, makes and distribute eletronic, computer and builiding products.</i> ”	34
Figura 6 – Proporção de classificadores utilizados	35
Figura 7 – Exemplo de SVM para dados linearmente separáveis	36
Figura 8 – Exemplo de classificador Naïve Bayes	38
Figura 9 – Árvore sintática de polaridade	40
Figura 10 – Exemplo de grafo construído utilizado para cálculo das probabilidades condicionais	42
Figura 11 – Janela de tamanho 3 produzido por uma sentença	60
Figura 12 – Gráficos representando a diferença entre a distância euclidiana (a) e distância de Mahalanobis (b)	82

Lista de algoritmos

Algoritmo 1 – Algoritmo de coleta de notícias	51
---	----

Lista de tabelas

Tabela 1 – Resultados das buscas por artigos em cada uma das bases	19
Tabela 2 – Exemplo de <i>frames</i> semânticos	27
Tabela 3 – Funções de similaridade	30
Tabela 4 – Palavras sementes com polaridade manualmente definida	30
Tabela 5 – Diferenças de estimação e probabilidade real	39
Tabela 6 – Características de polaridade e subjetividade	43
Tabela 7 – Perfis do Twitter selecionados	51
Tabela 8 – Detalhe dos anotadores	52
Tabela 9 – Concordância entre anotadores para polaridade e entidade alvo	53
Tabela 10 – Concordância por pares entre anotadores para a dimensão entidade alvo	54
Tabela 11 – Concordância por pares entre anotadores para a dimensão Polaridade ₂	55
Tabela 12 – Padrão ouro do processo de anotação	56
Tabela 13 – Resultados do classificador Naïve Bayes	62
Tabela 14 – Resultados do classificador SVM	63
Tabela 15 – Resultados do classificador por entropia	63
Tabela 16 – Características de polaridade e subjetividade	64
Tabela 17 – Resultados do classificador por entropia com dimensões adicionais	65
Tabela 18 – Resultados do classificador utilizando kernel polinomial	65
Tabela 19 – Resultados do classificador Naïve Bayes (entidade Dilma Rousseff)	68
Tabela 20 – Resultados do classificador SVM (entidade Dilma Rousseff)	68
Tabela 21 – Resultados do classificador por entropia (entidade Dilma Rousseff)	68
Tabela 22 – Resultados do classificador Naïve Bayes (entidade Marina Silva)	69
Tabela 23 – Resultados do classificador SVM (entidade Marina Silva)	69
Tabela 24 – Resultados do classificador por entropia (entidade Marina Silva)	70
Tabela 25 – Proporção de referências aos candidatos por perfil	76
Tabela 26 – Desvio do perfil em relação à proporção de referências aos candidatos por perfil	77
Tabela 27 – Proporção de referências aos candidatos nos tweets	78
Tabela 28 – Desvio em relação à proporção de referências aos candidatos nos tweets	78
Tabela 29 – Proporção de parágrafos classificados como positivos para cada candidato	79
Tabela 30 – Desvio em relação à proporção de parágrafos classificados como positivos	79

Tabela 31 – Proporção de parágrafos classificados como neutros para cada candidato	80
Tabela 32 – Desvio em relação à proporção de parágrafos classificados como neutros	80
Tabela 33 – Proporção de parágrafos classificados como negativos para cada candidato	80
Tabela 34 – Desvio em relação a proporção de parágrafos classificados como negativos	81
Tabela 35 – Distância de Mahalanobis em relação ao centro de gravidade	82
Tabela 36 – Correlação de Pearson entre os diferentes tipos de vieses	83

Sumário

1	Introdução	14
2	Objetivo	16
3	Conceitos Fundamentais	18
3.1	Definindo viés em notícias	19
3.2	Análise de sentimentos em notícias	20
3.2.1	Identificação de polaridade em notícias	21
3.2.2	Modelos de representação de textos	23
3.2.2.1	<i>Bag of Words</i>	23
3.2.2.2	Representação por árvores sintáticas	24
3.2.2.3	Representação por grafo	26
3.2.2.4	Representação por <i>Frames</i> semânticos	27
3.2.3	Classificação baseada em dicionários	28
3.2.3.1	Polaridade de palavras	28
3.2.3.2	Métodos de classificação	32
3.2.4	Classificação baseada em aprendizado de máquina	35
3.2.4.1	Classificadores	35
3.2.4.1.1	SVM	36
3.2.4.1.2	Naïve Bayes	38
3.2.4.2	Métodos de classificação	40
3.2.4.2.1	Classificação por árvore sintática de polaridade	40
3.2.4.2.2	Classificação por entropia de conexão entre as palavras	41
3.2.5	Avaliação dos métodos de classificação	44
3.3	Viés não presente no texto	46
4	Criação do corpus	49
4.1	Coleta de notícias e metodologia de anotação	50
4.2	Resultados e discussão	52
4.3	Padrão Ouro do Corpus	55
5	Experimentos de Análise de Sentimentos	57

5.1	Classificação Naïve Bayes e SVM	57
5.1.1	Stemming	58
5.1.2	Tipo de representação	59
5.2	Classificação por entropia	59
5.2.1	Etiquetador POS	59
5.3	Resultados da classificação por polaridade	61
5.3.1	Alterando o classificador por entropia	64
5.3.2	Classificação condicionada à entidade	66
5.4	Discussão	70
6	Análise de viés em notícias	73
6.1	Identificação de <i>outliers</i>	73
6.2	Viés de seleção	75
6.3	Viés de cobertura	77
6.4	Viés de afirmação	78
6.4.1	Distância de Mahalanobis para o viés de afirmação	80
6.5	Correlação entre os diferentes tipos de vieses	83
6.6	Discussão	84
7	Conclusão	86
7.1	Criação de Corpus	86
7.2	Análise de sentimentos sobre as notícias do corpus	86
7.3	Análise de viés a partir <i>outliers</i>	87
	Referências ¹	88

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A mídia tem um papel fundamental nas sociedades democráticas, provendo informações para a população. Entretanto, a confiança nos meios de comunicação não é plena. Nesse sentido, um estudo feito pela American Society of Newspaper Editors revelou que 78% do público acredita que há viés na publicação de notícias nos EUA (URBAN, 1999), indicando que o questionamento da isenção da mídia é comum nesse país. No Brasil, uma pesquisa do Datafolha¹, realizada na cidade de São Paulo, diagnosticou uma queda no prestígio da imprensa, em que 74% dos entrevistados atribuíram “muito prestígio” para a imprensa em 2003, tendo esse número caído para 61% em 2013.

A percepção da mídia não ser confiável não é uma exclusividade dos EUA ou Brasil. De fato, (PARK et al., 2009) definem o viés como um problema inerente à produção de notícias, não sendo possível para um meio de comunicação ser completamente neutro. Contudo, mesmo o problema não sendo completamente solúvel considerando-se apenas um meio de comunicação, isso não impede que seja possível reduzi-lo, procurando-se meios de comunicação distintos para melhor se informar.

Por outro lado, apesar do viés ser um problema inerente à produção de notícias, os leitores têm dificuldades em identificá-lo de forma objetiva. Nesse sentido, (SCHMITT, 2004) constatou que pessoas neutras e partidárias classificam notícias de formas distintas, à medida que um artigo considerado neutro por uma pessoa pode ser considerado positivo ou negativo por outra, de acordo com o seu posicionamento sobre o assunto. Além disso, existe também a possibilidade de existência do efeito câmara de eco (IYENGAR; HAHN, 2009), segundo o qual tendemos a ler notícias que servem para reforçar nossas atitudes e crenças pré-existentes. Ou seja, além do viés dos produtores de mídia, é necessário lidar com o viés de nossas próprias convicções.

Mais do que isso, é possível que os novos sistemas de recomendação de conteúdo estejam agravando o problema do viés na mídia (PARISER, 2011), à medida que serviços da internet, como a busca do Google ou o *feed* de novidades do Facebook, implementam filtros que trazem conteúdo personalizado de acordo com o perfil do usuário. Esses filtros não são claros para os usuários, agravando ainda mais o efeito da “câmara de eco”.

¹ <http://media.folha.uol.com.br/datafolha/2013/06/19/protestos-aumento-tarifa-ii.pdf>

Nesse sentido, como uma iniciativa atual que procura auxiliar nessa questão de parcialidade da mídia, o projeto Manchetômetro² da UERJ analisou as manchetes dos principais meios de comunicação nas últimas eleições para determinar o viés em relação aos candidatos à presidência. Nele, uma dupla de anotadores analisou se a manchete era positiva, negativa, neutra ou ambivalente em relação ao candidato. Se houvesse discordância entre os codificadores, um terceiro codificador funcionaria como critério de desempate.

Ainda que o Manchetômetro tenha procurado realizar o trabalho de forma objetiva, não é possível aplicar esse processo para toda a mídia. É necessário tempo para que o processo de codificação seja realizado, o que torna a análise de viés menos útil para temas urgentes. Nesse sentido, a automatização traz diversas vantagens para esse processo de análise de viés, como aumento da escala, agilidade na análise e redução do problema da subjetividade dos codificadores.

Trabalhos para construir sistemas que classificam notícias podem ser encontrados, ainda que em idiomas diferentes de português, como o japonês (ZHANG et al., 2011), coreano (PARK et al., 2012) e inglês (*e.g.* (BALAHUR et al., 2009), (LIN; BAGROW; LAZER, 2012) e (SUDHAHAR; LANSDALL-WELFARE, 2012)). Até onde foi possível verificar, não há nenhum trabalho similar para a língua portuguesa. Nesse projeto pretende-se então ajudar a preencher essa lacuna, estudando a classificação de notícias em português.

Classificar notícias, contudo, não abrange todo o problema do viés. Por esse motivo, o presente trabalho não se restringirá apenas ao conteúdo das notícias, uma vez que a omissão sobre determinado assunto, ou mesmo a atenção exagerada a algum outro, também configuram viés. Tais fatores, contudo, não podem ser analisados com uma simples análise de valência da notícia, dependendo também de métricas quantitativas, como a quantidade de referências a determinada entidade, ou a presença de determinado assunto em alguns provedores de notícias e em outros não, por exemplo. Neste trabalho, tais métricas serão utilizadas conjuntamente com a polaridade das notícias para a identificação de viés em notícias.

² <http://www.manchetometro.com.br>

2 Objetivo

O objetivo do trabalho é propor um método para análise de viés em notícias sobre política na língua portuguesa, combinando métricas quantitativas com a classificação de polaridade para a análise de viés apresentada pelo meio produtor de notícia em relação a entidades políticas. Para atingir esse objetivo, três grandes tarefas foram realizadas: construção de um corpus anotado contendo notícias de diferentes produtores de notícias, realização de experimentos com técnicas de análise de sentimentos e, por fim, o desenvolvimento e aplicação de uma metodologia para identificação de viés.

Para que fosse possível aplicar a metodologia de análise de viés, era necessário um corpus com uma variedade mínima de produtores de notícias e que as notícias tivessem sido selecionadas de forma sistemática, mas nenhum corpus encontrado em língua portuguesa atendia a todos esses requisitos. Portanto, tornou-se necessário o desenvolvimento de um corpus, que foi construído a partir da extração de notícias do perfil do Twitter de cinco produtores de notícias distintos durante o período das eleições de 2014. Após a extração das notícias, o corpus foi anotado em relação a polaridade por quatro anotadores.

Utilizando o corpus anotado em relação a polaridade, experimentos para avaliar o desempenho de algoritmos de análise de sentimentos foram realizados. A avaliação de desempenho desses algoritmos é importante pois as técnicas de análise de sentimentos são uma ferramenta importante para a automatização da análise de viés e, ao serem aplicadas em diferentes domínios e idiomas, podem apresentar variações de desempenho. Nesse sentido, é interessante avaliar os algoritmos em língua portuguesa. O algoritmo proposto por (SCHOLZ; CONRAD, 2013), baseado na entropia da conexão das palavras e com bom desempenho em língua alemã, foi avaliado em comparação a algoritmos linha de base baseados em representação *bag of words*.

Por fim, uma metodologia para classificação de viés foi proposta com base no conceito de outliers, a partir da comparação da cobertura de cada produtor de notícia em relação a determinadas entidades políticas. Foram analisados três tipos de viés sobre os dados do corpus construído: o viés de seleção, que avalia o quanto uma entidade é referenciada pelo meio de comunicação; o viés de cobertura, que avalia quanto destaque é destinado a entidade e o viés de afirmação, que avalia se estão falando mal ou bem da entidade.

O trabalho está organizado da seguinte forma: no capítulo 3 é apresentado o levantamento bibliográfico realizado sobre o tema de análise de viés; no capítulo 4 é apresentada a metodologia utilizada para construção de corpus e para a anotação do mesmo; no capítulo 5 são apresentados os resultados dos experimentos realizados com técnicas de análise de sentimentos; no capítulo 6 é proposta a metodologia de análise de viés e apresentados os resultados de sua aplicação sobre o corpus construído; e no capítulo 7 está a conclusão do trabalho e discussão sobre trabalhos futuros.

3 Conceitos Fundamentais

Neste capítulo, será apresentado um levantamento bibliográfico de assuntos fundamentais para o entendimento da análise de viés em notícias: o conceito de viés e como as técnicas de Análise de Sentimentos se aplicam ao problema.

O levantamento bibliográfico foi realizado na forma de uma revisão sistemática executada em Março de 2014. A revisão foi realizada com o objetivo de identificar as técnicas utilizadas para análise de viés em notícias. Nesse sentido, a seguinte questão de pesquisa foi formulada:

Quais são os métodos existentes para análise de viés em notícias?

A *string* de busca com os termos relacionados com análise de viés em notícias foi definida após uma revisão exploratória sobre o tema. Os termos inicialmente utilizados foram *Opinion Mining*, *Point of View* e *Bias*. Na revisão exploratória, encontraram-se estudos pertinentes classificando o problema de análise de viés como análise de sentimentos; por isso, os termos definidos por (PANG; LEE, 2008) em um *survey* sobre o tema foram adicionados à *string* de busca: *Sentiment Analysis* e *Subjectivity Analysis*. O termo *Viewpoint* foi adicionado como uma variação do termo *Point of View*. A *string* de busca foi definida da seguinte forma:

```
("Opinion Mining" OR "Sentiment Analysis" OR "Subjective Analysis" OR
"Subjectivity Analysis" OR "Bias" OR "Point of View" OR "Viewpoint")
AND news
```

Essa *string* foi submetida a três bases científicas: ACM Digital Libray¹, IEEE Explore² e ACL Anthology³. Os artigos retornados que, a partir da leitura de seu título e resumo não foram considerados relacionados à pergunta de pesquisa, foram excluídos. Não foi utilizado nenhum critério para excluir artigos por tempo de publicação. Os critérios de inclusão e exclusão foram aplicados após a leitura do título e resumo.

Os critérios de inclusão definidos foram:

1. Trabalhos publicados e disponíveis integralmente nas bases de dados pesquisadas.

¹ <http://dl.acm.org>

² <http://ieeexplore.ieee.org/Xplore/home.jsp>

³ <https://aclweb.org/anthology/>

2. Artigos que foram submetidos a revisão por pares.
3. Trabalhos que lidam com análise de viés em textos.

Os critérios de exclusão definidos foram:

4. Trabalhos que não abordam o domínio de notícias.
5. Estudos secundários.

Os resultados da execução da revisão sistemática em cada base científica estão na Tabela 1. Nela estão descritos a quantidade de artigos retornados pela busca, quantos foram excluídos, quantos foram incluídos, quantos incluídos se repetiram entre as bases e quando as buscas foram executadas. Nenhuma restrição aos resultados foi aplicada. Todos os artigos retornados pelas ferramentas de busca de cada base foram incluídos nessa contagem.

Tabela 1 – Resultados das buscas por artigos em cada uma das bases

Base	Total	Incluídos	Excluídos	Repetidos	Data
ACM Digital Libray	56	26	30	7	08/03/2014
IEEE Explore	129	38	91	4	08/03/2014
ACL Anthology	159	36	124	3	12/03/2014

Fonte: Gabriel Arruda, 2015

3.1 Definindo viés em notícias

O viés da mídia é uma falha inerente ao processo de produção de notícias, abrangendo as etapas de captação, redação e edição (PARK et al., 2009). Em cada uma dessas etapas, é provável que a notícia esteja sujeita a uma avaliação subjetiva do produtor de conteúdo ou a fatores externos, como proprietários e anunciantes. Todas essas influências fazem com que haja diversos recortes distintos da realidade, em que os produtores de conteúdo selecionam diferentes tópicos, cobrem esses tópicos em diversos tons e apresentam esses tópicos de diferentes formas.

A questão de viés é abordada de forma ampla por (PARK et al., 2009), incluindo aspectos que não estão presentes no texto da notícia. (SAEZ-TRUMPER; CASTILLO; LALMAS, 2013), por sua vez, consideraram três categorias distintas de viés em notícias sobre política:

- Viés de seleção: a preferência por escolher assuntos de um determinado partido político, por exemplo.

- Viés de cobertura: atribuir maior atenção (espaço/tempo) para um determinado partido político, por exemplo.
- Viés de afirmação: a preferência por expressar opiniões favoráveis (ou desfavoráveis) em relação a determinado partido político, por exemplo.

A partir dessas definições, pode-se analisar o viés em diversos contextos. (MEHLER et al., 2006), por exemplo, trataram apenas do chamado viés de cobertura e seleção em que, analisando dados geográficos das notícias, identificaram quais assuntos são abordados com mais frequência em quais lugares nos EUA. A maioria dos trabalhos encontrados na revisão sistemática; 78% dentre os 81 que realizam análise de algum dos tipos de viés definido por (SAEZ-TRUMPER; CASTILLO; LALMAS, 2013); trabalham com análise do viés de afirmação aplicando técnicas de Análise de Sentimentos (*e.g.* (BALAHUR et al., 2009) e (CAI et al., 2008)). O único trabalho encontrado que aborda os três tipos de viés é o de (SAEZ-TRUMPER; CASTILLO; LALMAS, 2013) para notícias em língua inglesa.

3.2 Análise de sentimentos em notícias

A Análise de Sentimentos é um campo de estudo que analisa as opiniões, sentimentos, avaliações, atitudes e emoções em relação a determinadas entidades, como produtos, serviços, organizações, indivíduos, questões e eventos (LIU, 2012). Como exemplo de aplicações de Análise de Sentimentos, é possível citar (JURAFSKY, 2014):

- Filmes: esta resenha é positiva ou negativa?
- Produtos: o que as pessoas pensam do novo iPhone?
- Sentimento público: como está a confiança do consumidor? Está decaindo ou crescendo?
- Política: o que as pessoas pensam sobre o candidato ou questão?
- Predição: prever os resultados das eleições ou tendências de mercado a partir de sentimentos.

A pesquisa em Análise de Sentimentos se tornou uma área de grande interesse a partir do ano 2000 (PANG; LEE, 2008). Os motivos para esse recente interesse na área são uma combinação de fatores, como:

- A crescente utilização de métodos de aprendizado de máquina em processamento de língua natural e recuperação de informação.
- A disponibilidade de grandes conjuntos de dados para algoritmos de aprendizado de máquina serem treinados, devido ao crescimento da World Wide Web e, especificamente, ao desenvolvimento de sites agregadores de resenhas.
- Percepção dos desafios intelectuais trazidos pela área e aplicações comerciais.

Sendo uma área de interesse recente, a nomenclatura utilizada ainda não é muito bem definida. Assim, embora o termo Análise de Sentimentos seja utilizado nesse trabalho, há outros nomes associados à mesma área de pesquisa, como Mineração de Opinião (*Opinion Mining*), Análise de Subjetividade (*Subjective Analysis*) e Computação afetiva (*Affective Computing*).

Por fim, ainda que Análise de Sentimentos trate de uma gama de dimensões como, por exemplo, subjetividade (KIM; HOVY, 2006; ABDUL-MAGEED; KORAYEM; YOUSSEFAGHA, 2011) e múltiplas emoções (raiva, nojo, medo, alegria, tristeza e surpresa) (CHAUMARTIN, 2007), neste trabalho o objetivo é utilizá-la para classificar notícias conforme sua polaridade, definindo se determinados trechos de notícias são positivos, negativos ou neutros.

3.2.1 Identificação de polaridade em notícias

O viés presente no texto da notícia é um conceito que possui uma série de nuances. A maneira mais simples de lidar com esse problema é usando uma classificação binária: positiva ou negativa, como é comumente aplicada ao domínio de resenhas. (TURNERY, 2001), por exemplo, utilizou a ideia de um produto ser recomendado como sendo da classe positiva e não recomendado como sendo da classe negativa. No domínio de resenhas de filmes, (PANG; LEE; VAITHYANATHAN, 2002) utilizaram as estrelas atribuídas pelo autor da resenha para classifica-las entre positivas, negativas e neutras.

Para o modelo de notícias há várias interpretações para as classes positiva e negativa. Por exemplo, no modelo de anotação definido por (ABDUL-MAGEED; DIAB, 2012) foi solicitado que os anotadores distinguíssem frases ruins e boas de frases subjetivas. Segundo esse modelo, frases como (1) “Gaddafi matou centenas de pessoas” e (2) “Infelizmente, Gaddafi definitivamente matou centenas de pessoas”, seriam, respectivamente, tidas como uma notícia negativa transmitida de forma neutra, e uma notícia subjetivamente

negativa, por conter a expressão “Infelizmente”. Outro ponto colocado pelos autores para os anotadores é considerar a perspectiva da notícia. Um ataque a Israel, por exemplo, pode ser analisado de perspectivas distintas dependendo do local de origem do jornal e seu posicionamento em relação ao conflito.

(JANG; SHIN, 2010b), por sua vez, solicitaram que anotadores classificassem sentenças extraídas de notícias como positivas ou negativas. Para que uma sentença recebesse uma das classificações, antes deveria ter sido classificada como subjetiva; caso contrário, ela não era considerada no processo de anotação. Os autores consideraram notícias subjetivas como as que contém alguma opinião referente a algum objeto. Se o anotador entendesse que a sentença era subjetiva pelo “tom de voz”⁴, ele deveria classificá-la como subjetiva, mesmo que não houvesse termos concretos expressando sentimentos.

Já (CURRAN; KOPRINSKA, 2013) dividiram o processo de anotação de um conjunto de citações em duas partes. Primeiramente, os anotadores deveriam classificar cada citação sem considerar o contexto, ou seja, deveriam utilizar apenas a citação em si como evidência para opinião, não considerando a opinião *a priori* do locutor ou o restante da notícia. Depois, eles deveriam continuar ignorando a opinião prévia do locutor, mas considerar o conteúdo da notícia para realizar a classificação. Como resultado, foi obtida uma concordância ligeiramente maior nas citações classificadas com contexto, isto é, considerando o conteúdo da notícia.

Nesse caso, a classificação sem contexto obteve kappa de Fleiss de 0,43, enquanto a classificação com contexto obteve kappa de Fleiss de 0,45. O kappa de Fleiss é uma métrica que avalia concordância entre anotadores baseada na proporção de concordância, podendo ser definido como $k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$, onde $\bar{P}$ é a probabilidade de concordância observada e $\bar{P}_e$ probabilidade de concordância esperada aleatoriamente.

