Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Caracterização de Notícias Falsas por Meio de Léxicos de Subjetividade

Caio Libânio Melo Jerônimo

Proposta de Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Metodologia e Técnicas da Computação

Leandro Balby Marinho
Cláudio E. C. Campelo

Campina Grande, Paraíba, Brasil

Resumo

Métodos de detecção de notícias falsas baseados unicamente em características textuais permitem uma detecção precoce deste tipo de conteúdo, sem necessitar de informações como número de curtidas ou quantidade de compartilhamentos, informações disponíveis apenas quando a notícia já tem se disseminando nas redes sociais. Nesta linha de pesquisa, muitos trabalhos que conseguem resultados expressivos utilizam representações baseadas na frequência de ocorrência de palavras ou termos nos documentos. Porém, tais representações possuem limitações, e podem, por exemplo, enviesar modelos de classificação em situações onde o assunto dos documentos possa influenciar na classificação. Nesta pesquisa, buscamos representações mais robustas que permitam uma melhor diferenciação e compreensão de notícias falsas e reais. Propomos, nesta etapa da pesquisa, considerar a subjetividade das notícias partindo do pressuposto de que os níveis de subjetividade das notícias falsas e reais são significativamente diferentes. Neste estudo, como estratégia para extração de subjetividade em texto, propomos uma representação textual baseada no distanciamento semântico entre as sentenças de documentos de notícias e léxicos que expressam diferentes dimensões de subjetividade. Para validar a abordagem apresentada nesta pesquisa, foram executados ensaios de classificação e análises baseadas na explicação dos modelos construídos. Os resultados das análises demonstram que os modelos propostos, apesar de ainda apresentarem resultados inferiores ao estado da arte, apresentam uma melhor generalização quando comparados com modelos clássicos baseados em frequência de palavras, especialmente em cenários onde o assunto dos documentos varia entre os conjuntos de treino e teste dos modelos.

Abstract

Fake news detection methods based solely on textual characteristics allow an early detection of this type of content, without requiring information such as number of likes or number of shares, information available only when the news has already spread on social networks. In this line of research, many works that achieve expressive results use representations based on the frequency of occurrence of words or terms in the documents. However, such representations have limitations, and can, for example, skew classification models in situations where the subject of the documents may influence the classification. In this research, we seek more robust representations that allow better differentiation and understanding of fake and real news. We propose, at this stage of the research, to consider the subjectivity of news based on the assumption that the levels of subjectivity of fake and real news are significantly different. In this study, as a strategy for extracting subjectivity from text, we propose a textual representation based on the semantic distance between the sentences in news documents and lexicons that express different dimensions of subjectivity. To validate the approach presented in this research, classification tests and analyzes were performed. The results demonstrate that the proposed models, although still presenting results inferior to the state of the art, present a better generalization when compared with classic models, especially in scenarios where the subject of the documents varies between the training and testing sets.

Conteúdo

1 Introdução **1**

1.1 Motivação e Justificativa 2

1.2 Questões de Pesquisa 5

1.3 Objetivos 6

1.3.1 Objetivo Geral 6

1.3.2 Objetivos Específicos 6

1.4 Estrutura do Documento 7

2 Fundamentação Teórica **8**

2.1 Notícias Falsas 8

2.1.1 Definição de Notícia Falsa 8

2.1.2 Objetividade vs Subjetividade em Notícias Falsas 10

2.2 Processamento de Linguagem Natural 11

2.2.1 Evolução da PLN 11

2.2.2 Bag of Words 12

2.2.3 Word Embeddings 14

2.2.4 Word Mover's Distance 16

2.2.5 Valores SHAP 17

3 Trabalhos Relacionados **18**

3.1 Identificação de Notícias Falsas 18

3.2 Identificação de Notícias Falsas utilizando subjetividade 22

3.3 Posicionamento desta pesquisa em relação aos trabalhos relacionados . . . 25

4 Análise de Subjetividade por Meio de Distâncias Semânticas **29**

4.1 Metodologia Geral de Experimentação 29
4.2 Processamento das Características de Subjetividade 30
4.2.1 Média de Subjetividade por Documentos 30
4.2.2 Vetores de Subjetividade por Sentenças 31
4.3 Base de Dados 31
4.3.1 Base de Dados em Português 31
4.3.2 Base de dados em Inglês 32
4.4 Léxicos de Subjetividade 33
4.4.1 Léxicos em Português 33
4.4.2 Léxicos em Inglês 34
4.5 Setup dos Experimentos 35
4.6 Resultados 36
4.6.1 Resultados para as notícias em Português 37
4.6.2 Resultados para as notícias em Inglês 47
4.7 Discussão dos Resultados 54