A despeito das diversas interpretações e nuances das classes “positiva” e “negativa” na classificação de notícias, as estratégias aplicadas para realizar essa classificação automaticamente não diferem muito na prática. Por isso, os trabalhos que realizam esse tipo de classificação serão analisados independentemente das interpretações dadas para as categorias.

⁴ Os autores não detalharam o conceito “tom de voz” em um texto escrito.

3.2.2 Modelos de representação de textos

A definição da forma de representação do texto é uma questão importante da tarefa de classificação de polaridade. Por exemplo, para que seja possível usar determinadas técnicas de aprendizado de máquina, é necessário representar o texto como um vetor de características. Como então fazer uma representação compatível com esse modelo e manter as informações relevantes de um texto? Nesta seção, serão apresentados os modelos de representação de textos encontrados na revisão bibliográfica.

3.2.2.1 *Bag of Words*

Uma solução simples para a representação de textos é o uso do conceito de *bag of words*. Nesse modelo, um texto é um conjunto de palavras independentes entre si. Dessa forma, é possível representar um texto como um vetor de características: cada dimensão do vetor de característica representa uma palavra do vocabulário do texto.

Para minimizar o problema de assumir independência entre as palavras, é possível utilizar uma representação baseada no modelo de n-gramas. Um n-grama é uma sequência de palavras: um 2-grama (mais comumente chamado um bigrama) é uma sequência de duas palavras como “Eu sou”, “Traga o”, e um 3-grama (mais comumente chamado um trigramma) é uma sequência de três palavras como “Eu sou brasileiro” ou “Traga o livro”. Nesse modelo, cada n-grama seria uma dimensão ao invés de cada palavra separadamente.

Em ambos os modelos, é possível considerar a frequência em que as palavras ocorrem ou apenas a presença delas no texto. O modelo booleano utilizado por (GAO et al., 2010) é definido da seguinte forma

$$w_{kj} = w(t_k, d_j) = \begin{cases} 0 & \text{se } \#(t_k, d_j) = 0 \\ 1 & \text{se } \#(t_k, d_j) > 0, \end{cases}$$

sendo $\#(t_k, d_j)$ a frequência de um termo t_k em um texto d_j , e w_{kj} o valor da dimensão k em um texto j .

No modelo que contabiliza frequência e não apenas presença, o vetor de características resultantes da contagem de palavras é normalizado entre 0 e 1,

$$w_{kj} = w(t_k, d_j) = \frac{\#(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} \#(t_s, d_j)^2}}$$

sendo $\#(t_k, d_j)$ a frequência de um termo t_k em um texto d_j , $|T|$ a quantidade de termos, e w_{kj} o valor da dimensão k em um texto j .

Um problema dessa abordagem é o custo de representar cada texto considerando todas as palavras do vocabulário, o que se torna ainda pior se forem utilizados modelos de n-gramas com $n > 1$, para mitigar esse problema, (SIERING, 2012) utilizou apenas as 500 palavras mais importantes, com base no ganho de informação de cada palavra, e contabilizou a frequência de cada palavra utilizando a medida tf-idf, ao invés de utilizar a simples contagem de palavras. O autor não detalhou como calculou o ganho de informação.

A medida tf-idf é uma forma que procura identificar a relevância da palavra dentro de um *corpus*: se uma palavra rara é encontrada em um documento, ela é mais importante para identificá-lo que uma palavra comum aos demais textos (*e.g.* preposições e pronomes). Uma das formas de calcular esse termo é utilizando uma escala logarítmica (MANNING; SCHÜTZE, 1999),

$$w_{kj} = \begin{cases} (1 + \log(\#(t_k, d_j))) \log \frac{N}{df_k} & \text{se } \#(t_k, d_j) \geq 1 \\ 0 & \text{se } \#(t_k, d_j) = 0, \end{cases}$$

sendo $\#(t_k, d_j)$ a frequência de um termo t_k em um texto d_j , N a quantidade de documentos no *corpus* e df_k a quantidade de documentos do *corpus* que contém o termo k . A função logarítmica é utilizada para reduzir o tamanho do texto no cálculo dessa métrica.

A representação *bag of words* tem como vantagem ser bastante simples, podendo ser utilizada em vários idiomas. O problema é que esse modelo assume independência entre as palavras do texto, o que não é verdade, já que a forma como as palavras estão organizadas é de extrema importância para o entendimento de um texto.

3.2.2.2 Representação por árvores sintáticas

Um texto pode ser representado por uma estrutura sintática derivada de sua gramática, como ilustrado na figura 1, na qual pode ser visto um exemplo de árvore sintática para a frase “Mary brought a cat.”. As árvores sintáticas representam o relacionamento entre as palavras de uma sentença, sendo assim uma forma de representar o texto mais completa que o *bag of words*, que assume independência entre as palavras.

Figura 1 – Árvore sintática da frase “Mary brought a cat.”

Fonte: (MOSCHITTI, 2006)

Uma forma de aproveitar as informações trazidas por esse tipo de estrutura é extrair apenas estruturas do tipo SVO (Sujeito-Verbo-Objeto) como, por exemplo, “Obama(S) - Accuse (V) - Republicans (O)” (SUDHAHAR; LANSALL-WELFARE, 2012). Dessa forma, é possível identificar quem está falando sobre quem no texto.

Apesar de ser uma estrutura mais rica em informações, a representação de árvores sintáticas na forma de um vetor de características não é simples como a representação *bag of words*. Para representar as árvores sintáticas como um vetor de características, (MALO et al., 2013) utilizaram *tree kernels*. Nesse método, cada árvore é construída como um grupo de sub-árvores. Na figura 2 é possível ver uma árvore sintática criada para a frase “Jeff ate the apple”, em que a) representa a árvore completa e b) algumas sub-árvores extraídas da árvore representada em a). Cada fragmento de árvore distinto se torna uma dimensão do vetor de características, fazendo com que esse tipo de estrutura tenha alta dimensionalidade ao ser representado como vetor.

Figura 2 – Exemplo de sub-árvores sintáticas

Fonte: (COLLINS; DUFFY, 2002)

Para geração de árvores sintáticas, são necessárias ferramentas de análise sintática, como o Stanford Parser⁵ para a língua inglesa, por exemplo. Dependendo do idioma,

⁵ <http://nlp.stanford.edu/software/lex-parser.shtml>

pode não haver esse tipo de ferramenta disponível, dificultando a portabilidade dessa representação para outras línguas.

3.2.2.3 Representação por grafo

A estratégia de (SCHOLZ; CONRAD, 2013) é utilizar uma estrutura de grafos conectando as palavras para identificar se aquelas palavras em conjunto são negativas, neutras ou positivas. Por exemplo, as três frases abaixo foram classificadas da seguinte forma:

- This solves the crisis. (positivo)
- This solves the crisis slowly. (neutro)
- This intensifies the crisis. (negativo)

No modelo do grafo, as arestas ε_{ij} conectam as palavras encontradas no conjunto de treinamento, contabilizando as co-ocorrências de palavras em sentenças de cada tipo. Temos então que

$$\varepsilon_{ij} = (y_{ij\pi}, y_{ijo}, y_{ijv}),$$

sendo $y_{ij\pi}$ o número de co-ocorrências do nó v_i e v_j em sentenças positivas. Analogamente, y_{ijo} representa a quantidade de co-ocorrência em sentenças neutras e y_{ijv} em sentenças negativas. O grafo gerado pelas três sentenças acima está representado na figura 3.

Figura 3 – Exemplo de grafo de relação entre as palavras

Fonte: (SCHOLZ; CONRAD, 2013)

A representação por grafo não assume independência entre os dados como o modelo *bag of words* e também não exige ferramentas de *parser* como a representação de árvores sintáticas. A partir dessa estrutura, métricas de entropia podem ser extraídas, como será discutido no modelo de classificação baseado em entropia (seção 3.2.4.2.2).

3.2.2.4 Representação por *Frames* semânticos

Os *frames* semânticos são representações esquemáticas das estruturas e padrões de crenças, práticas, instituições, imagens, etc. que promovem as bases para uma relação significativa em uma comunidade comunicativa (FILLMORE; JOHNSON; PETRUCK, 2003).

Dois exemplos de *frames* estão na tabela 2, um para o ato de compra e outro para acusação. *Target* são as palavras que identificam o *frame*. Para o *frame* Commerce_buy por exemplo, são considerados os verbos que remetem à ação de comprar algo. *Frame Element*, por sua vez, descreve os elementos que compõem o *frame*. No caso do Commerce_buy, os elementos são um comprador, um vendedor e bens.

Tabela 2 – Exemplo de *frames* semânticos

<i>Frame</i>	Judgement_comm	Commerce_buy
Target	accuse	buy
	sue	purchase
	charge	bid
Frame Element	COMMUNICATOR	BUYER
	EVALUEE	SELLER
	REASON	GOODS

Fonte: (XIE; PASSONNEAU, 2013)

(XIE; PASSONNEAU, 2013) utilizam *frames* semânticos para representar os textos, utilizando a ferramenta SEMAFOR⁶ para identificá-los. Para tornar os *frames* semânticos compatíveis com os algoritmos de aprendizado de máquina, os autores propuseram uma abordagem denominada SemTree para representação dos *frames* semânticos como árvores.

Na figura 4 é exemplificada a construção de uma árvore usando o SemTree sobre a sentença “Oracle sued Google in August 2010, saying Google’s Android mobile operating system infringes its copyrights and patents for the Java programming language”. Foram encontrados dois *frames* nessa sentença, um correspondendo ao verbo *sue* (1) e outro correspondendo ao verbo *say* (2). Construindo a árvore a partir da entidade *Oracle* (ORCL), que é parte constituinte de ambos os *frames* encontrados, a árvore é percorrida de forma reversa como em (3) e (4). Unificando os dois caminhos percorridos de forma reversa, tem-se com resultado a árvore (5).

Assim como as árvores sintáticas de (MALO et al., 2013), é possível representar uma árvore de *frames* semânticos como um conjunto de fragmentos de árvores. Dessa forma,

⁶ <http://www.ark.cs.cmu.edu/SEMAFOR/>

representações geradas com o SemTree podem ser utilizadas em técnicas baseadas em aprendizado de máquina.

Figura 4 – Árvores para representação de *frames* semânticos

Fonte: (XIE; PASSONNEAU, 2013)

3.2.3 Classificação baseada em dicionários

Uma forma para classificar as notícias entre positivas ou negativas é utilizando a polaridade das palavras. A partir da polaridade das palavras contidas na notícia, ela é classificada nas classes {positiva, neutra ou negativa} ou, alternativamente, {positiva, negativa}. Nesta seção, serão discutidos modelos para definição da polaridade das palavras e, com a polaridade das palavras, calcular a polaridade de um texto.

3.2.3.1 Polaridade de palavras

As técnicas baseadas em polaridade de palavras precisam de um dicionário que as definam. Para tal, uma solução simples é utilizar os dicionários que contêm dados sobre os sentimentos associados às palavras para definir tais palavras. Entretanto, essa solução

não pode ser adotada para qualquer idioma, já que nem todos dispõem de dicionários específicos como WordNet Affect ⁷ e SentiWordNet ⁸, criados para língua inglesa.

Uma alternativa ao uso desses dicionários, que exige pouco trabalho manual, é definir palavras sementes e construir o dicionário através de sinônimos e antônimos (HAIDER; MEHROTRA, 2011). Além de facilitar a criação de dicionários para línguas diversas, essa abordagem possibilita a inclusão de palavras de um domínio específico. Em um estudo de polaridade para notícias do mercado financeiro, por exemplo, é possível adicionar os termos “regulação” e “governo” (ZHANG; AHMAD, 2012).

Ainda que a forma mais simples de classificar as palavras seja de forma binária, como positivas e negativas, pode-se também atribuir uma escala de positividade e negatividade. Isso é feito utilizando a definição da palavra na WordNet, sendo um índice de polaridade calculado a partir da contagem de palavras negativas e positivas utilizadas na definição da palavra.

Nesse método, apresentado por (CAI et al., 2008), a palavra “*wonderful*”, por exemplo, possui um valor 13 de positividade, enquanto a palavra “*amnesty*” possui apenas 1,25 de positividade. Isso ocorre porque o termo “*wonderful*” possui 13 palavras positivas em sua única definição, enquanto a palavra “*amnesty*” possui apenas 5 palavras positivas em 4 definições. Apesar de exigir recursos menos sofisticados, buscar relações de sinônimos, hipônimos ou hiperônimos ainda exige dicionários como a WordNet. Para o português brasileiro, é possível utilizar essa estratégia, já que há a Wordnet.BR (SILVA, 2006) disponível para uso, muito embora ela ainda não contenha relações de hiponímia e hiperonímia.

Caso não existam ferramentas disponíveis, ou se opte por não utilizá-las, há alternativas baseadas em estatísticas sobre corpus que são independentes de idioma, como o modelo proposto por (MALANDRAKIS et al., 2013). De acordo com esse modelo, para definir a valência de palavras entre $[-1, 1]$, os autores propuseram a seguinte fórmula

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^N a_i v(w_i) f(d(w_i, w_j))$$

sendo w_j a palavra a ser caracterizada, $w_1, \dots, w_n$ as palavras sementes, $v(w_i)$ a valência para a palavra semente w_i , a_i o peso correspondente à palavra w_i (estimado através do

⁷ <http://wndomains.fbk.eu/wnaffect.html>

⁸ <http://sentiwordnet.isti.cnr.it>

método dos quadrados mínimos), $d(w_i, w_j)$ uma medida de similaridade entre as palavras w_i e w_j e $f(\bullet)$ uma função simples da tabela 3.

Tabela 3 – Funções de similaridade

Nome	Fórmula
linear	$f(d(\bullet)) = d(\bullet)$
exp	$f(d(\bullet)) = e^{d(\bullet)}$
log	$f(d(\bullet)) = \log(d(\bullet))$
sqrt	$f(d(\bullet)) = \sqrt{d(\bullet)}$

Fonte: (MALANDRAKIS et al., 2013)

As palavras sementes são manualmente selecionadas e classificadas, como os exemplos da tabela 4. Determinadas palavras podem possuir uma grande variância de valência ou possuir muitos significados e classes gramaticais distintas, tornando-as menos confiáveis como indicadoras de valência. Por isso, há o parâmetro α_i com o intuito de definir a relevância das palavras sementes.

Tabela 4 – Palavras sementes com polaridade manualmente definida

Classe	Palavras
positiva	good, superior, positive, correct, fortunate, nice, excellent
negativa	bad, inferior, negative, wrong, unfortunate, nasty, poor

Fonte: (MALANDRAKIS et al., 2013)

Os autores propuseram diversas funções de similaridade $d(w_i, w_j)$, sendo algumas baseadas na co-ocorrência das palavras em sentenças e documentos. Para estimar os valores de co-ocorrência das palavras, é possível utilizar buscadores da internet. Assim, a quantidade de resultados da busca de dois termos em conjunto indica o nível de co-ocorrência entre as duas palavras.

Uma outra alternativa apresentada pelos autores é utilizar o contexto em que a palavra aparece para calcular a função $d(w_i, w_j)$ entre dois termos. Para cada termo, é construído um vetor de características com base em seus k vizinhos à direita e à esquerda: $[v_{K,L} \dots v_{2,L} v_{1,L}] w [v_{1,R} v_{2,R} \dots v_{K,D}]$. O vetor de características para um termo w é definido como $T_{w,H} = (t_{w,1}, t_{w,2} \dots t_{w,V})$, sendo o i -ésimo valor $t_{w,i}$ a contagem de ocorrências do termo v_i do vocabulário V , em uma janela de contexto de tamanho H .

Utilizando um vocabulário $V = \{X, Y, Z\}$, todos os vetores de características teriam tamanho $|V|$. Por exemplo, considerando uma janela $H = 1$ para as sentenças “X X Z Y”; “X Y Z X” e “X Y Y X” teríamos o seguinte vetor de característica para o termo Y : $T_{Y,1} = (1, 1, 2)$.

Utilizando esses vetores, é possível calcular a função $d(w_i, w_j)$ pela similaridade de cossenos:

$$S(w_1, w_2) = \frac{\sum_{i=1}^V t_{w_1,i} t_{w_2,i}}{\sqrt{\sum_{i=1}^V (t_{w_1,i})} \sqrt{\sum_{i=1}^V (t_{w_2,i})}}$$

Uma outra abordagem, proposta por (VEALE; HAO, 2012), também abre mão do uso de dicionários, exigindo apenas a identificação da polaridade de algumas palavras sementes definidas pelo usuário. Segundo essa abordagem, a partir da base de *n-grams* do Google (BRANTS; FRANZ, 2006), é construído um grafo conectando as palavras encontradas. Utilizando as conexões desse grafo, a polaridade das palavras é estimada com base nas palavras sementes previamente definidas. Se uma palavra estiver conectada com mais palavras positivas, ela terá uma “positividade” maior.

A classificação das palavras no dicionário, contudo, não precisa ser necessariamente em relação à polaridade. Além de classificar os termos como sendo positivos ou negativos, (JANG; SHIN, 2010b) construíram manualmente um dicionário para o coreano, no qual atribuíram às palavras outras características:

- Cínico: enfatiza palavras negativas.
- Intensificador: dá ênfase para as expressões subsequentes.
- Conjectura: a opinião da expressão não é factual, mas hipotética.
- Citação: a opinião dada se refere a um terceiro.

Por exemplo, quando uma palavra classificada como “Cínico” aparece, as palavras positivas ao redor tem a polaridade multiplicada por -1 , e as palavras negativas tem a polaridade multiplicada por 2 .

A construção desses catálogos de polaridade para as palavras deve considerar o domínio no qual serão utilizados e os recursos disponíveis para o idioma. Utilizar recursos genéricos como WordNet Affect, por exemplo, pode ser útil para o domínio de notícias que aborda uma grande variedade de temas, mas pode ser inviável para outros idiomas e sub-domínios, como notícias de ações. Por isso, a construção desses catálogos com base em um *corpus* é uma alternativa interessante.

3.2.3.2 Métodos de classificação

Há vários métodos para calcular a polaridade de um texto a partir da polaridade das palavras que o compõem. (CAI et al., 2008), por exemplo, utilizaram um dicionário de palavras com polaridades contínuas, conforme descrito na seção 3.2.3.1. Seu método para calcular a polaridade de um trecho de texto funciona da seguinte forma

$$10 * (P - N) / \text{Math.sqrt}(\text{snippet.length}()),$$

sendo P e N a soma dos índices que indicam a positividade e negatividade de cada palavra e $\text{snippet.length}()$ o tamanho total do trecho sobre o qual se está calculando a polaridade.

Um modelo diferente, para palavras de polaridade binária, é definido por (SIERING, 2012). Nesse modelo, o sentimento de uma notícia é definido como

$$\text{Sent} = \frac{\text{pos} - \text{neg}}{\text{pos} + \text{neg}},$$

sendo pos a quantidade de palavras de polaridade positiva presentes no texto e neg a quantidade de palavras de polaridade negativa. O autor optou por normalizar o valor de Sent com base nas 5 notícias anteriores referentes ao mesmo assunto (no contexto do artigo, as últimas 5 notícias referentes às ações de uma determinada empresa

$$\text{sent} = \frac{\text{Sent} - \mu_{\text{Sent}}}{\sigma_{\text{Sent}}},$$

sendo μ_{Sent} a média de Sent e σ_{Sent} o desvio padrão.

(BALAHUR et al., 2009), por sua vez, optaram por um modelo mais simples. Em seu modelo, a polaridade de uma sentença é definida de forma binária, sendo que a polaridade com maior quantidade de palavras categoriza o texto. As palavras consideradas altamente polares, de acordo com a intuição dos autores, possuem 4 vezes mais peso que as demais palavras.

O problema de simplesmente contabilizar as palavras é que a relação entre as palavras contidas no texto não é levada em consideração. Por isso, alguns autores incluíram técnicas mais rebuscadas, com o intuito de reduzir as limitações da representação *bag of words*.

Ainda que a contabilização de palavras seja uma alternativa interessante para substantivos, adjetivos e até mesmo alguns verbos, há termos que trocam a valência de um

texto, como advérbios de negação por exemplo, que não podem ser identificados utilizando apenas contagem de palavras. Para tratar esse problema, uma alternativa simples adotada por (RAINA, 2013) consiste em inverter a polaridade de uma sentença que possui um número ímpar de palavras expressando negação. Outra alternativa, proposta por (FU et al., 2008), é a criação um conjunto de advérbios de negação que invertem a classificação de polaridade quando estão presentes em uma sentença.

Com relação à ambiguidade de certas palavras, (NASUKAWA; YI, 2003) utilizam um etiquetador *Part of Speech* (POS) para resolver o problema de palavras ambíguas e facilitar a identificação das entidades. Nesse caso, a ambiguidade é removida por meio das etiquetas POS, que classificam a classe gramatical da palavra (*e.g.* artigo, advérbio e verbo). Assim, a palavra inglesa “*like*”, por exemplo, que possui dois significados distintos, pode ser definida como referência ao verbo “gostar” ou à conjunção “como”.

Outro ponto a ser analisado no domínio de notícias é que, diferentemente do domínio de resenhas por exemplo, uma notícia não necessariamente se refere a um tema específico. Se estamos preocupados em definir trechos referentes a uma determinada entidade, como o presidente dos EUA por exemplo, não é bom utilizar a polaridade de todo o texto da notícia, mas apenas os trechos referentes ao presidente. Para lidar com esse problema, (SUDHAHAR; LANSALL-WELFARE, 2012) utilizaram a representação por árvores sintáticas para extrair apenas estruturas identificadas como SVO, discutidas na seção 3.2.2.2, nas quais o sujeito ou o objeto da estrutura fossem figuras políticas.

(CHAUMARTIN, 2007) também utilizou outro tipo de estrutura: o grafo de dependência entre as palavras de uma manchete. Diferentemente de representar a estrutura das sentenças como árvores sintáticas, o grafo de dependências representa a sentença como relacionamentos uniformes de dependência (MARNEFFE; MANNING, 2008). Isto é, triplas de relação entre as palavras, como por exemplo “O sujeito de *distributes* é *Bell*” na sentença “*Bell, based in Los Angeles, makes and distribute eletronic, computer and builiding products*”. Na figura 5 é apresentada uma representação gráfica do grafo gerado para a sentença.

Figura 5 – Representação gráfica da árvore de dependências da sentença “*Bell, based in Los Angeles, makes and distribute eletronic, computer and building products.*”

Fonte: (MARNEFFE; MANNING, 2008)

Utilizando o termo que forma a raiz do grafo, é identificada a palavra considerada mais importante do texto e a polaridade associada é multiplicada por 6. O autor também utiliza essa estrutura para detectar contrastes e nuances entre “bom ” ou “ruim”. Padrões como, por exemplo, [substantivo $\Rightarrow$ sujeito $\Rightarrow$ verbo] ou [verbo $\Rightarrow$ objeto direto $\Rightarrow$ substantivo] contendo verbos que aumentam ou decrementam quantidades são identificados. Usando a valência do substantivo, essas estruturas permitem ao sistema identificar notícias muito boas (ex: “aumenta o poder cerebral”), ou notícias boas em que algo ruim perde importância (ex: “reduz risco”, “furacão enfraquece”).

O uso de técnicas mais sofisticadas de PLN aumentam a capacidade do sistema de lidar com as nuances da língua natural. Porém, quanto mais complexas, mais associadas ficam a um determinado idioma, seja pela falta de ferramentas em outros idiomas, para atribuir etiquetas POS ou criar árvores sintáticas, seja pelas peculiaridades da linguagem. (JANG; SHIN, 2010b), por exemplo, utilizaram diversas características da língua coreana para melhorar a análise de sentimentos. Por ser uma linguagem rica em morfologia, não é possível utilizar ou adaptar as técnicas para a língua portuguesa. Por fim, além dos problemas descritos, algumas técnicas são computacionalmente caras, o que pode torná-las inadequadas para determinados contextos, em que seja necessário calcular a polaridade rapidamente.

3.2.4 Classificação baseada em aprendizado de máquina

Uma outra forma de identificar polaridade em textos é o uso de aprendizado supervisionado, segundo o qual, a partir de exemplos previamente anotados, um classificador é treinado para rotular as notícias como {positivo,negativo} ou {positivo,neutro,negativo}.

Nesta seção, será apresentado os classificadores utilizados para essa tarefa, assim como os modelos de representação comumente utilizados.

3.2.4.1 Classificadores

Uma das questões para realizar classificação por aprendizado de máquina é escolher um classificador adequado para a tarefa. Dois classificadores foram identificados como escolhas predominantes entre os artigos encontrados na revisão sistemática: SVM e Naïve Bayes. O SVM (*Support Vector Machines*) foi encontrado em 13 dos 20 trabalhos que utilizaram técnicas de aprendizado de máquina, enquanto o Naïve Bayes em 9 dos 20 trabalhos. Outros classificadores foram utilizados, mas com menor frequência: k-NN (GAO et al., 2010), Máxima Entropia (KAYA; FIDAN; TOROSLU, 2012; FONG et al., 2013; YU; KÜBLER, 2011), métodos baseados em árvore de decisão (DRURY; DIAS; TORGO, 2011; FONG et al., 2013; KAYA; FIDAN; TOROSLU, 2012) e Deep Learning (YANAGIMOTO; SHIMADA; YOSHIMURA, 2013). A proporção de classificadores encontrados na revisão sistemática está na figura 6. Por conta de sua popularidade, apenas SVM e Naïve Bayes serão apresentados em detalhes no que segue.

Figura 6 – Proporção de classificadores utilizados

Fonte: Gabriel Arruda, 2015

3.2.4.1.1 SVM

SVM é uma técnica de aprendizado de máquina que procura solucionar o seguinte problema de otimização: definir um hiper-plano entre duas classes, maximizando a margem entre os pontos mais próximos pertencente a classes distintas – os pontos sobre a borda da margem são chamados vetores de suporte, e o centro da margem é o hiper-plano ótimo que separa as classes (MEYER; WIEN, 2014). Dado um conjunto de treinamento rotulado de pares (x_i, y_i) , $i = 1, \dots, l$ onde $x_i \in R^n$ e $y_i \in \{1, -1\}$, o SVM procura a solução para o seguinte problema de otimização:

$$\begin{aligned} &\underset{w, b, \xi}{\text{minimizar}} && \frac{1}{2} \|w\|^2 + C \left(\sum_{i=1}^N \xi_i \right) \\ &\text{com as restrições} && y_i(w^T \cdot x_i + b) \geq 1 - \xi_i \\ &&& \xi \geq 0, \forall i, \dots, n. \end{aligned}$$

onde w define o hiper-plano que separa as duas classes, e $y_i(\bullet)$ se o exemplo x_i está dentro da margem mínima de distância para o hiper-plano. Na figura 7 um exemplo de classificador SVM para dados linearmente separáveis.

Figura 7 – Exemplo de SVM para dados linearmente separáveis

Fonte: (MEYER; WIEN, 2014)

Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis como apresentado na figura 7 (LORENA; CARVALHO, 2007). Para relaxar as

restrições do SVM, permitindo exemplos dentro da margem, é utilizado o termo ξ nas restrições do problema de otimização. A constante C é utilizada em conjunto com a variável para definir o peso dos exemplos que ficam dentro das margens: quanto maior o valor de C , menor a distância entre as margens.

(JOACHIMS, 1998) cita diversas características que tornam o SVM um classificador bem adaptado para a categorização de textos na representação *bag of words*:

- **Alta dimensionalidade no vetor de entrada:** no treinamento é comum lidar com muitas dimensões ao se trabalhar com textos. A definição do SVM que propõe a redução do *over-fitting* torna esse classificador adequado para lidar com espaços de alta dimensionalidade.
- **Poucas características irrelevantes:** um caminho para evitar espaços de alta dimensionalidade é selecionar as características mais relevantes. Infelizmente, em categorização de textos há poucas características irrelevantes. (JOACHIMS, 1998) selecionou as características mais relevantes utilizando o ganho de informação para classificar notícias em categorias. Mesmo utilizando as características menos relevantes para o processo de classificação, os resultados ainda foram consideravelmente superiores quando comparados à classificação aleatória.
- **Vetores representando textos são esparsos:** para cada documento, o vetor de características correspondente contém poucas características diferentes de zero. (KIVINEN; WARMUTH; AUER, 1997) apresentaram um modelo teórico e evidências empíricas para o limiar de erros de algoritmos aditivos, como os SVMs, serem bem adaptados para problemas com conceitos densos e instâncias esparsas.
- **A maioria dos problemas de categorização de texto são linearmente separáveis:** a maioria das categorias do corpus Oshumed de medicina (HERSH et al., 1994) e do corpus Reuters-21578 (LEWIS, 1997) utilizados pelo autor se mostraram linearmente separáveis.

Para as representações baseadas em árvores, como apresentado nas seções 3.2.2.4 e 3.2.2.2, algumas características da representação *bag of words*, como grande quantidade de dimensões e predominância de vetores esparsos, continuam verdadeiras.

3.2.4.1.2 Naïve Bayes

O classificador Naïve Bayes (NB) é a forma mais simples de rede bayesiana, na qual todos os atributos são independentes dado o valor da classe. Mesmo partindo dessa premissa de independência, raramente verdadeira em problemas do mundo real, esse classificador tem um desempenho surpreendentemente competitivo em tarefas de classificação (ZHANG, 2004).

Considerando um problema de classificação binária, a variável C representa a classe ($C = +$ ou $C = -$), e um exemplo E é representado por um vetor de características $E = (x_1, x_2, \dots, x_n)$. Pela regra de Bayes, a probabilidade de um exemplo E pertencer a uma classe C é obtido fazendo-se

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)}.$$

E é então classificado como a classe $C = +$ apenas, e apenas se

$$f_b(E) = \frac{p(C = +|E)}{p(C = -|E)} \geq 1,$$

sendo $f_b(E)$ um classificador bayesiano. Assumindo que todos os atributos são independentes dado o valor classe, isto é,

$$p(E|c) = p(x_1, x_2, \dots, x_n|c) = \prod_{i=1}^n p(x_i|c),$$

o classificador resultante é dado por:

$$f_{nb}(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^n \frac{p(x_i|C = +)}{p(x_i|C = -)}.$$

Na figura 8 é mostrada a representação gráfica de um classificador NB, no qual apenas a variável da classe é dependente dos atributos.

Figura 8 – Exemplo de classificador Naïve Bayes

Fonte: (ZHANG, 2004)

Para trabalhar com dados discretos, como a contagem de ocorrência de palavras em um texto no modelo *bag of words*, os parâmetros $\hat{P}(c)$ e $\hat{P}(E|c)$, que representam

as estimativas dos valores de $P(c)$ e $P(E|c)$ respectivamente, podem ser calculados por máxima verosimilhança, que utiliza a frequência relativa do conjunto de treinamento para estimar o valor das variáveis (MANNING; RAGHAVAN; SCHUTZE, 2009). Para estimar as probabilidades *a priori* $\hat{P}(c)$

$$\hat{P}(c) = \frac{N_c}{N},$$

sendo N_c o número de documentos da classe c e N o total de documentos do conjunto de treinamento. A probabilidade *a posteriori* é calculada a partir da frequência relativa do termo x em documentos pertencentes à classe c :

$$\hat{P}(x_i|c) = \frac{N_{cx} + \alpha}{N_c + \alpha n}$$

sendo N_{cx} o número de vezes que o termo x_i aparece nos documentos pertencentes à classe c e N_c o total de termos que aparecem nos documentos da classe c . O termo $\alpha \geq 0$ serve como termo de suavização para prevenir probabilidade igual a zero.

Ao utilizar o Naïve Bayes para classificação de texto utilizando a representação *bag of words*, duas premissas do classificador se mostram problemáticas: as palavras que ocorrem no texto não são independentes entre si e a representação *bag of words* ignora a ordem das palavras no texto (seção 3.2.2.1).

Apesar do classificador Naïve Bayes ter problemas para realizar boas estimativas de probabilidade devido às simplificações, as decisões de classificação são surpreendentemente boas (MANNING; RAGHAVAN; SCHUTZE, 2009). Por exemplo, assumamos um documento d com probabilidade reais $P(c_1|d) = 0,6$ e $P(c_2|d) = 0,4$. Supondo que d contenha muitos termos que são indicadores positivos para c_1 e muitos termos negativos para c_2 , ao se realizar as estimativas do classificador, a estimativa $\hat{P}(c_1) \prod_{i=1}^{n_d} \hat{P}(x_i|c_1)$ será muito maior que a estimativa $\hat{P}(c_2) \prod_{i=1}^{n_d} \hat{P}(x_i|c_2)$ (Tabela 5).

Tabela 5 – Diferenças de estimação e probabilidade real

	c_1	c_2
probabilidade correta $P(c d)$	0,6	0,4
$\hat{P}(c) \prod_{i=1}^{n_d} \hat{P}(x_i c)$	0,00099	0,00001

Fonte: (MANNING; RAGHAVAN; SCHUTZE, 2009)

A classe vencedora em uma classificação normalmente tem uma probabilidade muito maior que as demais classes e as estimativas divergem bastante das probabilidades reais.

Contudo, uma vez que a decisão da classificação é baseada em qual classe tem a maior estimativa, mesmo com estimativas ruins, os classificadores NB normalmente fazem boas classificações.

Por fim, assim como o classificador SVM, o Naïve Bayes tem algumas características interessantes para classificação de textos (MANNING; RAGHAVAN; SCHUTZE, 2009): ele se sobressai em problemas com muitas características relevantes e é robusto em relação a característica ruidosas.

3.2.4.2 Métodos de classificação

Um método de classificação é uma combinação de um classificador (Seção 3.2.4.1) com um modelo de representação (Seção 3.2.2). O método mais simples consiste em combinar a representação *bag of words* com um classificador, abordagem que pode ser aplicada a qualquer idioma, já que não necessita de nenhum recurso especial como, por exemplo, *parsers* ou dicionários de polaridade. Nessa seção, serão apresentadas outras abordagens que combinam aprendizado de máquina com outras formas de representação.

3.2.4.2.1 Classificação por árvore sintática de polaridade

O modelo de árvore sintática para representar palavras (Seção 3.2.2.2) foi utilizado por (MALO et al., 2013) em conjunto com um classificador SVM. As árvores sintáticas utilizadas nesse modelo contém a polaridade da palavra nas folhas e não a palavra em si (Figura 9). A polaridade das palavras foi extraída de um dicionário contendo termos classificados manualmente como positivos e negativos para o mercado financeiro.

Figura 9 – Árvore sintática de polaridade

Fonte: (MALO et al., 2013)

Para avaliar os resultados, os autores utilizaram as métricas precisão, abrangência (tradução para o termo *recall*), acurácia e a medida F⁹. Para comparar o desempenho desse modelo, os autores usaram como linha de base um método simples, baseado em contagem de palavras positivas e negativas. Segundo esse método, se 2/3 das palavras da sentença forem negativas, a sentença é classificada como negativa. Analogamente, se 2/3 das palavras contidas na sentença forem positivas, a sentença é classificada como positiva. Caso contrário, a sentença é classificada como neutra. Para o *corpus* de notícias do mercado financeiro, o modelo linha de base descrito foi superado em todas as métricas. Apenas na medida abrangência, para classificar sentenças neutras, o método linha de base superou os baseados em árvores léxicas.

O método, apesar de utilizar um dicionário de palavras pré-classificadas, também utiliza exemplos previamente anotados em um classificador SVM. Dessa forma, pode ser considerado um híbrido, já que usa ambas abordagens para resolver o problema: dicionário de polaridade e técnicas de aprendizado de máquina.

3.2.4.2.2 Classificação por entropia de conexão entre as palavras

O método proposto por (SCHOLZ; CONRAD, 2013) utilizou a representação do texto em grafo (Seção 3.2.2.3) para extrair métricas de entropia referentes à polaridade e subjetividade das palavras contidas no texto. As métricas extraídas foram submetidas a um classificador SVM. A figura 4 contém um exemplo de grafo construído sobre um pequeno conjunto de treinamento, com a sentença “There are structural factors behind the African growth story”, destacada com linhas sólidas representando os nós e arestas.

⁹ A precisão indica a proporção de sentenças classificadas corretamente em uma categoria. A abrangência indica, de todas as sentenças pertencentes a uma determinada classe, quantas foram classificadas corretamente. A acurácia indica a proporção de sentenças classificadas corretamente e a medida F é a média harmônica da precisão e abrangência. A definição completa das métricas acurácia, precisão e abrangência está na seção 5.3.

Figura 10 – Exemplo de grafo construído utilizado para cálculo das probabilidades condicionais

Fonte: (SCHOLZ; CONRAD, 2013)

A partir do grafo, construído com base no conjunto de treinamento, o algoritmo obtém a distribuição de probabilidade de novas sentenças para classifica-las conforme sua polaridade. Para uma determinada sentença, é construído um sub-grafo G_{sl} com todos os termos (vértices do grafo) encontrados e as arestas que os conectam. Usando esse sub-grafo, as seguintes probabilidades são calculadas para a sentença

$$\begin{aligned}
 P(pos|v_i, sub) &= \frac{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ij\pi}}{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ij\pi} + y_{ijv}} \\
 P(neg|v_i, sub) &= \frac{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ijv}}{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ij\pi} + y_{ijv}} \\
 P(sub|v_i) &= \frac{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ij\pi} + y_{ijv}}{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ij\pi} + y_{ijv} + y_{ijo}} \\
 P(neu|v_i) &= \frac{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ijo}}{\sum_{\varepsilon_{ij} \in G_{sl}} y_{ij\pi} + y_{ijv} + y_{ijo}}
 \end{aligned}$$

sendo ε_{ij} as arestas que conectam os termos i e j , $y_{ij\pi}$ a quantidade de sentenças positivas contendo os termos i e j , y_{ijo} a quantidade de sentenças neutras contendo os termos i e j , e y_{ijv} a quantidade de sentenças negativas contendo os termos i e j .

Para a classe (sub), é considerada a presença de sentenças afirmativas ($y_{ij\pi}$) e negativas (y_{ijv}). Os denominadores de polaridade se referem apenas às sentenças classificadas

como positivas e negativas, enquanto os denominadores de subjetividade se referem a todas as classificações.

Utilizando essas probabilidades condicionais, duas características são extraídas de cada sentença: polaridade ($f_{pol}(vi)$) e subjetividade ($f_{sub}(vi)$). O cálculo dessas características é feito utilizando a entropia de Shannon¹⁰ (SHANNON, 1948)

$$H(X) = - \sum_{i=1}^n p(x_i) \log_2(p(x_i)),$$

sendo $p(x_i)$ as probabilidades calculadas para polaridade e subjetividade. Essa métrica é normalizada entre 1 e -1 :

$$f_{pol}(v_i) = \begin{cases} 1 + P(pos|v_i) * \log_2(P(pos|v_i)) & \text{se } P(neg|v_i) \leq P(pos|v_i) \\ -1 - P(neg|v_i) * \log_2(P(neg|v_i)) & \text{caso contrário} \end{cases}$$

$$f_{sub}(v_i) = \begin{cases} 1 + P(sub|v_i) * \log_2(P(sub|v_i)) & \text{se } P(neu|v_i) \leq P(sub|v_i) \\ -1 - P(neu|v_i) * \log_2(P(neu|v_i)) & \text{caso contrário} \end{cases}$$

O cálculo dessas características é feito a partir da classe gramatical das palavras. Para cada sentença, são geradas quatro características conforme mostrado na tabela 6. As características obtidas são submetidas a um classificador SVM.

Tabela 6 – Características de polaridade e subjetividade

Características de polaridade	Características de subjetividade
$T_{v,pol}$: arestas com verbos	$T_{v,sub}$: arestas com verbos
$T_{s,pol}$: arestas com substantivos	$T_{s,sub}$: arestas com substantivos
$T_{adv,pol}$: arestas com advérbios	$T_{adv,sub}$: arestas com advérbios
$T_{adv,pol}$: arestas com adjetivos	$T_{adv,sub}$: arestas com adjetivos

Fonte: (SCHOLZ; CONRAD, 2013)

Para comparar os resultados, os autores utilizaram outras técnicas do estado da arte para a língua alemã. Foram utilizados os sistema Opinion Observer (DING; LIU; YU, 2008) e o SO-CAL(TABOADA; BROOKE; TOFILOSKI, 2011), ambos baseados em contagem

¹⁰ Os autores também utilizaram divergência de Kullback–Leibler, entretanto deram preferência à entropia de Shannon por ser uma métrica comutativa.

de palavras positivas e negativas. O sistema RSUNN ([SARVABHOTLA; PINGALI; VARMA, 2011](#)), baseado em aprendizado de máquina, também foi utilizado para a comparação. Todos os sistemas usados na comparação foram desenvolvidos para o domínio de resenhas.

A acurácia do modelo baseado em entropia superou a dos outros métodos, tanto quando testado com notícias do domínio de finanças quanto com notícias em geral. Para averiguar se a menor quantidade de dados de treinamento impactaria negativamente o processo de classificação, o algoritmo foi treinado com conjuntos de treinamento reduzidos e manteve bom desempenho comparado às demais técnicas.

O modelo de classificação baseado em grafos mitiga os problemas do modelo *bag of words* ao não assumir independência entre as palavras do texto, mantendo as vantagens de modelos mais simples. Nesse caso, o modelo exige apenas um etiquetador POS, não sendo necessária mais nenhuma ferramenta de PLN para que essa técnica seja aplicada.

A baixa dependência de idioma dessa técnica a torna bastante interessante para aplicação em língua portuguesa. Devido a essa característica e aos bons resultados obtidos pelos autores, essa técnica será utilizada nesse projeto para classificação de notícias em língua portuguesa.

3.2.5 Avaliação dos métodos de classificação

De modo geral, a avaliação dos classificadores é feita por meio de anotadores humanos (*e.g.* ([BALAHUR; STEINBERGER; KABADJOV, 2010](#)) e ([ZHANG et al., 2011](#))). Assim, dado o modelo de classificação e os algoritmos, os anotadores devem classificar os textos manualmente de acordo com a proposta do algoritmo. Para que esse método funcione, é necessário cuidado para que o modelo de anotação esteja claro, reduzindo assim a chance de erros por parte dos anotadores.

Se o algoritmo a ser avaliado for baseado em aprendizado de máquina, utiliza-se os exemplos anotados tanto para treinamento como para validação dos resultados, geralmente aplicando a técnica de validação cruzada (*e.g.* ([FONG et al., 2013](#)) e ([CARDEI; MANISOR; REBEDEA, 2013](#))). Para técnicas que não exigem conjunto de treinamento, todos os exemplos podem ser usados para validação dos resultados (*e.g.* ([BALAHUR; STEINBERGER; KABADJOV, 2010](#)) e ([CHAUMARTIN, 2007](#))).

Tal abordagem, contudo, sofre com a falta de concordância entre os anotadores. Por exemplo, ao classificar um conjunto de citações de forma binária, ([BALAHUR; STEINBERGER;](#)

KABADJOV, 2010) obtiveram menos de 50% de acordo entre os três anotadores designados para a tarefa. Após solicitar que os anotadores classifikassem as citações de acordo com o alvo, sem considerar se a notícia era boa ou ruim, a concordância chegou a 60%. Nesse caso, os autores perceberam que a delimitação de alguns aspectos tornava a tarefa muito mais clara. Dentre as instruções passadas, pediu-se que não fosse utilizado conhecimento prévio no julgamento. Ao final, foi obtido um consenso de 81% entre os classificadores. Outro fator determinante para o melhor desempenho dos métodos de classificação é a quantidade de anotadores. Nesse sentido, (DAS; BANDYOPADHYAY, 2010) perceberam que o aumento do número de anotadores reduz rapidamente o consenso entre eles.

Alternativamente ao uso de anotadores humanos, outra abordagem encontrada é o uso de fontes externas para classificação de notícias. Esta é a abordagem utilizada por (SIERING, 2012), que utilizou as flutuações do mercado para determinar a polaridade das notícias. Assim, se o preço das ações na bolsa aumentou após as notícias, então as notícias são classificadas como positivas; caso contrário, elas são classificadas como negativas. Apesar de isso resolver o problema de concordância entre os anotadores, esse tipo de abordagem tem outros problemas específicos. Nesse caso, é difícil determinar o tempo necessário para uma notícia produzir algum impacto mensurável na bolsa de valores, além de haver a possibilidade de confusão entre o impacto das notícias e outras variáveis externas que podem afetar o desempenho das ações na bolsa, mesmo que não estejam relacionadas com as notícias em si.

Além de definir a metodologia para anotação de notícias, outra questão relacionada é a granularidade com a qual as notícias serão anotadas. Considerando que notícias podem abordar múltiplos fatos e, conseqüentemente, possuir múltiplas polaridades, dividi-las em unidades menores de anotação pode auxiliar na análise individual desses fatos. Esta questão, entretanto, ainda é um problema em aberto, com abordagens indo da segmentação em sentenças (*e.g.* (BALAHUR; STEINBERGER; KABADJOV, 2010), (ABDUL-MAGEED; DIAB, 2012)) à extração de trechos, como citações de terceiros (*e.g.* (BALAHUR et al., 2009), (DRURY; ALMEIDA, 2012) e (CURRAN; KOPRINSKA, 2013)).

Com relação à métrica usada na avaliação, alguns trabalhos (*e.g.* (JANG; SHIN, 2010b)) consideram apenas o percentual de notícias classificadas como iguais. Entretanto, essa métrica tem o problema de não considerar a concordância casual entre os anotadores. Dessa forma, alguns trabalhos utilizam índices para concordância como *Kappa* de Cohen

(KIM; HOVY, 2005) e alfa de *Krippendorff* (DEVITT; AHMAD, 2007) ao invés do percentual de concordância entre os anotadores.