5 Conclusões e Trabalhos Futuros **58**
5.1 Conclusões 58
5.2 Trabalhos Futuros 59

Lista de Figuras

2.1 Arquitetura básica para geração de embeddings utilizando word2vec e Skip Gram. O objetivo é utilizar os pesos da camada escondida que a rede neural consegue aprender como embeddings para as palavras recebidas como entrada. As probabilidades retornadas na saída não são utilizadas para a geração do espaço vetorial. 16

3.1 Taxonomia proposta por Wang et al. (2018), baseada nas categorias presentes no *"Truth-O-Meter"* do PolitiFacts e na taxonomia denominada SHPT (RASHKIN et al., 2017). 25

4.1 Boxplot para as distâncias semânticas das notícias falsas em Português. . . 38
4.2 Boxplot para as distâncias semânticas das notícias reais em Português. . . . 39
4.3 *Summary plot* gerado através do SHAP, exibindo o peso que as features exercem sobre a decisão de classificação do modelo. No eixo y, estão listadas as cinco features de subjetividade que formam a representação vetorial de um documento, considerando as médias de subjetividade por documento como features. No eixo x, estão os *shap values*, onde valores maiores que zero representam uma maior chance para a classificação da classe alvo (classe 1), que neste caso, são as notícias falsas. Valores negativos (menores que zero), representam uma maior chance para a classificação de notícias reais (classe 0). 46
4.4 *Summary plot* gerado através do SHAP exibindo as features mais relevantes para a classificação do modelo baseado em BoW e TFIDF. 48
4.5 Boxplot para as distâncias semânticas das notícias falsas em Inglês. 49
4.6 Boxplot para as distâncias semânticas das notícias reais em Inglês. 50

4.7 *Summary plot* gerado através do SHAP, exibindo o peso que as features exercem sobre a decisão de classificação do modelo utilizando as médias de subjetividade por documento como features para a classificação das notícias em Inglês. 53

4.8 *Summary plot* gerado através do SHAP exibindo as features mais relevantes para a classificação do modelo baseado em BoW e TFIDF para as notícias em Inglês. 57

Lista de Tabelas

2.1 Representação de três documentos utilizando o modelo Bag of Words. Na tabela, os documentos são representados por vetores, onde cada posição do vetor corresponde à ocorrência (1) de uma palavra no documento, ou sua ausência (0) no mesmo. O tamanho do vetor corresponde ao tamanho do vocabulário presente nos documentos. 13

4.1 Distribuição das notícias reais coletadas para o Português. 32

4.2 Estatísticas descritivas das distâncias semânticas para as notícias falsas em Português 37

4.3 Estatísticas descritivas das distâncias semânticas para as notícias reais em Português 38

4.4 Testes de hipótese comparando as distâncias semânticas das sentenças para cada léxico, considerando as notícias falsas e reais em Português. Os resultados apresentam o número de ensaios em que foi reportado uma diferença significativa (p-value < 0,05) entre as distâncias semânticas das notícias falsa e reais em Português (coluna "# H0 rejeitada"). Na segunda coluna ("H1 two-sided") é apresentado os resultados (p-values significativos) de ensaio tradicional (execução simples do teste), comparando as distâncias das sentenças falsas e reais, onde a hipótese alternativa consiste em considerar que ambas as amostras pertencem a distribuições diferentes. Na terceira coluna ("H1 (Falsa >Real)") o mesmo ensaio é executado, porém agora a hipótese alternativa (H1) consiste em afirmar que os valores das distâncias semânticas das sentenças das notícias falsas são maiores (menos subjetivas) que os valores das sentenças das notícias reais. A quarta coluna ("H1 (Falsa <Real)") realiza a mesma testagem, porém, considerando que a hipótese alternativa considera que as distâncias semânticas das sentenças das notícias falsas são menores (mais subjetivas) que as distâncias das sentenças das notícias reais. 40