(PARK et al., 2012), por outro lado, optaram por uma análise qualitativa. Para testar o sistema, os autores definiram o seguinte critério para a efetividade do sistema NewsCube: a diversidade de artigos lidos pelos usuários em termos de aspecto e a influência do serviço na leitura de notícias balanceadas. Para avaliar esse critério de efetividade, os autores utilizaram três métodos distintos. Um dos métodos foi comparar quais notícias eram lidas utilizando outros sistemas para leitura de notícias como o Google News. Os outros dois métodos eram dois tipos diferentes de entrevista. A primeira entrevista procurava saber se o usuário acredita que o sistema NewsCube auxiliou a encontrar perspectivas diferentes sobre um determinado tópico. A segunda entrevista questionou sobre alguns tópicos específicos nos quais um grupo deveria se informar pelo sistema NewsCube e o outro com o sistema Google News.

3.3 Viés não presente no texto

A classificação do texto da notícia é a abordagem mais direta para a identificação de viés, mas não é suficiente para cobrir todo o problema. Uma fonte de notícias pode omitir um assunto em sua cobertura, o que pode constituir um viés positivo ou negativo, sendo contudo imperceptível para um sistema que apenas analise seu conteúdo.

Através da análise de *tweets* disponíveis publicamente, é possível confrontar quais tópicos estão sendo abordados pela mídia tradicional e nas redes sociais. Constatar, por exemplo, a omissão dos jornais egípcios em relação à revolta ocorrida no país em 2011, ao mesmo tempo em que os usuários do Twitter tinham grande interesse pelo assunto (YOUNUS et al., 2012) é um forte indicativo desse viés por omissão. Também comparando mídias sociais com mídia tradicional, (LIN; BAGROW; LAZER, 2012) compararam o viés dos blogs e da mídia tradicional através da quantidade de referências aos parlamentares estadunidenses.

(SAEZ-TRUMPER; CASTILLO; LALMAS, 2013), por sua vez, abordaram o problema do viés a partir de várias perspectivas, com métricas baseadas no Twitter. Foram definidos três tipos de viés: o viés de seleção, o viés de confirmação e o viés de afirmação (Seção 3.1). O viés da seleção de conteúdo foi analisado comparando os assuntos abordados por cada meio de comunicação, utilizando o coeficiente de *Jaccard* (JACCARD, 1901). O viés

de cobertura foi analisado a partir de contagem de referências a um determinado assunto e do tamanho das notícias referentes ao assunto. Por fim, o viés de afirmação foi analisado através da polaridade dos tweets e notícias, utilizando um modelo simples baseado em contagem de palavras positivas e negativas.

Lidando com o mesmo problema do viés de seleção, alguns trabalhos utilizaram o sistema Lydia (LLOYD; KECHAGIAS; SKIENA, 2005), que procura construir um modelo relacional de pessoas e lugares através de processamento de língua natural de fontes de notícias e análise estatística da frequência e co-localização das entidades. Através do Lydia, (WARD; BAUTIN; SKIENA, 2009) e (MEHLER et al., 2006) identificaram viés regional no tratamento de entidades. (WARD; BAUTIN; SKIENA, 2009), por sua vez, identificaram que os hispânicos geram grande quantidade de notícias locais no sudoeste dos EUA e na Flórida, regiões com maior número de imigrantes. Ao buscar por entidades específicas, (MEHLER et al., 2006) identificaram, por exemplo, que o jogador de basquete Steve Nash é mais comumente noticiado nas cidades dos times em que jogou.

Os trabalhos descritos nessa seção optaram por estratégias para identificação de viés que analisam o problema além da tarefa de classificação do texto em polaridade. Assim como (SAEZ-TRUMPER; CASTILLO; LALMAS, 2013) optou por combinar o uso de classificação em polaridade com outras métricas para análise do problema de viés, outros autores também optaram por uma abordagem híbrida para solucionar outros problemas. Ou seja, também utilizaram a classificação de polaridade em conjunto com outros fatores.

(ZHANG; SKIENA, 2009), por exemplo, utilizaram as informações de polaridade e subjetividade do Lydia de notícias sobre filmes para prever os valores de bilheteria dos mesmos. Além dos dados de notícias, foram utilizados outros dados sobre o filme, como país de origem, data de lançamento e orçamento. Através de modelos de regressão e um classificador k -NN, tanto os dados das notícias como os dados sobre o filme foram processados com o intuito de calcular o valor de bilheteria do filme, ou achar o filme com bilheteria mais similar. A combinação de informações das notícias com informações do filme obteve os melhores resultados, superando os métodos tradicionais que utilizam apenas dados sobre os filmes.

De forma similar, (DENG et al., 2011) utilizaram uma abordagem mista para prever o valor das ações na bolsa, combinando a polaridade das notícias com outros indicadores econômicos. Assim como para predição de bilheteria dos filmes, os melhores resulta-

dos também foram obtidos utilizando uma combinação de indicadores econômicos com informações de polaridade.

Por fim, (SCHARL; HERRING, 2013) também utilizaram diversos dados para analisar a questão climática, optando, contudo, por disponibilizar os resultados para análise do usuário ao invés de incorporar ao processo de Análise de Sentimentos. Nesse sentido, os autores disponibilizaram diversas informações como, por exemplo, a origem das notícias, palavras chaves mais utilizadas e tópicos em destaque, em um *dashboard* para análise pelo usuário.

4 Criação do corpus

A nossa proposta de análise de viés exige uma abordagem multidimensional das notícias: identificar o que está sendo dito, se está sendo dito, o quanto está sendo dito e sobre quem. Para ajudar a atender a essa necessidade, foi criado um corpus com um modelo de anotação bi-dimensional, que considera tanto a entidade sobre a qual se está comentando quanto a polaridade do texto (ARRUDA; ROMAN; MONTEIRO, 2015). Esse corpus foi desenvolvido com notícias sobre política, devido à natureza polarizada do embate político entre situação e oposição, sendo assim um tópico promissor para o estudo de viés.

Corpora de notícias anotadas já existem em alguns idiomas, como árabe (ABDULMAGEED; DIAB, 2012), português (ROCHA; SANTOS, 2000; ALEIXO; PARDO, 2008) e inglês (CURRAN; KOPRINSKA, 2013) por exemplo. Estes, entretanto, são gerais, não focando em um assunto específico, como notícias de política. Com esse intuito, apenas para a língua alemã parece haver um corpus dedicado a notícias de política (*cf.* (LI et al., 2008)).

Devido à necessidade de cobrir notícias políticas em português brasileiro de uma grande variedade de produtores de notícias (de modo a permitir uma comparação entre eles), foi necessário construir um novo corpus. Nesse sentido, projetos existentes, como o CSTNews (CARDOSO et al., 2011), CHAVE (ROCHA; SANTOS, 2000) e TeMário (PARDO; RINO, 2003), por exemplo, apesar de importantes, não atendem perfeitamente às necessidades do projeto, seja pela pequena quantidade de notícias de política ou porque estão restritos a poucos produtores de notícias brasileiros. O corpus CSTNews, por exemplo, contém notícias de quatro jornais brasileiros, sendo eles Folha de São Paulo, Estadão, Gazeta do Povo e Jornal do Brasil. Apesar de dispor de vários produtores de notícias, apenas 27 notícias do corpus tratam de política. O corpus CHAVE, por sua vez, contém uma grande quantidade de notícias, totalizando 726 edições do jornal português Público e 730 edições do jornal brasileiro Folha de São Paulo. Entretanto, todo o conteúdo vem de apenas dois produtores de notícias e um deles é português. Por fim, o TeMário, assim como o CHAVE, contém notícias provenientes de apenas dois produtos distintos: Folha de São Paulo e Jornal do Brasil.

4.1 Coleta de notícias e metodologia de anotação

As notícias de política foram extraídas de um conjunto de perfis públicos do Twitter¹. A escolha dos perfis do Twitter foi majoritariamente guiada pela importância subjetiva do produtor de notícias, ou seja, sua popularidade percebida. Como tal, foi selecionado um grupo de cinco produtores: Folha de São Paulo, Estado de São Paulo, G1, Veja e Carta Capital. Folha de São Paulo e Estado de São Paulo foram escolhidos por serem os maiores jornais do estado de São Paulo, sendo também dois dos maiores do Brasil. O G1 é um dos maiores portais online de notícias. Por fim, Veja e Carta Capital foram escolhidas por serem revistas semanais populares, que geralmente apresentam perfis editoriais opostos.

Durante o período de 06/09/2014 até 12/09/2014, todo dia às 20:00, um *crawler* recuperava os últimos 20 tweets de cada um dos perfis selecionados². Após filtrar os retweets (*i.e.* a republicação de um tweet já publicado) e tweets sem links para notícias, aqueles com links eram acessados e então o conteúdo original da notícia era recuperado como publicado no site de origem.

As notícias recuperadas eram então selecionadas pelo autor deste documento de acordo com a sua relevância para o corpus. Notícias eram consideradas relevantes toda vez que se referiam a um dos três principais candidatos concorrendo à presidência (*i.e.* Dilma Rousseff, Aécio Neves e Marina Silva), ou a um dos três principais candidatos concorrendo ao governo do estado de São Paulo (*i.e.* Geraldo Alckmin, Paulo Skaf e Alexandre Padilha). Ao final desse processo, 131 notícias foram selecionadas para formar o corpus, correspondendo a um total de 1.447 parágrafos com 65.675 palavras³. A tabela 7 resume os resultados para cada perfil analisado, em termos do número de tweets recuperados e selecionados, junto com o total de retweets, enquanto o algoritmo 1 descreve o processo de coleta dos resultados.

¹ twitter.com

² Notícias do dia 09/09/2014 não puderam ser extraídas devido a um problema técnico no sistema de extração naquele dia

³ A contagem de palavras foi realizada utilizando a biblioteca de tokens do NLTK (<http://www.nltk.org/api/nltk.tokenize.html>) e os parágrafos foram contabilizados pela ferramenta de anotação UAM Corpus Tool (<http://www.wagsoft.com/CorpusTool/index.html>)

Tabela 7 – Perfis do Twitter selecionados

Perfil	Nome	Tweets selecionados	Tweets recuperados	Retweets
@EstadaoPolitica	Política Estadão	7	17	1
@g1politica	G1 - Política	25	118	2
@folha_poder	Folha Poder	64	120	0
@cartacapital	Carta Capital	14	114	42
@VEJA	VEJA	21	118	8

Fonte: Gabriel Arruda, 2015

Algoritmo 1 Algoritmo de coleta de notícias

```

1: dataInicial ← 06/09/2014
2: dataFinal ← 12/09/2014
3: for dataReferencia ← dataInicial to dataFinal do
4:   noticiasDia ← extrairNoticiasTwitter(dataReferencia)
5:   for all noticia in noticiasDia do
6:     if elegivel(noticia) then
7:       adicionaCorpus(noticia)

```

Fonte: Gabriel Arruda, 2015

As notícias selecionadas foram segmentadas em parágrafos e apresentadas para um conjunto de quatro anotadores (ver tabela 8 para informações de sexo, idade, área de conhecimento e nível educacional dos anotadores). Tal segmentação foi escolhida para que se pudesse fornecer maior contexto para os anotadores, em comparação à segmentação em sentenças, por exemplo, sem que haja troca de tópico, uma vez que, segundo o dicionário Michaelis, o parágrafo é “pequena parte ou seção de discurso, capítulo, etc. que forma sentido completo e independente.” (MICHAELIS, 2015).

Para cada parágrafo, os anotadores deveriam identificar uma entidade alvo, determinando a polaridade do parágrafo em relação àquela entidade. Sendo assim, um parágrafo deveria ser classificado como positivo se ele trouxesse uma percepção positiva em relação à entidade para o anotador. Caso a percepção fosse negativa, então o parágrafo deveria ser classificado como tal. Parágrafos neutros, por sua vez, seriam os informativos que não alteram a percepção do autor em relação à entidade.

Tabela 8 – Detalhe dos anotadores

ID	Idade	Sexo	Área	Nível educacional
1	24	Feminino	Biológicas	Superior incompleto
2	24	Masculino	Exatas	Superior completo
3	31	Masculino	Exatas	Pós-graduação completa (mestrado)
4	26	Masculino	Exatas	Pós-graduação incompleta (mestrado)

Fonte: Gabriel Arruda, 2015

Os anotadores foram instruídos para considerar apenas pessoas como candidatos para uma entidade alvo, desconsiderando assim outras possibilidades, como empresas e locais por exemplo. A definição de um alvo é de grande importância, pois dependendo da entidade, a polaridade do parágrafo pode se inverter, já que algumas notícias positivas para determinados candidatos são potencialmente negativas para outros.

Os anotadores também deviam ter em mente que as entidades alvo são o assunto principal do parágrafo, a despeito de qualquer outra pessoa citada. Portanto, se o parágrafo apresenta uma crítica de um candidato em relação a outro, a entidade alvo deve ser o candidato criticado (assunto principal) ao invés do que está fazendo a crítica. Outro ponto importante é que as entidades alvos não necessariamente precisam ser explicitamente citadas no parágrafo. É necessário apenas que o conteúdo do parágrafo seja relacionado com alguma entidade. Por fim, se o anotador não identificasse nenhuma entidade alvo, então o parágrafo deveria ser considerado não classificado.

4.2 Resultados e discussão

Para verificar a concordância entre os anotadores do processo, tanto na identificação de entidades quanto na classificação de polaridade, foram utilizados três indicadores de concordância, sendo eles:

- O α de Krippendorff, que se baseia na proporção entre discordância observada e discordância esperada, sendo descrito da forma $\alpha = 1 - \frac{D_o}{D_e}$, onde D_o é a discordância observada e D_e a discordância esperada aleatoriamente. Entre suas vantagens, é possível citar a capacidade de lidar com dados faltantes e com qualquer número de anotadores e categorias. Quanto maior o valor do indicador, maior a concordância entre os anotadores. Seu valor varia entre -1 e 1.

- O κ de Fleiss, que se baseia na proporção de concordância entre os anotadores, podendo ser definido como $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$, onde $\bar{P}$ é a probabilidade de concordância observada e $\bar{P}_e$ probabilidade de concordância esperada aleatoriamente. Assim como o alpha de Krippendorff, essa métrica também é compatível com múltiplas categorias e anotadores. Os valores possíveis variam de -1 a 1, valores maiores indicam maior concordância entre os anotadores.
- O percentual de concordância, que consiste em calcular o total de parágrafos que foram anotados com a mesma categoria por todos os anotadores.

Para o nosso modelo de anotação, que permite que os anotadores se abstenham de classificar parágrafos caso não considerem pertinente, o α de Krippendorff é o indicador mais adequado, já que lida com classificações faltantes. Os resultados para esses três indicadores foram calculados com a ajuda do AgreeCalc ([ALVARES; ROMAN, 2013](#)) – uma ferramenta para calcular os vários índices de concordância entre múltiplos anotadores. A tabela 9 resume os resultados para polaridade e entidade.

Tabela 9 – Concordância entre anotadores para polaridade e entidade alvo

Índice de concordância	Polaridade ₁	Polaridade ₂	Entidade
α de Krippendorff	0,37	0,50	0,67
κ de Fleiss	0,26	0,28	0,39
Percentual de concordância	31,78	40,05	60,31

Fonte: Gabriel Arruda, 2015

Nessa tabela, Polaridade₁ refere-se à polaridade tratada como uma dimensão independente. Isso, entretanto, raramente é o caso, pois divergências na entidade alvo podem gerar divergências na classificação de polaridade, já que a classificação se refere à entidade alvo. Por essa razão, a concordância também foi calculada apenas para aqueles parágrafos nos quais os anotadores concordaram sobre a entidade alvo (Polaridade₂ na tabela 9). Os parágrafos contendo classificações discordantes foram considerados como não classificados.

A partir da tabela 9, é possível perceber que a concordância entre anotadores para a entidade alvo do parágrafo foi maior que para a sua polaridade. A concordância geral, entretanto, poderia ser maior se as entidades elegíveis fossem restritas aos principais candidatos. No trecho abaixo, por exemplo, dois anotadores escolheram “Guido Mantega”

como entidade alvo, enquanto outros dois escolheram “Dilma Rousseff”. Dado o relacionamento entre as duas entidades, os anotadores poderiam ter concordado em classificar como entidade alvo “Dilma Rousseff”, caso essa restrição tivesse sido aplicada.

*A presidenta Dilma Rousseff confirmou nesta segunda-feira 8 que, se for reeleita, o ministro da Fazenda, Guido Mantega, não vai permanecer no cargo. De acordo com Dilma, o próprio ministro não deseja continuar em um eventual segundo mandato.*⁴

Em relação à polaridade, a concordância foi maior quando calculada sobre parágrafos em que os anotadores concordaram com a entidade alvo (Polaridade₂ na tabela 9), do que quando calculada com entidade e polaridade como dimensões independentes. Esse resultado era esperado, pela razão apontada anteriormente, de que essas duas dimensões são, de fato, dependentes.

Os resultados de concordância par-a-par na entidade alvo, isto é a concordância calculada para cada combinação de pares de anotadores, podem ser vistos na tabela 10. Nesta tabela, o valor de α entre os anotadores 1 e 2 é 0,64, entre os anotadores 3 e 4 é 0,71 e assim por diante. O valor médio entre todos os pares é 0,68. A tabela 11, por sua vez, apresenta os valores de concordância para a dimensão Polaridade₂.

Tabela 10 – Concordância por pares entre anotadores para a dimensão entidade alvo

Índice	Média	Anotadores	2	3	4
α de Krippendorff	0,68	1	0,64	0,61	0,69
		2	–	0,72	0,74
		3		–	0,71
κ de Fleiss	0,46	1	0,43	0,41	0,47
		2	–	0,46	0,53
		3		–	0,47
Percentual de concordância	74,83	1	71,31	68,07	74,38
		2	–	78,38	80,04
		3		–	76,78

Fonte: Gabriel Arruda, 2015

Analisando os valores de α de Krippendorff na tabela 10, é possível perceber que a diferença entre o par com menor concordância ($\alpha = 0,61$ entre os anotadores 1 e 3) e maior concordância ($\alpha = 0,74$ entre os anotadores 2 e 4) gira em torno de 21%, para a

⁴ <http://www.cartacapital.com.br/blogs/carta-nas-eleicoes/mantega-nao-continua-em-eventual-segundo-mandato-diz-dilma-3791.html>

Tabela 11 – Concordância por pares entre anotadores para a dimensão Polaridade₂

Índice	Média	Anotadores	2	3	4
α de Krippendorff	0,48	1	0,50	0,39	0,40
		2	–	0,49	0,57
		3		–	0,51
κ de Fleiss	0,34	1	0,36	0,35	0,32
		2	–	0,34	0,31
		3		–	0,34
Percentual de concordância	65,72	1	67,23	59,82	59,89
		2	–	67,33	71,77
		3		–	68,30

Fonte: Gabriel Arruda, 2015

entidade alvo. A polaridade, por sua vez, mostra uma diferença de 46% (Tabela 11), entre os pares com menor concordância ($\alpha = 0,39$ entre os anotadores 2 e 4) e maior ($\alpha = 0,57$ entre os anotadores 2 e 4). Estas diferenças estão de acordo com pesquisas atuais (*e.g.* (ROMAN et al., 2015)), que encontraram uma diferença em torno de 32% (e, algumas vezes, até maior), em concordância por pares.

Por fim, mesmo que os resultados de concordância entre anotadores pareçam baixos, as pesquisas atuais em classificação de polaridade mostram resultados de concordância par a par indo de 66% (CURRAN; KOPRINSKA, 2013) até 81% (BALAHUR; STEINBERGER; KABADJOV, 2010) para um conjunto de três anotadores (ambos lidando com citações de terceiros encontradas em notícias), e 71% (JANG; SHIN, 2010b) para dois anotadores lidando com sentenças extraídas de notícias. Com uma média de concordância por pares de 74,83% para a entidade alvo e 65,72% para polaridade, com 80% de máxima concordância por pares para identificação de entidades e 72% para polaridade (ver tabelas 10 e 11), nossos resultados não parecem fora de escala.

4.3 Padrão Ouro do Corpus

Além da anotação propriamente dita, um padrão ouro do corpus também foi criado, para que fosse possível utilizá-lo para treinamento e validação de algoritmos de análise de sentimentos. Para a criação do padrão ouro, primeiramente foi associada a entidade escolhida pela maioria dos anotadores para cada parágrafo (o que inclui a opção “sem alvo”, que seria a opção de deixar o parágrafo como não classificado). Os empates foram

resolvidos pelo autor deste documento⁵. Na sequência, a polaridade do parágrafo foi definida a partir da polaridade escolhida pela maioria dos anotadores que concordaram sobre a entidade alvo do parágrafo (como determinado na etapa anterior da construção do padrão ouro). Classificações de polaridade associadas a outros alvos não foram consideradas para a maioria e, conseqüentemente, se nenhuma entidade foi associada ao parágrafo na etapa anterior, nenhuma polaridade foi associada. Novamente, os empates foram resolvidos pelo autor⁶.

A tabela 12 mostra os resultados da distribuição de polaridade entre os cinco perfis de Twitter utilizados. Como pode ser visto, a classificação de notícias de acordo com a sua polaridade é dependente do produtor de notícias com uma alta significância estatística ($\chi^2 = 110,5687$, $p < 0,01$, com nível de significância de 0,95). Este resultado, por sua vez, pode ser um indicador de viés em algum desses meios de produção.

Tabela 12 – Padrão ouro do processo de anotação

Perfil	Classificação			
	Positivo	Neutro	Negativo	Não classificado
@EstadaoPolitica	12	8	18	3
@g1politica	68	100	50	136
@folha_poder	187	177	232	148
@cartacapital	20	29	27	49
@VEJA	23	27	64	69
Total	310	341	391	405

⁵ 201 parágrafos dos 1,447 (13,9%) tiveram a entidade definida a partir do voto do autor

⁶ 160 parágrafos dos 1,042 (14,4%) em que os anotadores concordaram com a entidade-alvo tiveram a polaridade definida a partir do voto do autor

5 Experimentos de Análise de Sentimentos

Um dos tipos de viés descritos por (D’ALESSIO; ALLEN, 2000) é o viés de afirmação, que ocorre quando a opinião dos membros da mídia influencia os textos da cobertura jornalística. No contexto deste projeto, o viés de afirmação consiste em analisar como a opinião de determinados meios de comunicação influencia a cobertura jornalística em relação aos candidatos.

As técnicas de Análise de Sentimentos são uma forma de automatizar a análise do viés de afirmação. Entretanto a decisão de qual técnica se utilizar não se resume a identificar qual possui melhor desempenho. Como observado por (KAYA; FIDAN; TOROSLU, 2012), a aplicação de uma técnica de Análise de Sentimentos não necessariamente mantém o mesmo desempenho ao ser transportada para outro idioma, como ocorreu para a língua turca. Por outro lado, à medida em que se utiliza técnicas específicas para um idioma, como realizado para o coreano (JANG; SHIN, 2010a), por exemplo, elas se tornam naturalmente menos portáteis para outros idiomas.