4.5 Resultados médios da classificação de notícias falsas e reais utilizando as médias de subjetividade por documentos. Nesta representação, cada documento é representado por um vetor contendo cinco features, que consistem na média das distâncias reportadas pelo WMD em relação a cada um dos cinco léxicos de subjetividade utilizados para o Português. 41

4.6 Resultados médios da classificação de notícias falsas e reais utilizando os vetores subjetividade por sentenças. Nesta representação, cada documento é representado pelas distâncias semânticas de cada uma de suas sentenças relativa a um dos léxicos de subjetividade, considerando um tamanho máximo de 100 sentenças por documento. 41

4.7 Resultados médios da classificação entre notícias falsas e reais utilizando a representação clássica baseada em TFIDF para as notícias em Português. . . 42

4.8 Resultado médio da classificação de notícias falsas e reais utilizando os vetores subjetividade por sentenças, considerando o design domínio-cruzado. Neste design, os tópicos das notícias reais são variados entre os conjuntos de traino e teste. 43

4.9 Resultado médio da classificação de notícias falsas e reais utilizando os modelos baseados em features TFIDF, considerando o design de domínio-cruzado. Neste design, os tópicos das notícias reais são variados entre os conjuntos de treino e teste. 44

4.10 Estatísticas descritivas para as distâncias semânticas obtidas para as sentenças presentes na base de dados de notícias falsas em Inglês. 49

4.11 Estatísticas descritivas para as distâncias semânticas obtidas para as sentenças presentes na base de dados de notícias reais em Inglês. 49

4.12 Testes de hipótese comparando as distâncias semânticas das sentenças para cada léxico, considerando as notícias falsas e reais em Inglês. Nos resultados, é possível observar que para todos os léxicos utilizados, houve diferenças significativas (p-value < 0,05). Para estes casos, a maioria os testes demonstrou que as notícias falsas apresentaram distâncias semânticas menores que as notícias reais, o que demonstra uma maior similaridade semântica das notícias falsas com os léxicos de subjetividade, o que denota uma maior subjetividade das notícias falsas. 51

4.13 Resultado médio para a classificação de notícias falsas e reais em Inglês, considerando as médias de subjetividade por documento como features. . . 52

4.14 Resultado médio para a classificação de notícias falsas e reais em Inglês, considerando os vetores de subjetividade por sentenças como features. . . . 52

4.15 Resultados médios da classificação entre notícias falsas e reais utilizando a representação clássica baseada em TFIDF para as notícias em Inglês. 52

5.1 Cronograma inicial para a execução de atividades futuras desta pesquisa. . . 60

Capítulo 1

Introdução

Uma notícia falsa, ou *fake news*, é definida como um conteúdo jornalístico criado intencionalmente com o objetivo de transmitir uma informação falsa, buscando assim enganar a audiência (JR; LIM; LING, 2018; ALLCOTT; GENTZKOW, 2017a). Estas notícias usualmente tentam imitar o formato de uma notícia convencional. Porém, as notícias fabricadas com esse intuito não seguem os rigorosos processos editoriais de checagem de fatos, que garantem uma maior precisão da informação veiculada (LAZER et al., 2018).

A intensa disseminação de notícias falsas percebida nos últimos anos tem demonstrado a baixa credibilidade que veículos da grande mídia têm perante boa parcela da população (LAZER et al., 2018). Um indicador disso e um importante marco no tocante ao estudo de notícias falsas foi a eleição presidencial americana de 2016. Na ocasião, denúncias surgiram denotando uma possível intervenção Russa no processo eleitoral americano. Tal interferência teria ocorrido por meio da disseminação em massa, utilizando redes sociais, de notícias falsas relacionadas à então candidata do partido democrata, Hillary Clinton[1]. O processo eleitoral americano de 2016 também permitiu avaliar o impacto das redes sociais na disseminação de notícias falsas. Por exemplo, foi verificado que as notícias falsas mais populares foram disseminadas por meio do Facebook, inclusive superando o compartilhamento de notícias reais mais populares (SILVERMAN, 2016). E estas notícias, por sua vez, tendiam a favorecer Donald Trump na corrida eleitoral (SILVERMAN, 2016).

Zhou e Zafarani (2018) classificam as pesquisas no âmbito de notícias falsas considerando duas vertentes. A primeira vertente consiste em trabalhos que utilizam aspectos pre-

[1]https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html

sentes em redes sociais para a detecção de notícias falsas, como por exemplo, número de *likes* ou compartilhamentos. A segunda, considera trabalhos que utilizam apenas características textuais dos documentos. Grande parte das pesquisas que investigam as características textuais de notícias falsas consideram aspectos como frequência de ocorrência de determinadas palavras e tamanho dos documentos (AHMED; TRAORE; SAAD, 2017a; BOURGONJE; SCHNEIDER; REHM, 2017; HORNE; ADALI, 2017). Dentre estes trabalhos, muitos autores utilizam modelos baseados em Bag-of-Words (BoW), que consiste em uma representação textual onde os documentos são representados por vetores numéricos baseados no próprio vocabulário dos documentos. Tais modelos de classificação de notícias falsas reportam resultados expressivos, com acurácias atingindo cerca de 95% (KHAN et al., 2019). Porém, por serem baseados no vocabulário dos documentos utilizados no treinamento, esses modelos tendem a perder seu poder de generalização quando submetidos a classificação de documentos que abordam tópicos diferentes dos utilizados no conjunto de treino. Com base neste problema, apresentamos uma alternativa que, ao invés de utilizar diretamente os termos presentes nos documentos, se baseia nos níveis de subjetividade que estes termos podem carregar, considerando que notícias falsas e reais tenham diferentes níveis de subjetividade.

1.1 Motivação e Justificativa

A necessidade de modelos que permitam a detecção de notícias falsas de forma efetiva é clara, especialmente no momento atual, onde muitas destas notícias se disseminam pelas redes sociais. Um exemplo recente da disseminação de notícias falsas por meio de redes sociais aconteceu no último pleito eleitoral brasileiro, onde denúncias de disseminação em massa de boatos levaram a processos judiciais que ainda se encontram em tramitação na justiça brasileira[2].

Se, por um lado, as redes sociais permitem uma maior interação e participação política dos eleitores, por outro, estas mesmas redes favorecem a adesão dos mesmos a extremos políticos e ideológicos (LEE; SHIN; HONG, 2018). Este ambiente polarizado das redes sociais tende a favorecer a própria disseminação de notícias falsas (ALLCOTT; GENTZKOW, 2017b; FERRARA et al., 2016; MARCHI, 2012), bem como de boatos construídos com

[2]https://www.conjur.com.br/2019-out-17/tse-reabre-investigacao-uso-fake-news-massa

a intenção de enganar ou confundir. Outro exemplo de como a polarização política favorece a disseminação de notícias falsas pôde ser visto durante o processo de impeachment da ex-presidente Dilma Rouseff. Durante a semana de votação para o processo de seu impeachment, três das cinco notícias mais populares no Facebook eram falsas[3]. Corroborando com o exemplo descrito, a agência de checagem de fatos Lupa[4] mostrou que de Agosto a Outubro de 2018, durante o primeiro turno das eleições brasileiras, dez das mais populares notícias falsas tiveram cerca de 865.000 compartilhamentos apenas no Facebook[5].

Cenários como o descrito anteriormente demonstram a importância que as redes sociais têm ao se analisar notícias falsas. Porém, abordagens que utilizam marcadores sociais nas notícias (e.g. número de *likes* e compartilhamentos) acabam por apenas detectar estas notícias quando as mesmas já se propagaram pela rede. Visando contornar este problema, diferentes abordagens consideram a identificação de notícias falsas de forma precoce. Para tal, a análise precisa ser direcionada ao texto da notícia, representando assim, um desafio adicional, dado a ausência de informações como *likes* e compartilhamentos. Neste sentido, esta pesquisa tem como foco principal a análise textual dos documentos, permitindo assim, o estudo e detecção precoce de conteúdo enganoso.

Um caminho promissor, que permite revelar características textuais complexas das notícias, é a análise da subjetividade dos documentos. O estudo de subjetividade em notícias não é algo recente. Inclusive, normas jornalísticas preconizam que notícias devem prezar por uma forma objetiva, ou seja, priorizando uma linguagem que permita reportar fatos de forma imparcial, sem qualquer envolvimento com os fatos apresentados (SCHUDSON, 2001). Dentro deste conceito, a subjetividade seria uma forma de desvio da norma jornalística.