Nesse contexto, em que é necessário analisar o *trade-off* de aplicar técnicas mais sofisticadas mas menos portáteis, decidiu-se avaliar o desempenho das técnicas baseadas em aprendizado de máquina sobre o padrão ouro do corpus (Seção 4.3). Mais especificamente, avaliar o desempenho do classificador por entropia proposta por (SCHOLZ; CONRAD, 2013) (Seção 3.2.4.2.2) em comparação aos métodos linha de base (SVM e Naïve Bayes com representação *bag of words*). Em sua proposta original, o classificador por entropia apresentou diversas características desejáveis para lidar com o problema a ser tratado nesse projeto, como bom desempenho com poucos dados de treinamento, e uma solução para lidar com a dependência entre palavras (o que pode ser útil em domínios com textos mais complexos, como notícias) e baixa exigência de ferramentas de PLN, necessitando apenas de um classificador de etiquetas POS.

5.1 Classificação Naïve Bayes e SVM

Os classificadores SVM com *kernel* linear e Naïve Bayes têm qualidades adequadas para problemas de classificação de textos com representação *bag of words* (Seções 3.2.4.1.1 e 3.2.4.1.2), servindo como linha de base para comparação com o classificador por entropia. Além de variações na representação *bag of words* para os classificadores, também foi

utilizado *stemming* como técnica de pré-processamento para melhorar o desempenho desses classificadores.

5.1.1 Stemming

Stemming é o processo de converter variações de uma palavra em uma representação concisa e precisa. O objetivo do processo de *stemming* é fundir palavras que possuem um significado comum em uma única representação denominada *stem* (ALVARES; GARCIA; FERRAZ, 2005). Ao se utilizar *stems*, palavras que seriam dimensões separadas na representação *bag of words* serão fundidas em apenas uma, trazendo assim a vantagem de reduzir a quantidade de dimensões e, conseqüentemente, o problema da “maldição” da dimensionalidade¹.

Em busca exploratória, foram encontrados dois algoritmos desenvolvidos para *stemming* em língua portuguesa: STEMBR (ALVARES; GARCIA; FERRAZ, 2005) e RSLP Stemmer (ORENGO; HUYCK, 2001). Para comparar o desempenho desses algoritmos, é necessário avaliar os dois tipos de erro que podem ocorrer: o *overstemming*, quando uma palavra é reduzida a um *stem* menor do que o esperado, fundindo palavras que não têm o mesmo significado em um mesmo *stem*; e *understemming*, quando a palavra é reduzida a um *stem* maior que o esperado, deixando palavras de mesmo significado em *stems* separados.

Para comparar os algoritmos em relação ao erro médio, (PAICE, 1994) propôs um método – ERRT (Error Rate Relative to Truncation) – que compara as ocorrências de *understemming* e *overstemming* com um algoritmo trivial de truncamento que corta as k últimas letras da palavra. De acordo com essa métrica, o algoritmo RSLP obteve o melhor resultado em um corpus de textos jornalísticos (FLORES; MOREIRA; HEUSER, 2010). Dado o bom desempenho do RLSP Stemmer em um corpus de notícias, como o criado para este projeto, optou-se por utilizar essa técnica de *stemming*.

¹ A maldição da dimensionalidade é a dificuldade que existe em definir funções em ambientes de alta dimensão, que podem ser muito mais complexas que em ambientes de baixa dimensão (DUDA; HART; STORK, 2012)

5.1.2 Tipo de representação

O valor das características foi representado de três formas: presença das palavras, contagem das palavras e frequência relativa das palavras (TF-IDF). Para o SVM, foram utilizadas as três representações. Para o classificador Naïve Bayes, foram utilizadas apenas as representações de contagem e ocorrência de palavras, que são representações adequadas ao modelo de estimação multinominal.

A presença de palavras identificará as palavras de forma binária no vetor de características: se determinada palavra aparecer uma vez ou mais no parágrafo, terá o valor 1, caso contrário o valor será 0. A contagem das palavras representará quantas vezes cada palavra apareceu no parágrafo. Por fim, a frequência relativa utilizará a métrica TF-IDF (Seção 3.2.2.1) que atribui maior valor para as palavras mais relevantes dentro do corpus.

5.2 Classificação por entropia

O classificador por entropia utiliza um etiquetador POS para realizar os cálculos de entropia separadamente por classes gramaticais. Assim como nos classificadores baseados na representação *bag of words*, o *stemming* também foi aplicado após a classificação por classes gramaticais, já que o uso de *stems* pode colaborar para uma estimação mais precisa de entropia, unificando palavras distintas mas com significados similares.

5.2.1 Etiquetador POS

A técnica de classificação por entropia calcula os valores de entropia separadamente por classes gramaticais, ou seja, ao invés de calcular a entropia entre todas as palavras do parágrafo, são calculados 4 valores de entropia, um para cada uma das seguintes classes: verbos, substantivos, advérbios e adjetivos. As demais classes gramaticais são desconsideradas pelo algoritmo.

Para etiquetar as palavras do corpus, foi utilizado o algoritmo proposto por (FONSECA; ROSA, 2013) para a língua portuguesa, baseado no modelo de (COLLOBERT et al., 2011), no qual cada palavra é representada como um vetor de características e as classificações das etiquetas são realizadas utilizando uma rede neural.

O vetor de características de cada palavra é um vetor semântico, calculado utilizando o método Hyper-Space Analogue to Language (HAL) (LUND; BURGESS, 1996), que consiste na criação de um vetor de co-ocorrências para cada palavra do vocabulário. Uma palavra é representada por um vetor X em que cada dimensão representa uma palavra do vocabulário, o valor de cada dimensão x_i é a quantidade de vezes que a palavra i apareceu a uma distância k da palavra representada pelo vetor X . Dessa forma, palavras semanticamente similares possuem vetores de co-ocorrência similares entre si.

A rede neural utilizada para etiquetar uma palavra recebe de entrada, além do vetor da própria palavra, seus vizinhos definidos por uma janela de tamanho k . Na figura 11 estão representadas cada uma das janelas com $k = 1$ extraídas para as palavras da frase “Não sei.”. Para os casos em que não há uma palavra anterior ou posterior para preencher as posições, são utilizados dois vetores gerados aleatoriamente que representam o espaço em branco antes das palavras (*PADDING_LEFT*) e depois das palavras (*PADDING_RIGHT*). Para definir as etiquetas de uma sentença, a janela de cada uma de suas palavras deve ser analisada.

Figura 11 – Janela de tamanho 3 produzido por uma sentença

Fonte: (FONSECA; ROSA, 2013)

Além dos resultados da rede neural, outro recurso utilizado é a ordem em que as etiquetas comumente aparecem em uma frase como, por exemplo, ser comum um substantivo aparecer logo depois de um artigo. Essa relação pode ser representada como um Modelo Oculto de Markov. Os resultados gerados para cada palavra da sentença são armazenados em uma matriz de transição A , em que cada célula A_{ij} contém a probabilidade de um *token* etiquetado como i ser seguido por um outro com a tag j . Portanto, denotando os parâmetros da rede neural como θ e o valor dado pela rede neural para a etiqueta j no t -ésimo *token* como $f_{t,j}$, o valor dado para uma sentença x de tamanho T tendo uma sequência de etiquetas y é:

$$s(x, y, \theta) = \sum_{t=1}^T f_{t,y_t} + A_{y_{t-1}, y_t}$$

Dessa forma, ao invés de definir como etiqueta o resultado dado pela rede neural para uma palavra, é calculada a etiqueta mais adequada considerando tanto os resultados da rede neural quanto a probabilidade de ocorrência das etiquetas utilizando o algoritmo de Viterbi (FORNEY G.D., 1973).

O algoritmo proposto foi treinado e avaliado com o corpus Mac-Morpho (ALUÍSIO; PELIZZONI; MARCHI, 2003) que contém notícias do jornal Folha de São Paulo. Para palavras conhecidas, o algoritmo utilizando a representação por sentenças e informações de capitalização na representação vetorial obteve 96,48% de acurácia. Para palavras desconhecidas, a acurácia desse modelo ficou em 87,34%.

O conjunto de etiquetas do corpus Mac-Morpho é diferente do utilizado pelo classificador por entropia. Entretanto, a diferença trata apenas da granularidade em que as etiquetas do corpus Mac-Morpho estão divididas. Por isso, as seguintes etiquetas do corpus Mac-Morpho (FONSECA; ROSA; ALUÍSIO, 2015) foram mapeadas para um conjunto de etiquetas compatível com as 4 categorias do classificador:

- VERB (Verbo) e VAUX (verbo auxiliar) $\Rightarrow$ Verbos.
- N (substantivo), NPROP (nome próprio) e PROSUB (pronome nominal) $\Rightarrow$ Substantivos.
- ADV (advérbio) $\Rightarrow$ Advérbios.
- ADJ (adjetivo), PROADJ (pronome adjetivo) e PCP (Particípio) $\Rightarrow$ Adjetivos.

5.3 Resultados da classificação por polaridade

As técnicas de classificação por polaridade dos parágrafos foram validadas utilizando validação cruzada com $k = 10$. Para todos os testes realizados, foram usadas as mesmas 10 partições do conjunto de dados, variando assim apenas os algoritmos aplicados e suas configurações. Para cada uma das três polaridades, foram extraídas as seguintes métricas:

- acurácia, indicando a proporção de parágrafos classificados corretamente, definida como $\frac{VP+VN}{VP+VN+FP+FN}$, sendo VP os parágrafos pertencentes à polaridade e classificados como tal; FP os parágrafos não pertencentes à polaridade e classificados como pertencentes; VN os classificados como não pertencentes à polaridade e classificados como tal; e FN os parágrafos pertencentes à polaridade e classificados como não pertencentes;

- precisão, indicando, de todos os parágrafos classificados em uma determinada categoria, quantos pertenciam realmente a essa categoria, definida como $\frac{VP}{VP+FP}$;
- abrangência indicando, de todos os parágrafos pertencentes a uma determinada classe, quantos o classificador rotulou corretamente, definida como $\frac{VP}{VP+FN}$.

Para cada uma dessas métricas, foi calculada a média aritmética entre os valores obtidos por cada uma das três polaridades. Para os classificadores baseados na representação *bag of words*, foram realizados testes variando o tipo de representação utilizada (Contagem, Presença ou TF-IDF) e aplicação ou não do *stemming*² no conjunto de dados. Para comparação dos resultados entre os classificadores, foi utilizado o teste *t* de Student com *folds* pareados e intervalo de confiança de 95%.

Os resultados do classificador Naïve Bayes com e sem *stemming* estão na tabela 13. A partir desses resultados, é possível perceber que o *stemming* propicia um pequeno ganho de desempenho nos resultados de classificação. De forma geral, tanto o *stemming* quanto as diferentes representações do texto pouco influenciaram nos resultados de classificação. A diferença de acurácia média entre os resultados para configuração de melhor desempenho ($bayes_m$), em que foi aplicado *stemming* e utilizada representação binária, e pior desempenho ($bayes_p$), em que não foi aplicado *stemming* e utilizada a representação binária, é de apenas 1%, portanto a hipótese $bayes_m = bayes_p$ não pode ser rejeitada ($t = 0,927, p = 0,378$). As demais métricas também não apresentaram diferença de desempenho estatisticamente relevante entre as configurações de melhor e pior desempenho do Naïve Bayes ($t = 1,075$ e $p = 0,31$ para precisão e $t = 0,917$ e $p = 0,383$ para abrangência).

Tabela 13 – Resultados do classificador Naïve Bayes

Stemming	Representação	Acurácia	Precisão	Abrangência
Não	Contagem	0,6018	0,6023	0,6001
Não	Presença	0,5951	0,596	0,5944
Sim	Contagem	0,6027	0,6038	0,6019
Sim	Presença	0,6037	0,6052	0,6029

Fonte: Gabriel Arruda, 2015

Os resultados para o classificador SVM estão na tabela 14. Eles foram ligeiramente piores que os obtidos a partir do classificador Naïve Bayes. A diferença entre a acurácia média da melhor configuração do SVM (svm_m), em que foi aplicado *stemming* e utilizada

² O conjunto original sem *stemming* tem 3472 dimensões, após a aplicação do *stemming* esse valor é reduzido para 2136.

a representação TF-IDF, e a melhor configuração do Naïve Bayes, foi de apenas 3%. A hipótese $bayes_m = svm_m$ não pode ser rejeitada ($t = 1,124, p = 0,29$). Para as demais métricas, também não houve diferença de desempenho estatisticamente relevante entre a configuração de melhor desempenho do SVM e a configuração de melhor desempenho do Naïve Bayes ($t = 1,046$ e $p = 0,323$ para precisão e $t = 1,128$ e $p = 0,2314$ para abrangência). O melhor desempenho para o classificador SVM foi obtido utilizando a representação TF-IDF e aplicando *stemming*.

Tabela 14 – Resultados do classificador SVM

Stemming	Representação	Acurácia	Precisão	Abrangência
Não	TF-IDF	0,5778	0,58	0,5723
Não	Contagem	0,5414	0,5423	0,5382
Não	Presença	0,5509	0,5491	0,5474
Sim	TF-IDF	0,5863	0,5897	0,5832
Sim	Contagem	0,546	0,5459	0,5449
Sim	Presença	0,5431	0,5404	0,5426

Fonte: Gabriel Arruda, 2015

Os resultados do classificador por entropia (Tabela 15) foram inferiores aos classificadores usados como linha de base. O classificador por entropia $entropia_m$ teve desempenho consideravelmente abaixo do resultado obtido pela melhor configuração do classificador Naïve Bayes; a hipótese $entropia_m = bayes_m$ é rejeitada para $p = 0,01$, sendo $entropia_m$ a melhor configuração do classificador por entropia, sem aplicação de *stemming*, considerando as três medidas: acurácia ($t = -7,855, p = 2,562 \times 10^{-5}$), precisão ($t = -7,739, p = 2,882 \times 10^{-5}$) e abrangência ($t = -7,839, p = 2,602 \times 10^{-5}$). O desempenho do SVM também foi bastante superior ao classificador por entropia; a hipótese $entropia_m = SVM_m$ é rejeitada para $p = 0,05$ considerando as três medidas: acurácia ($t = -3,172, p = 0,011$), precisão ($t = -2,815, p = 0,02$) e abrangência ($t = -3,32, p = 0,009$).

Tabela 15 – Resultados do classificador por entropia

Stemming	Acurácia	Precisão	Abrangência
Sim	0,5077	0,521	0,4966
Não	0,523	0,5312	0,5143

Fonte: Gabriel Arruda, 2015

5.3.1 Alterando o classificador por entropia

Devido às características do problema, a hipótese inicial era de que o classificador baseado em entropia obteria um melhor desempenho em comparação aos métodos linha de base. Esperava-se que o classificador por entropia mantivesse um bom desempenho mesmo com um volume reduzido de dados de treinamento (SCHOLZ; CONRAD, 2013), o que é uma característica importante, já que nosso *corpus* também não possui um volume muito grande de dados. Além disso, a proposta de considerar a relação entre as palavras, e não trata-las como dimensões independentes, é uma característica interessante para o domínio de notícias, que aborda diversos tópicos distintos, tornando assim o contexto em que as palavras aparecem importante. Entretanto, mesmo com essas características, não foi obtido um bom desempenho em comparação com modelos mais simples.

Para identificar se palavras relevantes para o processo de classificação não estavam sendo desconsideradas, devido às dimensões segregadas por classe gramatical (Seção 5.2.1), foram utilizadas duas dimensões adicionais para calcular as métricas de subjetividade e polaridade para as palavras que não entraram em nenhuma das 4 classes (verbos, substantivos, advérbios e adjetivos). Ou seja, mais duas características, $T_{out,pol}$ e $T_{out,sub}$, foram adicionadas ao modelo original apresentado na tabela 6. O novo conjunto de características do modelo está na tabela 16.

Tabela 16 – Características de polaridade e subjetividade

Características de polaridade	Características de subjetividade
$T_{v,pol}$: arestas com verbos	$T_{v,sub}$: arestas com verbos
$T_{s,pol}$: arestas com substantivos	$T_{s,sub}$: arestas com substantivos
$T_{adv,pol}$: arestas com advérbios	$T_{adv,sub}$: arestas com advérbios
$T_{adv,pol}$: arestas com adjetivos	$T_{adv,sub}$: arestas com adjetivos
$T_{out,pol}$: arestas com demais classes	$T_{out,sub}$: aresta com demais classes

Fonte: Gabriel Arruda, 2015

Os resultados dessa adição estão na tabela 17 e não trouxeram melhoras; pelo contrário, piorou o desempenho do classificador. A adição das novas palavras ao modelo, como preposições (de, para) e artigos (o, um) por exemplo, antes desconsideradas por não estarem em nenhuma das quatro classes, influi no cálculo de entropia das demais

categorias: caso palavras de categorias distintas sejam encontradas em um mesmo parágrafo, o cálculo de entropia resultante é considerado para as duas dimensões. Nesse contexto, é importante perceber que a adição de informação pode apenas adicionar “ruído” ao modelo atrapalhando o processo de classificação. A hipótese $entropia_{de} = entropia_m$, sendo $entropia_{de}$ o classificador por entropia com as dimensões extras, teve os seguintes resultados para o teste t : acurácia ($t = -1,946, p = 0,0833$), precisão ($t = -1,963, p = 0,081$) e abrangência ($t = -1,6587, p = 0,131$)

Tabela 17 – Resultados do classificador por entropia com dimensões adicionais

Stemming	Acurácia	Precisão	Abrangência
Sim	0,4866	0,4989	0,4757
Não	0,4951	0,5001	0,4873

Fonte: Gabriel Arruda, 2015

Outra estratégia aplicada foi utilizar um diferente tipo de *kernel* para o classificador SVM aplicado sobre as dimensões de entropia. Dada a reduzida quantidade de dimensões desse modelo, diferentemente dos modelos de representação *bag of words* de alta dimensionalidade, as vantagens do *kernel* linear podem não se aplicar mais ao classificador por entropia (Seção 3.2.4.1.1). O uso de um *kernel* polinomial, entretanto, não trouxe melhores resultados como pode ser visto nas tabela 18. A hipótese $entropia_m = entropia_{plk}$, sendo $entropia_{plk}$ o classificador por entropia usando *kernel* polinomial de grau k , pode ser rejeitada para $p = 0,05^3$, considerando acurácia e os três graus de polinômios experimentados (2, 3 e 4), indicando que o uso de *kernel* polinomial não teve grande influência nos resultados obtidos.

Tabela 18 – Resultados do classificador utilizando *kernel* polinomial

Grau polinômio	Stemming	Acurácia	Precisão	Abrangência
2	Sim	0,477	0,475	0,4687
2	Não	0,4712	0,4804	0,4668
3	Sim	0,4712	0,4849	0,468
3	Não	0,4683	0,4983	0,4672
4	Sim	0,428	0,5021	0,4283
4	Não	0,4126	0,5119	0,4159

Fonte: Gabriel Arruda, 2015

³ A hipótese $entropia_{pl2} = entropia_m$ teve os seguintes resultados: $t = -2,769$ e $p = 0,0218$ para acurácia, $t = -2,144$ e $p = 0,0607$ para precisão e $t = -2,447$ e $p = 0,037$ para abrangência. A hipótese $entropia_{pl3} = entropia_m$ teve os seguintes resultados: $t = -2,985$ e $p = 0,0153$ para acurácia, $t = -1,365$ e $p = 0,2056$ para precisão e $t = -2,569$ e $p = 0,03$ para abrangência. A hipótese $entropia_{pl4} = entropia_m$ teve os seguintes resultados: $t = -5,3367$ e $p = 0,0005$ para acurácia, $t = -1,14$ e $p = 0,284$ para precisão e $t = -5,9$ e $p = 0,0002$ para abrangência.

5.3.2 Classificação condicionada à entidade

No processo de anotação do corpus, foi solicitado que os anotadores definissem a polaridade dos parágrafos em relação a determinada pessoa, devido à hipótese de que um texto negativo para um candidato pode não ser visto da mesma forma por um opositor. Ou seja, uma notícia não é positiva ou negativa *a priori*. Essa classificação depende de diferentes perspectivas. Nesse contexto, faz sentido criar um classificador de polaridade para cada entidade.

Para verificar o desempenho dos classificadores por entidade, foram escolhidas as duas entidades com mais parágrafos associados: Dilma Rousseff com 367 parágrafos e Marina Silva com 272 parágrafos. O conjunto completo tem 1042 parágrafos com entidade associada. Os resultados desses dois conjuntos reduzidos serão comparados aos resultados utilizando o conjunto completo de notícias. Cada classificador foi treinado e validado considerando apenas parágrafos referentes a uma mesma entidade. A validação desses conjuntos também foi realizada usando validação cruzada com $k = 10$. Para comparação dos resultados entre os conjuntos reduzidos de treinamento e o conjunto completo também foi utilizado o teste t de Student. Entretanto, não mais com os *folds* pareados já que os conjuntos de dados são distintos entre si.

Nas tabelas 19, 20 e 21 são apresentados os resultados de classificação para a entidade Dilma Rousseff. O desempenho obtido ao utilizar como corpus apenas notícias referentes à Dilma Rousseff foi superior em comparação ao conjunto completo de notícias, o classificador Naïve Bayes obteve uma acurácia média 13,64% melhor, o classificador SVM 12,50% melhor e o classificador baseado em entropia 14,6% melhor. Em comparação a configuração de melhor desempenho no conjunto completo de parágrafos.

Para o classificador Naïve Bayes, comparando a configuração de melhor desempenho utilizando todos os parágrafos ($bayes_m$) e a configuração de melhor desempenho para o conjunto contendo apenas os parágrafos referentes à candidata Dilma Rousseff ($bayes_{md}$), em que não foi aplicado *stemming* e utilizada representação por contagem, foram obtidos os seguintes resultados para a hipótese $bayes_{md} = bayes_m$: $t = 3,076$ e $p = 0,007$ para acurácia, $t = 2,3201$ e $p = 0,032$ para precisão e $t = 2,32$ e $p = 0,0323$ para abrangência, indicando um desempenho superior com o conjunto de dados reduzido em comparação com o conjunto de dados completo com $p = 0,05$ considerando todas as métricas.

Para o classificador SVM, comparando a configuração de melhor desempenho utilizando todos os parágrafos (SVM_m) e a configuração de melhor desempenho para o conjunto contendo apenas os parágrafos referentes à candidata Dilma Rousseff (SVM_{md}), em que não foi aplicado *stemming* e utilizada representação TF-IDF, foram obtidos os seguintes resultados para a hipótese $SVM_{md} = SVM_m$: $t = 3,523$ e $p = 0,003$ para acurácia, $t = 2,59$ e $p = 0,021$ para precisão e $t = 2,819$ e $p = 0,0128$ para abrangência, indicando um desempenho superior com o conjunto de dados reduzido em comparação com o conjunto de dados completo com $p = 0,05$ considerando todas as métricas.

Para o classificador por entropia, comparando a configuração de melhor desempenho utilizando todos os parágrafos ($entropia_m$) e a configuração de melhor desempenho para o conjunto contendo apenas os parágrafos referentes à candidata Dilma Rousseff ($entropia_{md}$), em que foi aplicado *stemming*, foram obtidos os seguintes resultados para a hipótese $entropia_{md} = entropia_m$: $t = 2,232$ e $p = 0,0389$ para acurácia, $t = 2,643$ e $p = 0,017$ para precisão e $t = 1,301$ e $p = 0,21$ para abrangência, indicando um desempenho superior com o conjunto de dados reduzido em comparação com o conjunto de dados completo com $p = 0,05$ considerando todas as métricas.