Apesar de ser uma área de estudo bastante explorada em outros domínios, como jornalismo e linguística, a subjetividade representa um grande desafio no tocante à sua análise automática, por meio e métodos computacionais. Apenas recentemente, os avanços de áreas como Processamento de Linguagem Natural (PLN), bem como a aplicabilidade de técnicas de aprendizagem profunda (*Deep Learning*) têm permitido que novos estudos adentrem em áreas antes inexploradas na Ciência da Computação.

Nesta pesquisa, por meio do estudo de subjetividade presentes em documentos de no-

[3]http://www.businessinsider.com/brazil-is-more-worried-about-fake-news-thanany-other-country-chart-2017-9
[4]https://piaui.folha.uol.com.br/lupa/
[5]https://piaui.folha.uol.com.br/lupa/2018/10/07/artigo-epoca-noticias-falsas-1-turno/

tícias falsas e reais, são construídos modelos que, ao invés de considerarem as palavras de forma literal (e.g. modelos baseados em BoW), se baseiam nos diferentes níveis de subjetividade que estes documentos podem trazer. Ainda como escopo desta pesquisa, e visando suprir uma carência de trabalhos que consideram a explicabilidade de modelos preditivos para classificação de notícias falsas, esta proposta também analisa os modelos do ponto de vista explicativo. Basicamente, os modelos construídos serão comparados, a nível de explicação de suas características de classificação (referidas neste trabalho com a nomenclatura mais usual de *features*), com modelos baseados em BoW. Em resumo, essas características de classificação consistem em um conjunto de informações que os modelos de Aprendizado de Máquina utilizam como base para o aprendizado automático. Este estudo a nível de explicabilidade tem como objetivo avaliar o real impacto que as *features* exercem sobre os modelos, e como este impacto pode ajudar a entender as características das notícias falsas.

A princípio, nesta proposta, o método de extração de subjetividade utilizado é uma adaptação da aplicação bem sucedida utilizada por Amorim, Cançado e Veloso (2018a), onde busca-se analisar níveis de subjetividade presentes nas correções de redações do ENEM. A abordagem utilizada tem como principal diferencial, o uso de distâncias semânticas calculadas utilizando *Word Movers Distance* (WMD) (KUSNER et al., 2015) entre um grupo de palavras (i.e. léxico) que expressa alguma ideia de subjetividade (e.g. sentimentos, argumentações, pressuposições), aqui denominados de léxicos de subjetividade, e os documentos textuais, que nesta pesquisa, são notícias falsas e reais. Basicamente, a ideia é que, quanto mais subjetivo seja um documento, menor será a "distância" semântica entre o tal documento e os léxicos de subjetividade. Para a análise proposta nesta pesquisa, o método é adaptado para a extração de subjetividade a nível de suas sentenças, o que permite uma análise mais localizada de trechos subjetivos nos documentos.

Esta pesquisa tem como principal contribuição, a análise focada em abordagens que buscam extrair características semânticas das notícias, permitindo assim, acessar possíveis elementos de subjetividade presentes em notícias falsas e reais. Na pesquisa, é proposto um conjunto de estratégias para tal análise.

1.2 Questões de Pesquisa

Esta pesquisa de doutorado tem como foco principal, a investigação de características de subjetividade em documentos de notícias falsas, tendo como base, distâncias semânticas existentes entre estes documentos e léxicos de subjetividade existentes na literatura. A pesquisa também busca avaliar a aplicabilidade do método de extração de subjetividade no tocante à detecção de notícias falsas emergentes, por meio da construção de modelos preditivos que utilizam os níveis de subjetividade encontrados como *features* para classificação dos documentos entre notícias falsas e reais. De forma a atender as demandas propostas por esta pesquisa, foram formuladas questões básicas de investigação:

QP1 A utilização de métodos de extração de subjetividade baseados em semântica permitem revelar diferenças significativas entre notícias falsas e reais?

H1-0 Métodos de extração de subjetividade baseados em semântica, não permitem diferenciar, de forma significativa, notícias falsas e reais.

H1-1 Métodos de extração de subjetividade baseados em semântica, permitem diferenciar, de forma significativa, notícias falsas e reais.

Caso a hipótese nula seja refutada, significa que a abordagem proposta para a análise de subjetividade pode ser aplicada no contexto de detecção precoce de notícias falsas. Com isso, a linha de estudo que utiliza espaços semânticos para a extração de subjetividade em documentos fica validada, estatisticamente.

QP2 É possível determinar, de forma significativa, que notícias falsas são mais subjetivas que notícias reais?

H2-0 Não é possível afirmar que notícias falsas são mais subjetivas que notícias reais.

H2-1 É possível afirmar que notícias falsas são mais subjetivas que notícias reais.

Esta questão de pesquisa parte da intuição de que notícias falsas seriam mais subjetivas que notícias reais. Com a refutação da hipótese nula, esta ideia geral poderá ser validada, ao menos nos cenários apresentados. Caso não seja, hipóteses interessantes de estudo ainda podem se abrir, considerando que notícias reais podem ser tão subjetivas quanto as falsas, ou até mesmo mais subjetivas em determinados casos.

QP3 É possível, com uso das *features* de subjetividade propostas, construir modelos de classificação mais generalizáveis em relação a modelos baseados em BoW?

H3-0 Não é possível construir modelos mais generalizáveis utilizando as *features* de subjetividade.

H3-1 É possível construir modelos mais generalizáveis utilizando as *features* de subjetividade.

Esta questão de pesquisa busca avaliar se modelos de classificação baseados nas *features* de subjetividade propostas podem gerar modelos que sejam mais generalizáveis, quando comparados com modelos baseados em BoW. Para tal avaliação, os modelos serão avaliados em cenários onde os assuntos das notícias variam entre os conjuntos de treino e teste dos modelos.

1.3 Objetivos

Esta seção apresenta os objetivos gerais desta proposta de pesquisa, bem como seus objetivos específicos.

1.3.1 Objetivo Geral

Caracterizar notícias falsas e reais por meio do seus níveis de subjetividade, permitindo assim, a construção de modelos preditivos que possam, utilizando características de subjetividade, classificar esses dois tipos de notícia.

1.3.2 Objetivos Específicos

- Levantar a literatura disponível relativa à classificação de notícias falsas, bem como a análise de subjetividade presente neste tipo de conteúdo.
- Utilizar e adaptar abordagens de extração de subjetividade baseados em análise semântica, e aplicá-las para o estudo de notícias falsas.
- Realizar uma análise de dados que permita a caracterização, com base em subjetivi-

dade, das notícias falsas e reais.

- Construir modelos preditivos baseados em subjetividade.
- Avaliar os modelos construídos, e seu posicionamento na literatura.
- Analisar a capacidade de explicação das *features* de subjetividade utilizadas, e compará-las com modelos que utilizam BoW como *features* principais.

1.4 Estrutura do Documento

O restante do documento está organizado da seguinte maneira. No Capítulo 2, são apresentados conceitos básicos que permeiam esta proposta de pesquisa, auxiliando assim, o entendimento de conceitos importantes. No Capítulo 3, é apresentada uma ampla revisão da literatura que abrange o tópico principal destra proposta, ou seja, caracterização e identificação de notícias falsas. No Capítulo 4, é apresentada uma análise exploratória dos dados de subjetividade, e como os documentos falsos e reais se caracterizam no tocante a este aspecto textual. No Capítulo 5, é apresentado experimentos e principais resultados reportados pelos modelos construídos. No Capítulo 6, são apresentadas as conclusões, bem como os delineamentos futuros desta pesquisa, bem como um cronograma detalhado para atividades futuras.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são abordados os principais tópicos que permeiam o escopo principal desta pesquisa. São descritos assuntos que abrangem desde a temática jornalística e conceitual acerca das notícias falsas, até conceitos voltados ao processamento de linguagem natural.

2.1 Notícias Falsas

Esta sub-seção apresenta conceitos básicos sobre Notícias Falsas, e suas diferentes definições presentes na literatura. Também será apresentada uma breve evolução histórica destas, bem como conceitos presentes no jornalismo e que são de fundamental importância para o entendimento deste fenômeno que, embora não sendo novo, emergiu de forma rápida e contundente em anos recentes.