Os algoritmos SVM e Naïve Bayes novamente obtiveram desempenho bastante similares, o Naïve Bayes tendo acurácia média apenas 0,4% melhor que o SVM. A hipótese $bayes_{md} = SVM_{md}$ não pode ser rejeitada para as três medidas: $t = 0,25$ e $p = 0,808$ para acurácia, $t = 0,159$ e $p = 0,877$ para precisão e $t = 0,488$ e $p = 0,637$ para abrangência. O algoritmo de classificação por entropia novamente teve desempenho inferior comparado aos linha de base, a hipótese $bayes_{md} = entropia_{md}$ pode ser rejeitada para as três medidas: $t = 3,668$ e $p = 0,005$ para acurácia, $t = 2,691$ e $p = 0,025$ para precisão e $t = 5,297$ e $p = 0,0005$.

As configurações de melhor desempenho, entretanto, foram diferentes para os classificadores SVM e Naïve Bayes. O classificador Naïve Bayes obteve melhor resultado usando contagem e sem aplicação de *stemming* para o conjunto contendo apenas parágrafos da entidade Dilma Rousseff, enquanto para o conjunto completo foi utilizando *stemming* e a presença das palavras somente. Apesar de diferentes configurações alcançarem o melhor desempenho, a diferença entre as várias representações no conjunto reduzido é pouco relevante. Sendo $bayes_{spd}$ o classificador para o conjunto contendo apenas exemplos para a candidata Dilma Rousseff com aplicação de *stemming* e considerando apenas a presença das palavras, a hipótese $bayes_{md} = bayes_{spd}$ não pode ser rejeitada para três medidas:

$t = 0,32$ e $p = 0,756$ para acurácia, $t = 0,449$ e $p = 0,664$ para precisão e $t = 0,559$ e $p = 0,59$ para abrangência.

Para o classificador SVM, a configuração de melhor desempenho para o conjunto completo foi aplicando *stemming* e utilizando representação TF-IDF, enquanto para o conjunto reduzido foi sem aplicação de *stemming* e também utilizando representação TF-IDF. Novamente, apesar de diferentes configurações alcançarem o melhor desempenho, a diferença entre as várias representações no conjunto reduzido é pouco relevante. Sendo SVM_{std} o classificador para o conjunto contendo apenas exemplos para Dilma Rousseff com aplicação de *stemming* e utilizando a representação TF-IDF, a hipótese $SVM_{md} = SVM_{std}$ não pode ser rejeitada para três medidas: $t = 1,465$ e $p = 0,177$ para acurácia, $t = 1,489$ e $p = 0,171$ para precisão e $t = 1,378$ e $p = 0,201$ para abrangência.

Tabela 19 – Resultados do classificador Naïve Bayes (entidade Dilma Rousseff)

Stemming	Representação	Acurácia	Precisão	Abrangência
Não	Contagem	0,6839	0,6764	0,6681
Não	Presença	0,6621	0,6551	0,6532
Sim	Contagem	0,6785	0,6641	0,6545
Sim	Presença	0,665	0,6541	0,644
Sim	Contagem	0,6785	0,6641	0,6545

Fonte: Gabriel Arruda, 2015

Tabela 20 – Resultados do classificador SVM (entidade Dilma Rousseff)

Stemming	Representação	Acurácia	Precisão	Abrangência
Não	Contagem	0,6264	0,6145	0,6089
Não	Presença	0,6595	0,652	0,6553
Não	TF-IDF	0,6811	0,6739	0,6626
Sim	Contagem	0,5939	0,5704	0,5727
Sim	Presença	0,6351	0,6182	0,6258
Sim	TF-IDF	0,6596	0,6373	0,6353

Fonte: Gabriel Arruda, 2015

Tabela 21 – Resultados do classificador por entropia (entidade Dilma Rousseff)

Stemming	Acurácia	Precisão	Abrangência
Sim	0,5752	0,5998	0,5245
Não	0,5996	0,62	0,5527

Fonte: Gabriel Arruda, 2015

Nas tabelas 22, 23 e 24 estão os resultados de classificação para a entidade Marina Silva.. O classificador Naïve Bayes obteve uma acurácia 11,5% melhor, o classificador SVM 11,6% melhor e o classificador por entropia 2% melhor.

Para o classificador Naïve Bayes, comparando a configuração de melhor desempenho utilizando todos os parágrafos ($bayes_m$) e a configuração de melhor desempenho para o conjunto contendo apenas os parágrafos referentes à candidata Marina Silva ($bayes_{mm}$), em que não foi aplicado *stemming* e utilizada a representação binária, foram obtidos os seguintes resultados para a hipótese $bayes_{mm} = bayes_m$: $t = 1,4954$ e $p = 0,154$ para acurácia, $t = 1,096$ e $p = 0,2886$ para precisão e $t = 1,406$ e $p = 0,177$ para abrangência.

Para o classificador SVM, comparando a configuração de melhor desempenho utilizando todos os parágrafos (SVM_m) e a configuração de melhor desempenho para o conjunto contendo apenas os parágrafos referentes à candidata Marina Silva (SVM_{mm}), em que foi aplicado *stemming* e utilizada a representação TF-IDF, foram obtidos os seguintes resultados para a hipótese $SVM_{mm} = SVM_m$: $t = 2,013$ e $p = 0,065$ para acurácia, $t = 1,868$ e $p = 0,0836$ para precisão e $t = 1,636$ e $p = 0,127$ para abrangência.

Para o classificador por entropia, comparando a configuração de melhor desempenho utilizando todos os parágrafos ($entropia_m$) e a configuração de melhor desempenho para o conjunto contendo apenas os parágrafos referentes à candidata Marina Silva ($entropia_{mm}$), em que não foi aplicado *stemming*, foram obtidos os seguintes resultados para a hipótese $entropia_{mm} = entropia_m$: $t = 0,229$ e $p = 0,822$ para acurácia, $t = 0,193$ e $p = 0,85$ para precisão e $t = 0,0741$ e $p = 0,942$ para abrangência.

Tabela 22 – Resultados do classificador Naïve Bayes (entidade Marina Silva)

Stemming	Representação	Acurácia	Precisão	Abrangência
Não	Contagem	0,6585	0,6429	0,6466
Não	Presença	0,6733	0,6583	0,6506
Sim	Contagem	0,6471	0,6458	0,6373
Sim	Presença	0,6437	0,6318	0,6294

Fonte: Gabriel Arruda, 2015

Tabela 23 – Resultados do classificador SVM (entidade Marina Silva)

Stemming	Representação	Acurácia	Precisão	Abrangência
Não	TF-IDF	0,6467	0,6608	0,6424
Não	Contagem	0,6022	0,5949	0,6029
Não	Presença	0,5878	0,5891	0,5872
Sim	TF-IDF	0,6545	0,6544	0,6426
Sim	Contagem	0,6251	0,6117	0,6179
Sim	Presença	0,5996	0,5999	0,5995

Fonte: Gabriel Arruda, 2015

Tabela 24 – Resultados do classificador por entropia (entidade Marina Silva)

Stemming	Acurácia	Precisão	Abrangência
Sim	0,5079	0,5056	0,5
Não	0,5337	0,5224	0,5176

Fonte: Gabriel Arruda, 2015

Assim como para o conjunto reduzido para a entidade Dilma Rousseff, os algoritmos SVM e Naïve Bayes novamente obtiveram desempenho bastante similares, o Naïve Bayes tendo acurácia média apenas 2,8% melhor que o SVM. A hipótese $bayes_{mm} = SVM_{mm}$ não pode ser rejeitada para as três medidas: $t = 0,199$ e $p = 0,847$ para acurácia, $t = -0,422$ e $p = 0,683$ para precisão e $t = 0,133$ e $p = 0,897$ para abrangência. O algoritmo de classificação por entropia novamente teve desempenho inferior comparado aos linha de base; a hipótese $bayes_{mm} = entropia_{mm}$ pode ser rejeitada para as três medidas: $t = 2,945$ e $p = 0,016$ para acurácia, $t = 2,303$ e $p = 0,047$ para precisão e $t = 2,707$ e $p = 0,024$.

O classificador Naïve Bayes obteve melhor resultado usando presença e sem aplicação de *stemming* para o conjunto contendo apenas parágrafos da entidade Marina Silva, enquanto para o conjunto completo foi utilizado *stemming* e também considerada a presença das palavras somente. Apesar de diferentes configurações alcançarem o melhor desempenho, a diferença entre as várias representações no conjunto reduzido é pouco relevante. Sendo $bayes_{spm}$ o classificador para o conjunto contendo apenas exemplos para Marina Silva com aplicação de *stemming* e considerando a presença das palavras, a hipótese $bayes_{mm} = bayes_{spm}$ não pode ser rejeitada para três medidas: $t = 0,534$ e $p = 0,606$ para acurácia, $t = -0,124$ e $p = 0,904$ para precisão e $t = 0,444$ e $p = 0,668$ para abrangência.

5.4 Discussão

Nessa seção, foram estudados algoritmos de classificação de polaridade para análise automática de uma das dimensões de viés a ser analisada nesse trabalho – o viés de afirmação. A partir do corpus desenvolvido, contendo tanto as dimensões de polaridade quanto entidade alvo, foi possível avaliar o desempenho dos algoritmos de classificação e identificar formas de analisar o problema. Foram utilizados algoritmos de aprendizado de máquina, que podem ser aplicados facilmente em língua portuguesa sem grandes adaptações.

Os resultados dos classificadores linha de base, ao contrário do esperado, superaram o desempenho do classificador baseado em entropia. As alterações propostas ao modelo na seção 5.3.1 não trouxeram melhores resultados de classificação. Apesar das características do classificador parecerem adequadas à tarefa de classificação do *corpus* (Seção 3.2.4.2.2), soluções mais simples baseadas na representação *bag of words* obtiveram melhores resultados.

Os resultados de classificação restritos à entidade Dilma Rousseff se mostraram superiores à classificação de polaridade sem entidade definida, os resultados da entidade Marina Silva não apresentaram diferença estatisticamente relevante. Vale ressaltar que os resultados foram superiores para entidade Dilma Rousseff e similares para a entidade Marina Silva mesmo com conjunto de treinamento bastante reduzido: 367 parágrafos para Dilma Rousseff e 272 parágrafos para Marina Silva, enquanto o conjunto completo contém 1042 parágrafos.

Para as técnicas de análise de sentimentos, o domínio de notícias tem características que tornam a tarefa especialmente desafiadora como, por exemplo, a predominância de textos objetivos e a existência de múltiplos tópicos e assuntos em uma mesma notícia (BALAHUR et al., 2009). Nesse sentido, os melhores resultados para o corpus restrito à entidade Dilma Rousseff, mostram que a restrição de domínio é uma questão importante a ser debatido para a classificação de notícias.

Em nosso corpus, a definição de um alvo foi feita de forma manual pelos anotadores. Apesar de mais simples que a definição de polaridade, nem mesmo para anotadores humanos é simples definir o alvo de um parágrafo, já que houve alguma discordância no processo de anotação (Tabela 10). Como forma de mitigar essa dificuldade, é possível fazer uso de citações de terceiros contidas em notícias ao invés do corpo da notícia, como proposto por (BALAHUR et al., 2009) e (CURRAN; KOPRINSKA, 2013). As citações são uma alternativa interessante para trabalhos futuros por não partirem da premissa de ser um texto objetivo, como é o caso do texto jornalístico, e por geralmente possuírem uma fonte e um alvo bem claros.

Ao trabalhar com um corpus multi-dimensional, tendo tanto a perspectiva como a polaridade do texto, talvez uma abordagem com múltiplos classificadores treinados para cada perspectiva seja mais simples do que o uso de um único classificador que consiga classificar textos como negativos ou positivos independentemente de perspectiva. Os nossos resultados vão parcialmente de encontro a esse hipótese, para a entidade Dilma Rousseff

foram obtidos resultados de classificação superiores, enquanto para Marina Silva o resultado apenas se manteve.

6 Análise de viés em notícias

A proposta desse projeto é analisar o viés na mídia, utilizando a classificação bi-dimensional do *corpus* criado. O viés será analisado a partir das três dimensões descritas por (SAEZ-TRUMPER; CASTILLO; LALMAS, 2013):

- Viés de seleção: a preferência por escolher assuntos relacionados a um político em particular.
- Viés de cobertura: atribuir maior atenção (espaço/tempo) para determinado político.
- Viés de afirmação: a preferência por expressar opiniões favoráveis (ou desfavoráveis) em relação a determinado político.

Embora não seja possível definir *a priori* se um produtor de notícias é enviesado, é possível comparar a diferença entre os diversos produtores de notícia. As notícias dos cinco produtores de notícia foram selecionadas no mesmo período de tempo e a partir da mesma ferramenta de distribuição de notícias (a rede social Twitter), portanto é esperado que não haja grandes diferenças em relação às entidades abordadas e ao teor das notícias. Nesse contexto, para analisar a diferença entre os diferentes meios de comunicação, será utilizado o conceito de *outliers*: se a cobertura de algum produtor de notícia se mostrar demasiadamente diferente das demais, ela será classificada como enviesada.

Para essa análise, serão considerados os três principais candidatos à presidência (Dilma Rousseff, Aécio Neves e Marina Silva) e os três principais candidatos ao governo de São Paulo (Geraldo Alckmin, Alexandre Padilha e Paulo Skaf) nas eleições de 2014.

6.1 Identificação de *outliers*

O conceito de *outliers* não tem uma definição precisa amplamente aceita (HODGE; AUSTIN, 2004; COUSINEAU; CHARTIER, 2010). No contexto deste projeto, a definição de (GRUBBS, 1969) para *outliers* parece ser mais adequada, dada a sua similaridade ao problema de identificar produtores de notícias enviesados: um *outlier* é uma observação que aparenta um desvio nítido de outras observações da amostra na qual ocorre. Segundo (HODGE; AUSTIN, 2004), há três tipos de estratégias diferentes para identificação de *outliers*:

- **Tipo 1:** identificação de *outliers* sem conhecimento prévio dos dados. Esse tipo de estratégia é essencialmente análoga a técnicas de classificação não supervisionada.

Esse tipo de abordagem analisa os dados como uma distribuição estatística, encontra os pontos mais remotos e os classifica como potenciais *outliers*. O Tipo 1 assume que os erros ou falhas estarão separados dos dados “normais” e, portanto, aparecerão como *outliers*.

- **Tipo 2:** modela a normalidade e anormalidade. Esta abordagem é análoga à classificação supervisionada e exige dados previamente anotados, classificados como normais ou *outliers*. Dispondo de exemplos suficientes de *outliers* e casos comuns, é possível aplicar estratégias de classificação supervisionada, como redes neurais por exemplo.
- **Tipo 3:** modela apenas normalidade ou, em alguns casos, apenas anormalidade. Os autores geralmente definem essa técnica como detecção de falhas¹. Essa abordagem é análoga às técnicas de aprendizado semi-supervisionado, sendo necessários apenas dados classificados como normais para que o algoritmo consiga determinar *outliers* a partir de um limiar de normalidade. É uma estratégia interessante para sistemas dinâmicos que podem aprimorar o modelo a partir do surgimento de mais dados.

Neste trabalho, está sendo utilizado o conceito de *outlier* para definição de viés e não o contrário. Ou seja, não está definido o que é um meio de comunicação enviesado (*outlier*) e não enviesado; portanto a abordagem do tipo 1 é a mais adequada, já que não há definição *a priori* do que é um dado normal e anormal em nosso conjunto de dados.

Para as métricas univariadas em que os dados seguem uma distribuição normal, o método mais comum é o uso da medida z-score (COUSINEAU; CHARTIER, 2010), que consiste na quantidade de desvios-padrão que uma observação está distante da média. Por exemplo, ao excluir todas as observações que estão 4 desvios de distância da média, seriam eliminadas 3 observações de cada 100.000. Entretanto, o uso de média como indicador de tendência central traz três problemas para essa abordagem (LEYS et al., 2013), uma vez que:

- Assume uma distribuição normal com os *outliers* inclusos.
- A média e o desvio padrão são fortemente impactados pela presença de *outliers*.
- É improvável que esse método identifique *outliers* em pequenas amostras.

¹ Tradução para *novelty detection*

Uma alternativa é o uso da mediana como indicador de tendência central que, diferentemente da média, é pouco sensível à presença de *outliers*. O Desvio Absoluto Médio (DAM) (HAMPEL, 1974), baseado na mediana como indicador de tendência central, é vantajoso pela sua robustez em relação à presença de *outliers*, além de ser imune ao problema do tamanho da amostra, sendo dado por

$$DAM = b \sum M_i (|x_i - M_j(x_j)|)$$

onde x_j são as n observações iniciais, M_j a mediana da série original e M_i a mediana da série original subtraída de M_j . Em geral, $b = 1,4826$ é uma constante derivada da premissa de normalidade dos dados, a despeito da anormalidade induzida pela presença de *outliers*.

Para exemplificar, considere o cálculo de DAM para uma série de tamanho $n = 8$ com os valores $\{1, 3, 3, 6, 8, 10, 10, 1.000\}$. Cada termo da série original é subtraído da mediana ($M_j = 7$) e se torna a seguinte série de valores absolutos: $\{6, 4, 4, 1, 1, 3, 3, 993\}$. A mediana dessa nova série $M_i = 3,5$ multiplicada por $b = 1,4826$ é igual a DAM de 5,1891.

Após o cálculo da DAM, é necessário definir um critério de rejeição para definir uma observação como *outlier*. Esse tipo de decisão continua sendo invariavelmente subjetiva. (MILLER, 1991), por exemplo, propõe três valores: 3 desvios (bastante conservador), 2,5 desvios (moderadamente conservador) e 2 desvios (pouco conservador). Para a proposta do projeto, em que os *outliers* são interpretados como presença de viés, será utilizada uma medida pouco conservadora de 2 desvios.

6.2 Viés de seleção

O viés de seleção lida com o problema do *gatekeeping*: escritores e editores selecionam, a partir de um corpo de notícias potenciais, aquelas que serão apresentadas para o público e, por extensão, descartam notícias que não serão publicadas (D'ALESSIO; ALLEN, 2000). Para análise desse tipo de viés em nosso corpus será utilizada a quantidade de referências que cada produtor de notícia fez a cada candidato.

Na tabela 25 está a relação de candidatos por perfil do Twitter. Para cada candidato, conforme análise do corpus descrito no capítulo 4, foi calculado o percentual de parágrafos em que ele foi selecionado como entidade alvo por produtor de notícia.

Tabela 25 – Proporção de referências aos candidatos por perfil

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	14,63%	2,44%	0,0%	14,63%	12,2%	19,51%
@g1politica	1,69%	1,98%	4,52%	2,82%	23,45%	12,99%
@folha_poder	4,03%	3,76%	3,23%	8,33%	20,56%	20,56%
@cartacapital	5,6%	0,0%	0,0%	0,0%	23,2%	16,0%
@VEJA	0,0%	0,0%	0,0%	3,28%	16,94%	24,59%
Média	5,08%	1,64%	1,55%	5,36%	18,27%	22,53%

Como pode ser visto, os candidatos ao governo de São Paulo recebem menos atenção que os candidatos à presidência. Em média, Geraldo Alckmin foi selecionado como entidade em 5,4% dos parágrafos, Alexandre Padilha 1,6% e Paulo Skaf 1,6%. Os perfis de revistas semanais, @cartacapital e @VEJA, pouco abordam os candidatos ao governo, apenas a Carta Capital teve 5,6% de seus parágrafos associadas ao candidato Geraldo Alckmin. Os candidatos à presidência, por outro lado, receberam maior atenção: em média, Aécio Neves foi selecionado como entidade do parágrafo 5,6%, Dilma Rousseff 19,2% e Marina Silva 16% das vezes.

A partir da proporção de parágrafos associada a cada candidato por perfil apresentado na tabela 25, foi calculado o desvio de cada meio em relação à mediana de cada meio em relação à entidade. Por exemplo, para a candidata Marina Silva foram obtidas as seguintes proporções de referência dos perfis: {0,1951 , 0,1299 , 0,2056 0,16 , 0,2459}. Ordenando os valores, é obtida a seguinte sequência {0,1299 , 0,16 , 0,1951 , 0,2056 , 0,2459}. O valor da mediana é $M = 0,1951$; subtraindo M da sequência é obtida a seguinte sequência em valores absolutos: {0,065 , 0,035 , 0 , 0,01 , 0,051}. A sequência é ordenada novamente ({0 , 0,01 , 0,035 , 0,0507 , 0,065}) e o valor da DAM é obtido após multiplicar a mediana da nova série (0,035) pela constante $B = 1,4826$, resultando em $DAM = 0,052$. Por fim, basta calcular a quantidade de desvios com a equação $(x_i - M)/DAM$: {0 , -1,25 , 0,2 , -0,675 , 0,98}.

A quantidade de desvios dos perfis em relação à quantidade de referência aos candidatos está na tabela 26. Apenas a cobertura do @EstadaoPolitica ultrapassou o valor limite de 2 desvios para os candidatos Geraldo Alckmin e Aécio Neves. Em ambos os casos, o desvio foi positivo, indicando que, durante o período de extração de notícias, o perfil @EstadaoPolitica abordou ambos os candidatos mais vezes que os demais produtores de

notícia. Para o candidato Paulo Skaf, devido ao valor da mediana ser igual a zero, qualquer valor é considerado *outlier*, já que a razão de desvios em relação a mediana é sempre infinito.

Tabela 26 – Desvio do perfil em relação à proporção de referências aos candidatos por perfil

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	3,06	0,17	∞	2,34	-1,96	0,0
@g1politica	-0,67	0,0	∞	-0,09	0,67	-1,25
@folha_poder	0,0	0,67	∞	1,04	0,0	0,2
@cartacapital	0,45	-0,75	∞	-0,67	0,62	-0,67
@VEJA	-1,16	-0,75	∞	0,0	-0,85	0,98

6.3 Viés de cobertura

O viés de cobertura é uma tentativa de analisar a quantidade de cobertura recebida por cada entidade, que é tipicamente medido analisando o tamanho de colunas, as fotografias e as manchetes (D’ALESSIO; ALLEN, 2000). Para notícias distribuídas pela internet, o efeito desse tipo de viés é menor, já que há menos restrições de espaço e tempo em comparação com outros meios de comunicação. No contexto da rede social Twitter, uma métrica possível é identificar se a entidade foi referenciada diretamente no tweet associado à notícia. A presença da entidade já no tweet é importante pois, de acordo com o princípio da pirâmide invertida, os jornalistas devem organizar as informações e fatos em ordem decrescente de importância (PARK et al., 2012). Ou seja, é mais importante a presença da entidade no tweet, que funciona como manchete da notícia, do que no corpo da notícia.