2.1.1 Definição de Notícia Falsa

A definição de notícia falsa está longe de ser um consenso absoluto. Zhou e Zafarani (2018) apresentam duas definições para notícias falsas. A primeira, mais ampla, define uma notícia falsa como sendo uma declaração, uma fala, postagem ou qualquer outra forma de comunicação que seja essencialmente falsa. Esta definição torna o termo "notícia" mais amplo e generalizado, sendo esta comumente utilizada pela grande mídia.

A segunda definição para notícia falsa, sendo mais restritiva, considera uma notícia falsa como sendo uma publicação com características jornalísticas (i.e. notícia em formato de

artigo, contendo título, autor e corpo da notícia), porém, com o objetivo de enganar a audiência. Estas notícias costumam ser publicadas por veículos de comunicação (jornais ou blogs de notícias), porém, com o intuito verificado de desinformar ou enganar. Esta definição, por ser a mais amplamente utilizada em trabalhos relacionados ao tema (ALLCOTT; GENTZKOW, 2017b; SHU et al., 2017), é a definição adotada nesta proposta de pesquisa. Quanto possíveis tipos de notícias falsas, Jr, Lim e Ling (2018) definem os seguintes tipos:

- **Sátiras:** Tipo de notícia falsa mais comum, é baseado em uma linguagem humorística e exagerada. Neste tipo de notícia, o leitor percebe que o conteúdo não é verídico, mas sim, uma piada acerca de um fato real. O objetivo deste tipo de notícia é puramente de entretenimento.

- **Paródias:** Similar às sátiras, as paródias também fazem uso do humor para entreter o leitor, porém, se diferem das sátiras por utilizar informações fictícias (não sendo baseadas em um fato pré-existente). Neste tipo de documento, o leitor percebe o cunho humorístico da notícia.

- **Notícia Falsa:** Notícia falsa que apresenta informações sem base factual, e com a intenção deliberada de enganar. Ao contrário das sátiras e paródias, os autores deste tipo de notícia tentam criar um ambiente que imita grandes veículos de imprensa, para assim, passar uma suposta credibilidade ao material publicado.

- **Manipulação em Fotos:** Este tipo de conteúdo consiste na manipulação de imagens e vídeos com o objetivo de criar uma falsa narrativa. Softwares de edição de imagens e vídeos permitiram a popularização deste tipo de conteúdo, permitindo, inclusive, a adição de pequenas passagens de texto nas imagens.

- **Publicidade Enganosa:** Este tipo de conteúdo é caracterizado pela inserção de material publicitário em um jornal, em formato de notícia, porém, com o intuito de persuadir acerca de uma marca ou produto. Um exemplo deste tipo de material são os populares “clickbaits”, que visam persuadir o leitor a se direcionar para uma página comercial.

- ***Propaganda:*** Este tipo de material se caracteriza como notícias ou mesmo narrativas criadas por uma entidade política, cujo objetivo é influenciar a percepção pública

acerca de uma figura pública, organização ou governos. Este tipo de conteúdo tente a se parecer com uma publicidade enganosa, porém, este último está mais direcionado à venda de produtos ou ganho financeiro.

2.1.2 Objetividade vs Subjetividade em Notícias Falsas

Apesar de ter ganhado grande notoriedade nos últimos anos, especialmente após as eleições presidenciais americanas de 2016, as notícias falsas e suas consequências não são fenômenos novos. Rumores e contos fictícios, com intuito de enganar, sempre permearam a história humana, especialmente em sociedades baseadas em poder e ascensão social, até mesmo bem antes da invenção da imprensa. A era chamada "pós-imprensa", a partir do Renascimento, com o advento da imprensa escrita, veio a favorecer a disseminação de notícias para regiões mais distantes e de forma mais rápida, onde a forma verbal de disseminação de notícias antes não permitia. Este fato, consequentemente, favoreceu às aqueles que, dominando a leitura e escrita, podiam facilmente manipular a informação, exercendo poder sobre as populações menos letradas (BURKHARDT, 2017).

O conceito de objetividade veio surgir, como um elemento necessário ao jornalismo, apenas após o uso massivo da imprensa para fins de propaganda relativa à Primeira Guerra Mundial (LAZER et al., 2018). Tuchman (1993 apud HENRIQUES, 2016) entende que um bom produto jornalístico, guiado pelo conceito de objetividade, contém todos os esforços e estratégias que permitem, se não anular, minimizar qualquer viés subjetivo na notícia.