Na tabela 27 está a relação de candidatos por perfil do Twitter. Para cada candidato foi calculado o percentual de tweets em que foi citado. Como esperado, os resultados foram similares ao viés de seleção, com o destaque prevalecendo em relação aos candidatos à presidência.

A quantidade de desvios dos perfis em relação a proporção de citações dos candidatos nos tweets está na tabela 28. Como esperado, nos tweets também não houve referências ao candidato Paulo Skaf por três dos cinco perfis, tornando novamente o cálculo de desvios igual a infinito devido à mediana igual a zero. O perfil @EstadaoPolitica novamente ultrapassou positivamente o valor limite de 2 desvios para os candidatos Geraldo Alckmin

e Aécio Neves e também ultrapassou negativamente o valor limite de 2 desvios para a candidata Dilma Rousseff. O perfil @VEJA ultrapassou o valor limite positivamente para a candidata Marina Silva.

Tabela 27 – Proporção de referências aos candidatos nos tweets

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	14,63%	2,44%	0,0%	29,27%	12,2%	29,27%
@g1politica	1,69%	1,98%	4,52%	5,65%	42,09%	23,73%
@folha_poder	4,3%	3,76%	3,23%	14,52%	32,93%	40,32%
@cartacapital	5,6%	0,0%	0,0%	0,0%	33,6%	27,2%
@VEJA	0,0%	0,0%	0,0%	6,56%	26,23%	53,01%
Média	5,24%	1,64%	1,35%	11,2%	29,41%	34,71%

Tabela 28 – Desvio em relação à proporção de referências aos candidatos nos tweets

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	2,67	0,17	∞	2,34	-2,09	0,0
@g1politica	-0,67	0,0	∞	-0,09	0,92	-0,67
@folha_poder	0,0	0,67	∞	0,82	0,0	1,35
@cartacapital	0,34	-0,75	∞	-0,67	0,07	-0,25
@VEJA	-1,11	-0,75	∞	0,0	-0,67	2,89

6.4 Viés de afirmação

O viés de afirmação ocorre quando os membros da mídia procuram colocar suas próprias opiniões na cobertura jornalística, em geral é analisado comparando textos “positivos” e “negativos” em relação a um tema (D’ALESSIO; ALLEN, 2000). Em nosso corpus, esse tipo de viés será analisado utilizando as três polaridades em que os parágrafos do corpus foram anotados: positivo, neutro e negativo. Para cada uma das polaridades, será considerada a proporção de parágrafos de cada perfil em relação às entidades selecionadas para análise.

Os dados para a proporção de parágrafos classificados como positivos em relação às entidades estão na tabela 29, enquanto que a quantidade de desvios dos perfis em relação aos parágrafos classificados como positivo estão na tabela 30. O limite de desvios foi ultrapassado negativamente para a candidata Dilma Rousseff pelo perfil @VEJA e para

o candidato Aécio Neves pelo perfil @cartacapital². O perfil @EstadaoPolitica ultrapassou positivamente o limite de desvios para a candidata Marina Silva. Para os candidatos Alexandre Padilha e Paulo Skaf, os desvios são infinitos pois para ambos os candidatos a mediana é igual a zero.

Tabela 29 – Proporção de parágrafos classificados como positivos para cada candidato

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	50,0%	0,0%	0,0%	50,0%	20,0%	62,5%
@g1politica	0,0%	57,14%	56,25%	40,0%	25,3%	50,0%
@folha_poder	26,67%	28,57%	4,17%	35,48%	35,95%	40,52%
@cartacapital	14,29%	0,0%	0,0%	0,0%	31,03%	35,0%
@VEJA	0,0%	0,0%	0,0%	50,0%	3,23%	33,33%
Média	18,19%	17,14%	12,08%	35,1%	23,1%	44,27%

Tabela 30 – Desvio em relação à proporção de parágrafos classificados como positivos

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	1,69	∞	∞	0,67	-0,62	2,06
@g1politica	-0,67	∞	∞	0,0	0,0	0,89
@folha_poder	0,58	∞	∞	-0,3	1,25	0,0
@cartacapital	0,0	∞	∞	-2,7	0,67	-0,52
@VEJA	-0,67	∞	∞	0,67	-2,6	-0,67

Os dados para a proporção de parágrafos classificados como neutros em relação às entidades estão na tabela 31, enquanto que a quantidade de desvios dos perfis em relação aos parágrafos classificados como neutro estão na tabela 32. O único desvio que passou do limiar definido foi do perfil @cartacapital, de forma negativa, em relação ao candidato Aécio Neves. Entretanto vale lembrar que o perfil @cartacapital não fez referências ao candidato, ou seja, o viés de afirmação não pode ser analisado. Para os três candidatos ao governo a mediana é igual a zero, tornando a quantidade de desvios infinita.

Os dados para a proporção de parágrafos classificados como negativo em relação às entidades estão na tabela 33, enquanto que a quantidade de desvios dos perfis em relação aos parágrafos classificados como negativos estão na tabela 34. Para esse tipo de viés, o perfil @VEJA chegou a 6 desvios na proporção de parágrafos negativos em relação à

² O perfil @cartacapital não teve parágrafos associados ao candidato Aécio Neves, logo esse desvio é decorrência da omissão sobre o candidato

Tabela 31 – Proporção de parágrafos classificados como neutros para cada candidato

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	0,0%	0,0%	0,0%	50,0%	80,0%	12,5%
@g1politica	100,0%	42,86%	43,75%	30,0%	45,78%	30,43%
@folha_poder	26,67%	35,71%	16,67%	38,71%	29,41%	24,18%
@cartacapital	0,0%	0,0%	0,0%	0,0%	44,83%	35,0%
@VEJA	0,0%	0,0%	0,0%	50,0%	16,13%	17,78%
Média	25,33%	15,71%	12,08%	33,74%	43,23%	23,98%

Tabela 32 – Desvio em relação à proporção de parágrafos classificados como neutros

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	∞	∞	∞	0,67	1,54	-1,23
@g1politica	∞	∞	∞	-0,52	0,04	0,66
@folha_poder	∞	∞	∞	0,0	-0,67	0,0
@cartacapital	∞	∞	∞	-2,31	0,0	1,14
@VEJA	∞	∞	∞	0,67	-1,26	-0,67

candidata Dilma Rousseff. O perfil @VEJA também superou o limiar positivamente em relação à candidata Marina Silva. O @EstadaoPolitica superou negativamente o limiar em relação a candidata Dilma Rousseff. Para os candidatos Alexandre Padilha e Paulo Skaf, os desvios são infinitos pois para ambos os candidatos a mediana é igual a zero.

Tabela 33 – Proporção de parágrafos classificados como negativos para cada candidato

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	50,0%	100,0%	0,0%	0,0%	0,0%	25,0%
@g1politica	0,0%	0,0%	0,0%	30,0%	28,92%	19,57%
@folha_poder	46,67%	35,71%	79,17%	25,81%	34,64%	35,29%
@cartacapital	85,71%	0,0%	0,0%	0,0%	24,14%	30,0%
@VEJA	0,0%	0,0%	0,0%	0,0%	80,65%	48,89%
Média	36,48%	27,14%	15,83%	11,62%	33,67%	31,75%

6.4.1 Distância de Mahalanobis para o viés de afirmação

Para a análise de cada tipo de viés, foi utilizada uma técnica para variáveis univariadas, possibilitando assim a análise individual de cada um dos tipos de vieses.

Tabela 34 – Desvio em relação a proporção de parágrafos classificados como negativos

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	0,06	∞	∞	∞	-3,41	-0,64
@g1politica	-0,81	∞	∞	∞	0,0	-1,33
@folha_poder	0,0	∞	∞	∞	0,67	0,67
@cartacapital	0,67	∞	∞	∞	-0,56	0,0
@VEJA	-0,81	∞	∞	∞	6,09	2,41

Entretanto, o viés de afirmação é composto por três métricas complementares: a proporção de parágrafos positivos, neutros e negativos. Nesse caso, é interessante considerar uma análise multi-dimensional para o problema.

Uma forma de realizar esse tipo de classificação é a partir do uso da distância de Mahalanobis ([MAESSCHALCK](#); [JOUAN-RIMBAUD](#); [MASSART, 2000](#)) de um vetor de variáveis aleatórias em relação ao restante do grupo, sendo possível assim definir o quão distante da média um ponto está para múltiplas variáveis. A distância de Mahalanobis (MD) de um ponto i é definida da seguinte forma

$$MD_i = \sqrt{(x_i - \bar{x})\mathbf{C}^{-1}(x_i - \bar{x})^T}$$

sendo x_i um vetor de observações, $\bar{x}$ o vetor com a média observações e $\mathbf{C}^{-1}$ a matriz de covariância entre as variáveis que compõem o modelo. Ao considerar a variância e covariância das variáveis do conjunto, a distância de Mahalanobis tem como vantagem, em relação à distância euclidiana, interpretar a distribuição de dados de forma “não esférica”. Na figura 12 o gráfico (a) contém um conjunto de dados com duas variáveis x_1 e x_2 e os círculos representando distâncias euclidianas iguais em relação ao centro de gravidade, o gráfico (b) contém o mesmo conjunto de dados e as elipses representando distâncias de Mahalanobis iguais em relação ao centro de gravidade.

Figura 12 – Gráficos representando a diferença entre a distância euclidiana (a) e distância de Mahalanobis (b)

Fonte: (MAESSCHALCK; JOUAN-RIMBAUD; MASSART, 2000)

Na tabela 35 é apresentada a distância de Mahalanobis de cada perfil por candidato em relação ao centro de gravidade. Nesse tipo de métrica, não é possível analisar o sentido do desvio, mas é possível identificar quais meios de comunicação tiveram uma abordagem mais distante do centro de gravidade, ou seja, da proporção média de parágrafos positivos, neutros e negativos.

Tabela 35 – Distância de Mahalanobis em relação ao centro de gravidade

Perfil	Entidade					
	Alckmin	Padilha	Skaf	Neves	Rousseff	Silva
@EstadaoPolitica	2,593	3,2	0,533	1,2	2,732	2,66
@g1politica	2,991	3,2	3,2	3,2	0,0353	1,333
@folha.poder	0,248	3,2	3,2	3,2	1,710	0,1156
@cartacapital	2,969	1,2	0,533	3,2	0,4	1,508
@VEJA	3,2	1,2	0,533	1,2	3,130	2,382

Analisando os resultados é possível identificar, por exemplo, que o perfil @VEJA diverge mais que os demais em relação ao candidato Dilma Rousseff, assim como o perfil @EstadaoPolitica. A natureza do viés é diferente; enquanto @VEJA contém uma proporção muito alta de parágrafos negativos em relação à candidata Dilma Rousseff, o perfil @EstadaoPolitica tem uma proporção muito alta de parágrafos neutros. Uma grande diferença também pode ser observada no @EstadaoPolitica e @VEJA em relação à candidata Marina Silva; o perfil @VEJA novamente tem uma alta proporção de parágrafos negativos em relação à candidata e o perfil @EstadaoPolitica tem uma grande proporção de

parágrafos positivos. Naturalmente, esses dados já poderiam ser identificados nas análises unidimensionais, mas a distância de Mahalanobis fornece uma visão global do viés de afirmação.

6.5 Correlação entre os diferentes tipos de vieses

As métricas indicadoras de viés foram extraídas e comparadas entre os diferentes produtores de notícia. Entretanto não foi identificado se existe uma relação entre os diferentes tipos de vieses. Uma relação possível seria, por exemplo, uma entidade comumente destacada nas manchetes (viés de cobertura) ter mais parágrafos negativos, positivos ou neutros associadas à ela em relação aos demais.

Para analisar os vieses em conjunto, foi calculada a correlação entre as métricas extraídas. Na tabela 36 está o cálculo do coeficiente de correlação de Pearson (r) (FILHO; JÚNIOR, 2009) e o respectivo valor de p , no formato (r, p) , entre as métricas de vieses: o viés de seleção (Tabela 25), viés de cobertura (Tabela 27) e o viés de afirmação separado em positivo, neutro e negativo (Tabelas 29, 31 e 33 respectivamente), com intervalo de confiança de 95%.

Tabela 36 – Correlação de Pearson entre os diferentes tipos de vieses

	Seleção	Cobertura	Positivo	Neutro	Negativo
Seleção	(1 , 0)	(0,97 , $2,36 \times 10^{-18}$)	(0,49 , 0,006)	(0,23 , 0,22)	(0,27 , 0,14)
Cobertura	–	(1 , 0)	(0,46 , 0,01)	(0,2 , 0,28)	(0,22 , 0,23)
Positivo	–	–	(1 , 0)	(0,34 , 0,06)	(–0,08 , 0,69)
Neutro	–	–	–	(1 , 0)	(–0,27, 0,15)
Negativo	–	–	–	–	(1 , 0)

Assim como para a definição de *outliers*, a definição de um valor para identificar uma correlação forte ou fraca é subjetiva. Para (COHEN, 2013), valores de r no intervalo $0,1 \leq r \leq 0,29$ podem ser considerados baixos; no intervalo $0,3 \leq r \leq 0,49$ podem ser considerados médios; e valores no intervalo $0,5 \leq r \leq 1$ podem ser interpretados como fortes. (DANCEY; REIDY, 2007) apontam para uma classificação ligeiramente diferente: valores de r no intervalo $0,1 \leq r \leq 0,39$ podem ser considerados fracos; valores no intervalo $0,4 \leq r \leq 0,69$ podem ser considerados médios; e valores no intervalo $0,7 \leq r \leq 1$ podem ser interpretados como fortes.

Analisando esses limiares, é possível identificar uma forte correlação entre viés de cobertura e viés de seleção, indicando que as entidades destacadas nos *tweets* são os candidatos mais abordados pelas notícias. Além disso, há uma correlação mediana do viés de seleção/cobertura e o viés de afirmação positivo.

6.6 Discussão

Nessa seção foi proposta uma metodologia para identificação de *outliers* baseada na comparação entre diferentes meios de comunicação. Optou-se por uma abordagem comparativa pois não é possível definir para cada métrica um valor “neutro”. Se alguma métrica for demasiadamente distinta das demais, o meio de comunicação é classificado como enviesado em relação a ela.

As métricas de viés de seleção e cobertura mostraram uma forte correlação. A quantidade de vezes que uma entidade é citada em um tweet é proporcional à quantidade de vezes que a entidade é definida como alvo dos parágrafos. Apesar do viés de seleção e afirmação trazerem informações muito similares entre si, essa forte correlação tem relevância para fins práticos. Os tweets estão restritos a 140 caracteres, o que torna mais simples seu uso em técnicas de PLN em comparação ao uso do corpo das notícias, por exemplo, pois é mais simples extrair entidades alvos de tweets que do corpo das notícias. Ou seja, em relação às entidades abordadas nas notícias, é possível trabalhar apenas com os tweets sem grandes perdas de informação.

Os resultados do viés de seleção/cobertura mostraram um maior enfoque dos meios de comunicação aos candidatos à presidência em comparação aos candidatos ao governo de São Paulo. Os candidatos à re-eleição, Dilma Rousseff para presidência e Geraldo Alckmin para o governo, também receberam mais atenção que os demais candidatos ao mesmo cargo. O @EstadaoPolitica atribuiu maior atenção aos candidatos do PSDB para presidência e governo, sendo classificado como *outliers* para Geraldo Alckmin e Aécio Neves. Além disso, o @EstadaoPolitica também abordou menos os demais candidatos (Marina Silva e Dilma Rousseff) à presidência em comparação aos demais meios de comunicação, apesar de não ultrapassar o limiar de 2 desvios.

Não foi encontrada uma relação clara entre viés de afirmação e viés de seleção/cobertura, apenas uma correlação, entre média e forte, do viés de seleção/cobertura e do viés de afirmação positivo. Ou seja, quanto mais abordada a entidade, mais positiva a abordagem

em relação a ela. Entretanto também não se observou uma relação contrária, ou seja, quanto menor o destaque à entidade, maior a proporção de parágrafos classificados como negativos em relação a mesma. Uma possibilidade, nesse caso, é que a omissão seja uma outra estratégia. @VEJA e @CartaCapital são perfis de linhas editoriais distintas, a @VEJA teve um grande viés positivo de parágrafos negativos em relação à Dilma Rousseff (6 desvios) enquanto a @CartaCapital ficou levemente abaixo da mediana em relação a parágrafos negativos. Em contrapartida, a @VEJA teve um viés negativo de parágrafos positivos em relação à Dilma Rousseff e a @CartaCapital ficou ligeiramente acima da média. Em relação ao candidato da oposição Aécio Neves, ao invés de uma grande quantidade de textos negativos, como feito pela @VEJA, a @CartaCapital não deu destaque ao candidato da oposição com zero referência.

Para o viés de afirmação, também foi calculado uma métrica multi-dimensional para análise de viés, a distância de Mahalanobis. É uma abordagem interessante, por permitir uma análise conjunta das três variáveis que constituem esse tipo de viés utilizando o mesmo conceito de desvio em relação a um indicador de tendência central. Entretanto, a análise multi-dimensional dificulta a análise comparativa em relação aos demais tipos de viés já que a distância não identifica se o desvio é positivo ou negativo em relação à média. No contexto da análise de viés, a distância de Mahalanobis pode ser uma ferramenta importante caso se opte por adicionar mais métricas à análise (para os vieses de seleção e cobertura foi utilizada uma métrica indicadora para cada tipo de viés, mas poderiam ser utilizadas mais métricas). À medida que se aumenta a quantidade de dimensões do problema, uma métrica multi-dimensional se torna mais interessante para facilitar a análise dos resultados.

7 Conclusão

Nesse projeto foi estudado a análise de viés dos produtores de notícia em relação a figuras políticas. A execução desse projeto pode ser separada em três etapas: criação de um corpus de notícias sobre política extraídas de cinco meios de comunicação distintos e anotadas em relação a polaridade e entidade; avaliação de técnicas de análise de sentimentos sobre o corpus construído; e proposição de uma metodologia para análise de viés baseada em desvios a partir de um indicador de tendência central.

7.1 Criação de Corpus

Para o projeto, era necessário um corpus de notícias que contivesse notícias sobre política de diversos produtores, extraídas durante um mesmo período. Até onde foi possível verificar, não foi encontrado nenhum que atendesse essa necessidade. Foram extraídas notícias de cinco produtores de notícias a partir do Twitter e, após segmentá-las em parágrafos, foi realizada a classificação dos parágrafos em relação à polaridade da entidade associada ao parágrafo.

O corpus desenvolvido será disponibilizado publicamente ([ARRUDA; ROMAN; MONTEIRO, 2015](#)), tornando-se assim uma ferramenta de linha de base para outros trabalhos que procurem aplicar técnicas de análise de sentimentos em língua portuguesa utilizando o padrão ouro do corpus. O uso do conceito de alvo da polaridade no método de anotação é um aspecto do corpus que difere de outros modelos de anotação e que pode oferecer novas possibilidades, como a classificação por entidade, por exemplo (Seção [5.3.2](#)).

7.2 Análise de sentimentos sobre as notícias do corpus

A análise de sentimentos é um aspecto importante da análise de viés em notícias. Técnicas que possibilitem a classificação automática de texto em relação a polaridade são importantes para que seja possível automatizar a análise de viés. Devido a essa necessidade, técnicas de análise de sentimentos baseadas em aprendizado de máquina foram aplicadas sobre o corpus criado.

Os resultados mostraram que o algoritmo baseado em entropia não superou o desempenho dos algoritmos linha de base e também que a classificação de notícias é uma

tarefa complexa para esses algoritmos (a acurácia não ultrapassou 0,7 em nenhum teste). A classificação restrita à Dilma Rousseff apresentou melhores resultados que a classificação geral, enquanto a classificação restrita a Marina Silva apenas manteve o desempenho. Pelos resultados obtidos, a classificação de notícias em relação a polaridade se mostra um problema em aberto, incluindo a granularidade da classificação e o escopo de treinamento.

7.3 Análise de viés a partir *outliers*

A análise de viés é um processo que precisa ser realizado de forma sistemática, devido à dificuldade em se analisar esse problema objetivamente. Para lidar com essa questão, foi proposto um modelo baseado em comparação de diversos meios de comunicação, uma vez que não é possível definir *a priori* se determinado meio de comunicação está sendo demasiadamente crítico ou dando muito destaque a determinado candidato. Entretanto é possível analisá-los comparativamente.

O viés foi analisado a partir de três perspectivas: o viés de seleção, indicado pela proporção de parágrafos em que a entidade foi definida como alvo; o viés de cobertura, definido pela presença da entidade nos tweets das notícias e, por fim, o viés de afirmação definido pela proporção de parágrafos classificados como positivo, neutro ou negativo em relação à entidade. Para cada métrica, foi analisada a quantidade de desvios em relação à mediana e, para o viés afirmação, também foram analisadas em conjunto as três métricas utilizando uma abordagem multi-dimensional.

Para trabalhos futuros, seria interessante a automatização desse método: além do uso da análise de sentimentos estudadas neste projeto para o viés de afirmação, seria interessante o uso de técnicas de identificação de entidades para o viés de seleção e cobertura. Além de automatizar a aplicação das técnicas, uma outra possibilidade seria utilizar o teste de Kolmogorov-Smirnov para identificar quão diferentes são as coberturas dos jornais em relação a cada métrica.

Os resultados mostraram que há dificuldades na aplicação de análise de sentimentos no corpo das notícias e que as entidades destacadas nos tweets são as mais abordadas na notícia, indicando que trabalhar apenas com os tweets pode ser uma abordagem interessante do ponto de vista prático, já que é um domínio mais simples de se trabalhar.

Referências¹

- ABDUL-MAGEED, M.; DIAB, M. AWATIF : A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. *LREC*, p. 3907–3914, 2012. Disponível em: <<http://www.seas.gwu.edu/~mtdiab/files/publications/refereed/13.pdf>>. Citado 3 vezes nas páginas 21, 45 e 49.
- ABDUL-MAGEED, M.; KORAYEM, M.; YOUSSEFAGHA, A. "Yes we can?": Subjectivity Annotation and Tagging for the Health Domain. *RANLP*, n. September, p. 666–671, 2011. Disponível em: <<http://www.aclweb.org/anthology-new/R/R11/R11-1096>>. Citado na página 21.
- ALEIXO, P.; PARDO, T. A. S. *CSTNews: um cópulus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory)*. [S.l.], 2008. Citado na página 49.
- ALUÍSIO, S.; PELIZZONI, J.; MARCHI, a. An account of the challenge of tagging a reference corpus for brazilian portuguese. ... of the Portuguese ..., n. i, p. 110–117, 2003. ISSN 03029743. Disponível em: <<http://www.springerlink.com/index/5vh6lfqfdnlqlek.pdf>>. Citado na página 61.
- ALVARES, A. R.; ROMAN, N. T. AgreeCalc : Uma Ferramenta para Análise da Concordância entre Múltiplos Anotadores. In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2013. p. 1–10. Citado na página 53.
- ALVARES, R. V.; GARCIA, A. C. B.; FERRAZ, I. Stembr: A stemming algorithm for the brazilian portuguese language. In: *Progress in Artificial Intelligence*. [S.l.]: Springer, 2005. p. 693–701. Citado na página 58.
- ARRUDA, G. D.; ROMAN, N. T. R.; MONTEIRO, A. An annotated corpus for sentiment analysis in political news. In: *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 49 e 86.
- BALAHUR, A. et al. Opinion Mining on Newspaper Quotations. In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2009. v. 3, p. 523–526. ISBN 978-0-7695-3801-3. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5285040>
<<http://dl.acm.org/citation.cfm?id=1632189.1632232>>. Citado 5 vezes nas páginas 15, 20, 32, 45 e 71.
- BALAHUR, A.; STEINBERGER, R.; KABADJOV, M. Sentiment Analysis in the News. *LREC*, p. 2216–2220, 2010. Disponível em: <http://www.researchgate.net/publication/220746038_Sentiment_Analysis_in_the_News/file/50463516f07379dc26.pdf>. Citado 3 vezes nas páginas 44, 45 e 55.
- BRANTS, T.; FRANZ, A. {Web 1T 5-gram Version 1}. Linguistic Data Consortium, Philadelphia, 2006. Citado na página 31.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- CAI, K. et al. Leveraging Sentiment Analysis for Topic Detection. In: *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2008. p. 265–271. Disponível em: <http://dl.acm.org/citation.cfm?id=1486927.1487036>. Citado 3 vezes nas páginas 20, 29 e 32.
- CARDEI, C.; MANISOR, F.; REBEDEA, T. Opinion mining for social media and news items in Romanian. In: *2nd International Conference on Systems and Computer Science*. IEEE, 2013. p. 240–245. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6632054>. Citado na página 44.
- CARDOSO, P. C. et al. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: *the Proceedings of the 3rd RST Brazilian Meeting*. [S.l.: s.n.], 2011. p. 88–105. Citado na página 49.
- CHAUMARTIN, F.-R. UPAR7: a knowledge-based system for headline sentiment tagging. Association for Computational Linguistics, p. 422–425, 2007. Disponível em: <http://dl.acm.org/citation.cfm?id=1621474.1621568>. Citado 3 vezes nas páginas 21, 33 e 44.
- COHEN, J. *Statistical power analysis for the behavioral sciences*. [S.l.]: Academic press, 2013. Citado na página 83.
- COLLINS, M.; DUFFY, N. Convolution kernels for natural language. In: DIETTERICH, T.; BECKER, S.; GHAHRAMANI, Z. (Ed.). *Advances in Neural Information Processing Systems 14*. MIT Press, 2002. p. 625–632. Disponível em: <http://papers.nips.cc/paper/2089-convolution-kernels-for-natural-language.pdf>. Citado na página 25.
- COLLOBERT, R. et al. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, v. 12, p. 2493–2537, 2011. ISSN 1532-4435. Disponível em: <http://arxiv.org/abs/1103.0398>. Citado na página 59.
- COUSINEAU, D.; CHARTIER, S. Outliers detection and treatment: a review. *International Journal of Psychological Research*, v. 3, n. 1, p. 58–67, 2010. ISSN 2011-7922. Disponível em: <http://mvint.usbmed.edu.co:8002/ojs/index.php/web/article/view/460>. Citado 2 vezes nas páginas 73 e 74.
- CURRAN, T.; KOPRINSKA, P. An Annotated Corpus of Quoted Opinions in News Articles. *tokeefe.org*, p. 516–520, 2013. Disponível em: <http://www.tokeefe.org/blog/wp-content/uploads/2013/08/acl13shortopinions.pdf>. Citado 5 vezes nas páginas 22, 45, 49, 55 e 71.
- D’ALESSIO, D.; ALLEN, M. Media bias in presidential elections: a meta-analysis. *Journal of Communication*, v. 50, n. 4, p. 133–156, 2000. ISSN 0021-9916. Disponível em: <http://doi.wiley.com/10.1111/j.1460-2466.2000.tb02866.x>. Citado 4 vezes nas páginas 57, 75, 77 e 78.
- DANCEY, C. P.; REIDY, J. *Statistics without maths for psychology*. [S.l.]: Pearson Education, 2007. Citado na página 83.

DAS, A.; BANDYOPADHYAY, S. Topic-based Bengali opinion summarization. ... of the 23rd International Conference on ..., n. August, p. 232–240, 2010. Disponível em: <<http://dl.acm.org/citation.cfm?id=1944593>>. Citado na página 45.

DENG, S. et al. Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction. In: *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 2011. p. 800–807. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6118898>>. Citado na página 47.

DEVITT, A.; AHMAD, K. Sentiment polarity identification in financial news: A cohesion-based approach. *ACL*, n. June, p. 984–991, 2007. Disponível em: <<http://acl.ldc.upenn.edu/P/P07/P07-1124.pdf>>. Citado na página 46.

DING, X.; LIU, B.; YU, P. A holistic lexicon-based approach to opinion mining. *Proceedings of the 2008 International Conference ...*, 2008. Disponível em: <<http://dl.acm.org/citation.cfm?id=1341561>>. Citado na página 43.

DRURY, B.; ALMEIDA, J. The Minho Quotation Resource. *LREC*, p. 2280–2285, 2012. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/611_Paper.pdf>. Citado na página 45.

DRURY, B.; DIAS, G.; TORGO, L. A Contextual Classification Strategy for Polarity Analysis of Direct Quotations from Financial News. *RANLP*, v. 2011, n. September, p. 434–440, 2011. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.4167&rep=rep1&type=pdf#page=468>>. Citado na página 35.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012. Citado na página 58.

FILHO, D. B. F.; JÚNIOR, J. A. d. S. Desvendando os mistérios do coeficiente de correlação de Pearson (r). *Revista Política Hoje*, v. 18, n. 1, p. 115–146, 2009. ISSN 0104-7094. Citado na página 83.

FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. R. L. Background to Framenet. *International Journal of Lexicography*, v. 16, n. 3, p. 235–250, set. 2003. Disponível em: <<http://ijl.oxfordjournals.org/content/16/3/235.abstract>>. Citado na página 27.

FLORES, F. N.; MOREIRA, V. P.; HEUSER, C. a. Assessing the impact of stemming accuracy on information retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 6001 LNAI, p. 11–20, 2010. ISSN 03029743. Citado na página 58.

FONG, S. et al. Sentiment Analysis of Online News Using MALLET. In: *2013 International Symposium on Computational and Business Intelligence*. IEEE, 2013. p. 301–304. ISBN 978-0-7695-5066-4. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6724372>>. Citado 2 vezes nas páginas 35 e 44.

FONSECA, E. R.; ROSA, G. Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging. p. 98–107, 2013. Citado 2 vezes nas páginas 59 e 60.

FONSECA, E. R.; ROSA, J. a. L. G.; ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. *Journal of the Brazilian Computer Society*, v. 21, 2015. ISSN 0104-6500. Disponível em: <<http://www.journal-bcs.com/content/21/1/2>>. Citado na página 61.

FORNEY G.D., J. The viterbi algorithm. *Proceedings of the IEEE*, v. 61, n. 3, 1973. ISSN 0018-9219. Citado na página 61.

FU, T.-c. et al. Discovering the Correlation between Stock Time Series and Financial News. In: *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2008. v. 1, p. 880–883. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4740568>>. Citado na página 33.

GAO, Y. et al. Sentiment classification for stock news. In: *5th International Conference on Pervasive Computing and Applications*. IEEE, 2010. p. 99–104. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5704082>>. Citado 2 vezes nas páginas 23 e 35.

GRUBBS, F. E. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, v. 11, n. 1, p. 1–21, 1969. ISSN 0040-1706. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657>>. Citado na página 73.

HAIDER, S.; MEHROTRA, R. Corporate news classification and valence prediction: A supervised approach. ... *2nd Workshop on Computational Approaches* ..., p. 175–181, 2011. Disponível em: <<http://dl.acm.org/citation.cfm?id=2107676>>. Citado na página 29.

HAMPEL, F. R. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, v. 69, n. 346, p. 383–393, 1974. ISSN 0162-1459. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962>>. Citado na página 75.

HERSH, W. et al. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Springer-Verlag New York, Inc., 1994. (SIGIR '94), p. 192–201. ISBN 0-387-19889-X. Disponível em: <<http://dl.acm.org/citation.cfm?id=188490.188557>>. Citado na página 37.

HODGE, V. J.; AUSTIN, J. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, v. 22, n. 2, p. 85–126, 2004. Citado na página 73.

IYENGAR, S.; HAHN, K. S. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, v. 59, n. 1, p. 19–39, mar. 2009. ISSN 00219916. Disponível em: <<http://doi.wiley.com/10.1111/j.1460-2466.2008.01402.x>>. Citado na página 14.

JACCARD, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, v. 37, p. 241–272, 1901. Citado na página 46.

JANG, H.; SHIN, H. Effective Use of Linguistic Features for Sentiment Analysis of Korean. *PACLIC*, p. 173–182, 2010. Disponível em: <<http://www.aclweb.org/anthology-Y/Y10/Y10-1021.pdf>>. Citado na página 57.

JANG, H.; SHIN, H. Language-specific sentiment analysis in morphologically rich languages. Association for Computational Linguistics, p. 498–506, 2010. Disponível em: <<http://dl.acm.org/citation.cfm?id=1944566.1944623>>. Citado 5 vezes nas páginas 22, 31, 34, 45 e 55.

JOACHIMS, T. Text categorization with suport vector machines: Learning with many relevant features. In: *Proceedings of the 10th European Conference on Machine Learning*. London, UK, UK: Springer-Verlag, 1998. (ECML '98), p. 137–142. ISBN 3-540-64417-2. Disponível em: <<http://dl.acm.org/citation.cfm?id=645326.649721>>. Citado na página 37.

JURAFSKY, D. *Sentiment Analysis: What is Sentiment Analysis?* 2014. Disponível em: <<http://web.stanford.edu/class/cs124/lec/sentiment.pdf>>. Citado na página 20.

KAYA, M.; FIDAN, G.; TOROSLU, I. H. Sentiment Analysis of Turkish Political News. In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE, 2012. v. 1, p. 174–180. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6511881>>. Citado 2 vezes nas páginas 35 e 57.

KIM, S.; HOVY, E. Automatic detection of opinion bearing words and sentences. *Proceedings of IJCNLP*, p. 61–66, 2005. Disponível em: <http://acl.ldc.upenn.edu/I/I05/I05-2011.pdf?origin=publication_detail>. Citado na página 46.

KIM, S.; HOVY, E. Extracting opinions, opinion holders, and topics expressed in online news media text. ... of the *Workshop on Sentiment and Subjectivity in Text*, n. July, p. 1–8, 2006. Disponível em: <<http://dl.acm.org/citation.cfm?id=1654642>>. Citado na página 21.

KIVINEN, J.; WARMUTH, M.; AUER, P. The perceptron algorithm versus winnow: linear versus logarithmic mistake bounds when few input variables are relevant. *Artificial Intelligence*, v. 97, n. 1-2, p. 325–343, 1997. ISSN 00043702. Citado na página 37.

LEWIS, D. D. *Reuters-21578, distribution 1.0*. 1997. Citado na página 37.

LEYS, C. et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, Elsevier Inc., v. 49, n. 4, p. 764–766, jul. 2013. ISSN 00221031. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0022103113000668>>. Citado na página 74.

LI, H. et al. Annotating Opinions in German Political News. *LREC*, p. 1183–1188, 2008. Citado na página 49.

LIN, Y.-R.; BAGROW, J. P.; LAZER, D. "Quantifying Bias in Social and Mainstream Media" by Yu-Ru Lin, James P. Bagrow, and David Lazer with Ching-man Au Yeung as coordinator. *ACM SIGWEB Newsletter*, ACM, n. Summer, p. 1–6, 2012. ISSN 19311745. Disponível em: <<http://dl.acm.org/citation.cfm?id=2246063.2246068>>. Citado 2 vezes nas páginas 15 e 46.

LIU, B. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. Disponível em: <<http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016>>. Citado na página 20.

LLOYD, L.; KECHAGIAS, D.; SKIENA, S. Lydia : A system for large-scale news analysis. *Lecture notes in computer science*, Springer, p. 161–166, 2005. ISSN 0302-9743. Disponível em: <<http://cat.inist.fr/?aModele=afficheN&cpsidt=17415995>>. Citado na página 47.

LORENA a.C.; CARVALHO a.C.P.L.F. de. Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Disponível em: <http://seer.ufrgs.br/index.php/rita/article/viewArticle/rita_v14_n2_p43-67>. Citado na página 36.

LUND, K.; BURGESS, C. *Producing high-dimensional semantic spaces from lexical co-occurrence*. 1996. 203–208 p. Citado na página 60.

MAESSCHALCK, R. D.; JOUAN-RIMBAUD, D.; MASSART, D. L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, v. 50, n. 1, p. 1–18, 2000. ISSN 0169-7439. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0169743999000477nbackslashhttp://linkinghub.elsevier.com/retrieve/pii/S0169743999000477>>. Citado 2 vezes nas páginas 81 e 82.

MALANDRAKIS, N. et al. Distributional Semantic Models for Affective Text Analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 21, n. 11, p. 2379–2392, 2013. ISSN 1558-7916. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6578101>>. Citado 2 vezes nas páginas 29 e 30.

MALO, P. et al. Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis. In: *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013. p. 945–954. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6754023>>. Citado 3 vezes nas páginas 25, 27 e 40.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. *An Introduction to Information Retrieval*. [s.n.], 2009. 569 p. ISSN 13864564. ISBN 0521865719. Disponível em: <<http://dspace.cusat.ac.in/dspace/handle/123456789/2538>>. Citado 2 vezes nas páginas 39 e 40.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1. Citado na página 24.

MARNEFFE, M.-C. D.; MANNING, C. D. Stanford typed dependencies manual. URL http://nlp.stanford.edu/software/dependencies_manual.pdf, 2008. Citado 2 vezes nas páginas 33 e 34.

MEHLER, A. et al. Spatial analysis of news sources. *IEEE transactions on visualization and computer graphics*, v. 12, n. 5, p. 765–771, 2006. ISSN 1077-2626. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17080798>>. Citado 2 vezes nas páginas 20 e 47.

MEYER, D.; WIEN, F. T. Support vector machines. *The Interface to libsvm in package e1071*, 2014. Citado na página 36.

MICHAELIS. *Moderno dicionário*. 2015. Disponível em: <<http://michaelis.uol.com.br/moderno/portugues/index.php?lingua=portugues-portugues&palavra=par%E1grafo>>. Citado na página 51.

MILLER, J. Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, Taylor & Francis, v. 43, n. 4, p. 907–912, 1991. Citado na página 75.

MOSCHITTI, A. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. p. 318–329, 2006. Citado na página 25.

NASUKAWA, T.; YI, J. Sentiment analysis. In: *Proceedings of the international conference on Knowledge capture - K-CAP '03*. ACM Press, 2003. p. 70. Disponível em: <<http://dl.acm.org/citation.cfm?id=945645.945658>>. Citado na página 33.

ORENGO, V.; HUYCK, C. A stemming algorithm for the portuguese language. *Proceedings Eighth Symposium on String Processing and Information Retrieval*, 2001. Citado na página 58.

PAICE, C. D. An Evaluation Method for Stemming Algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 42–50, 1994. Citado na página 58.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, v. 2, n. 1–2, p. 1–135, 2008. ISSN 1554-0669. Disponível em: <<http://www.nowpublishers.com/product.aspx?product=INR\&doi=1500000011>>. Citado 2 vezes nas páginas 18 e 20.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. ...-02 *conference on Empirical methods ...*, 2002. Disponível em: <<http://dl.acm.org/citation.cfm?id=1118704>>. Citado na página 21.

PARDO, T. A. S.; RINO, L. H. M. *TeMário: Um Corpus para Sumarização Automática de Textos*. [S.l.], 2003. Citado na página 49.

PARISER, E. *The filter bubble : what the Internet is hiding from you*. New York: Penguin Press, 2011. ISBN 9781594203008 1594203008. Citado na página 14.

PARK, S. et al. NewsCube. In: *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*. ACM Press, 2009. p. 443. Disponível em: <<http://dl.acm.org/citation.cfm?id=1518701.1518772>>. Citado 2 vezes nas páginas 14 e 19.

PARK, S. et al. A Computational Framework for Media Bias Mitigation. *ACM Transactions on Interactive Intelligent Systems*, ACM, v. 2, n. 2, p. 1–32, 2012. ISSN 21606455. Disponível em: <<http://dl.acm.org/citation.cfm?id=2209310.2209311>>. Citado 3 vezes nas páginas 15, 46 e 77.

RAINA, P. Sentiment Analysis in News Articles Using Sentic Computing. In: *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013. p. 959–962. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6754025>>. Citado na página 33.

ROCHA, P.; SANTOS, D. Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR*, v. 2000, p. 131–140, 2000. Citado na página 49.

ROMAN, N. T. et al. Sentiment and behaviour annotation in a corpus of dialogue summaries. *Journal of Universal Computer Science (J.UCS)*, v. 21, n. 4, p. 561–586, April 2015. ISSN 0948-695x (Online Edition: ISSN 0948-6968). Citado na página 55.

SAEZ-TRUMPER, D.; CASTILLO, C.; LALMAS, M. Social media news communities. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*. ACM Press, 2013. p. 1679–1684. Disponível em: <http://dl.acm.org/citation.cfm?id=2505515.2505623>. Citado 5 vezes nas páginas 19, 20, 46, 47 e 73.

SARVABHOTLA, K.; PINGALI, P.; VARMA, V. Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents. *Information Retrieval*, v. 14, n. 3, p. 337–353, fev. 2011. ISSN 1386-4564. Disponível em: <http://link.springer.com/10.1007/s10791-010-9161-5>. Citado na página 44.

SCHARL, A.; HERRING, D. D. Extracting knowledge from the web and social media for progress monitoring in public outreach and science communication. In: *Proceedings of the 19th Brazilian symposium on Multimedia and the web - WebMedia '13*. ACM Press, 2013. p. 121–124. Disponível em: <http://dl.acm.org/citation.cfm?id=2526188.2526219>. Citado na página 48.

SCHMITT, K. M. Why Partisans See Mass Media as Biased. *Communication Research*, v. 31, n. 6, p. 623–641, dez. 2004. ISSN 0093-6502. Disponível em: <http://crx.sagepub.com/cgi/doi/10.1177/0093650204269390>. Citado na página 14.

SCHOLZ, T.; CONRAD, S. Opinion Mining in Newspaper Articles by Entropy-Based Word Connections. *EMNLP*, n. October, p. 1828–1839, 2013. Disponível em: http://www.researchgate.net/publication/257985909_Opinion_Mining_in_Newspaper_Articles_by_Entropy-Based_Word_Connections/file/3deec5268d54864768.pdf. Citado 7 vezes nas páginas 16, 26, 41, 42, 43, 57 e 64.

SHANNON, C. A mathematical theory of communication. *Bell System Technical Journal*, v. 27, p. 379–423, 623–656, July, October 1948. Disponível em: <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>. Citado na página 43.

SIERING, M. "Boom" or "Ruin"--Does It Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions. In: *2012 45th Hawaii International Conference on System Sciences*. IEEE, 2012. p. 1050–1059. ISSN 1530-1605. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6149156>. Citado 3 vezes nas páginas 24, 32 e 45.

SILVA, B. C. Dias-da. Wordnet.Br: An Exercise of Human Language Technology Research. *Proceedings of The Third International WordNet Conference (GWC 2006)*, n. 1998, p. 301–303, 2006. Citado na página 29.

- SUDHAHAR, S.; LANSDALL-WELFARE, T. ElectionWatch: detecting patterns in news coverage of US elections. *Proceedings of the ...*, p. 82–86, 2012. Disponível em: <<http://dl.acm.org/citation.cfm?id=2380938>>. Citado 3 vezes nas páginas 15, 25 e 33.
- TABOADA, M.; BROOKE, J.; TOFILOSKI, M. Lexicon-based methods for sentiment analysis. *Computational ...*, n. September 2010, 2011. Disponível em: <http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049>. Citado na página 43.
- TURNEY, P. D. Thumbs up or thumbs down? In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Morristown, NJ, USA: Association for Computational Linguistics, 2001. p. 417. Disponível em: <<http://dl.acm.org/citation.cfm?id=1073083.1073153>>. Citado na página 21.
- URBAN, C. D. *Examining Our Credibility: Perspectives of the Public and the Press*. [S.l.]: Asne Foundation, 1999. Citado na página 14.
- VEALE, T.; HAO, Y. In the mood for affective search with web stereotypes. In: *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*. ACM Press, 2012. p. 429. Disponível em: <<http://dl.acm.org/citation.cfm?id=2187980.2188066>>. Citado na página 31.
- WARD, C. B.; BAUTIN, M.; SKIENA, S. Identifying Differences in News Coverage between Cultural/Ethnic Groups. In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2009. v. 3, p. 511–514. ISBN 978-0-7695-3801-3. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5285037> <http://dl.acm.org/citation.cfm?id=1632189.1632229>>. Citado na página 47.
- XIE, B.; PASSONNEAU, R. Semantic frames to predict stock price movement. *Proceedings of the 51st ...*, p. 873–883, 2013. Disponível em: <<http://oldsite.aclweb.org/anthology-new/P/P13/P13-1086.pdf>>. Citado 2 vezes nas páginas 27 e 28.
- YANAGIMOTO, H.; SHIMADA, M.; YOSHIMURA, A. Document similarity estimation for sentiment analysis using neural network. In: *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*. IEEE, 2013. p. 105–110. ISBN 978-1-4799-0174-6. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6607825>>. Citado na página 35.
- YOUNUS, A. et al. Investigating bias in traditional media through social media. In: *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*. ACM Press, 2012. p. 643. Disponível em: <<http://dl.acm.org/citation.cfm?id=2187980.2188168>>. Citado na página 46.
- YU, N.; KÜBLER, S. Filling the Gap : Semi-Supervised Learning for Opinion Detection Across Domains. *... on Computational Natural Language Learning*, n. June, p. 200–209, 2011. Disponível em: <<http://dl.acm.org/citation.cfm?id=2018959>>. Citado na página 35.
- ZHANG, H. The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference FLAIRS 2004*, v. 1, n. 2, p. 1 –

6, 2004. ISSN 01678655. Disponível em: <<http://www.aaai.org/Papers/FLAIRS/2004-/Flairs04-097.pdf>>. Citado na página 38.

ZHANG, J. et al. Sentiment Bias Detection in Support of News Credibility Judgment. In: *2011 44th Hawaii International Conference on System Sciences*. IEEE, 2011. p. 1–10. ISSN 1530-1605. Disponível em: <<http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5718623>>. Citado 2 vezes nas páginas 15 e 44.

ZHANG, W.; SKIENA, S. Improving Movie Gross Prediction through News Analysis. In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2009. v. 1, p. 301–304. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5286056>>. Citado na página 47.

ZHANG, X.; AHMAD, K. Proceedings of the 2nd workshop on sentiment analysis where ai meets psychology. In: _____. The COLING 2012 Organizing Committee, 2012. cap. Affect Proxies and Ontological Change: A finance case study, p. 99–114. Disponível em: <<http://aclweb.org/anthology/W12-5311>>. Citado na página 29.