Hoje, as normas de objetividade jornalística são amplamente difundidas como base para o trabalho de investigação e apuração dos fatos, para assim, gerar uma notícia bem produzida. Contudo, é plenamente sabido que, diferentes veículos de notícias podem possuir diferentes visões acerca dos fatos e acontecimentos (e.g. veículos de notícias mais alinhados ao campo político de esquerda tenderão a ter visões sobre fatos que diferem de veículos mais alinhados à direita), o que pode colocar o conceito de objetividade jornalística em questionamento.

Dentro do aspecto histórico que permeia o jornalismo e a massiva disseminação de notícias falsas, está o surgimento e massificação da Internet. A Internet permitiu que diversos outros veículos de imprensa, que muitas vezes não seguem as normas de objetividade jornalística, se estabelecessem a um custo financeiro muito baixo, tornando-se assim, competidores de grandes veículos de imprensa. Como consequência, culmina-se nas eleições

presidenciais americanas de 2016, os mais baixos índices de confiança nos grandes veículos de imprensa, onde 51% dos Democratas e 14% dos Republicanos alegaram confiar na grande mídia americana (LAZER et al., 2018). Esse evento demonstrou pela primeira vez, e em larga escala, o poder e alcance que as notícias falsas podem ter, especialmente em um cenário de grande descrédito relacionado aos grandes meios de comunicação.

2.2 Processamento de Linguagem Natural

Esta seção apresenta um levantamento histórico da evolução do Processamento de Linguagem Natural (PLN) dentro da Ciência da Computação, bem como apresenta também os principais conceitos e técnicas utilizadas para o desenvolvimento desta pesquisa.

2.2.1 Evolução da PLN

A área de Processamento de Linguagem Natural busca investigar o uso de computadores para processar ou entender linguagens humanas (i.e. linguagem natural). Em termos gerais, a PLN busca modelar os processos cognitivos envolvidos no entendimento e produção das linguagens humanas (DENG; LIU, 2018).

Os primeiros estudos sobre PLN tiveram início ainda na década de 50, como a junção das áreas de Inteligência Artificial (IA) e Linguística (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). Deng e Liu (2018) descreve as três etapas de evolução da PLN. A primeira, denominada de "Racionalista", que perdurou até meados da década de 80, foi caracterizada pela construção de sistemas inteligentes baseados em regras manualmente definidas. Estas regras tinham como objetivo incorporar algum nível de conhecimento linguístico nestes sistemas. Já a segunda fase do processo de evolução da PLN, denominada de "Empiricista", teve como principal característica, o uso de conjuntos de dados, modelos de aprendizado de máquina e modelagens estatísticas para a construção de sistemas inteligentes. Esta etapa dominou os estudos relacionados à PLN até a década de 90, onde as abordagens desenvolvidas eram denominadas de empíricas, ou pragmáticas, por se basearem fundamentalmente em dados (i.e. exemplos) ao invés de regras pré-definidas (CHURCH; MERCER, 1993). Uma das principais conferências na área de PLN, a "*Empirical Methods in Natual Language Processing*" (EMNLP) demonstra, em seu próprio título, a importância que esta fase teve, e

ainda tem, no desenvolvimento da PLN.

Já a terceira etapa, descrita por Deng e Liu (2018) como "Deep Learning", representa o estado mais atual de desenvolvimento da PLN. Nos modelos de aprendizado de máquina tradicionais, as características, ou *features*, usadas nos modelos de predição são extraídas por humanos, sendo este um processo custoso e demorado. Modelos baseados em aprendizagem profunda, ou *Deep Learning*, permitem ir além destas limitações, proporcionando, por meio de redes neurais de múltiplas camadas, modelos com um grande poder de generalização, permitindo a representação de funções de alta complexidade (GOODFELLOW; BENGIO; COURVILLE, 2016). A atual disponibilidade de grandes volumes de dados torna estes modelos ainda mais poderosos, quando comparados com modelos de aprendizado de máquina tradicionais. Os principais avanços da atualidade, no campo de processamento de linguagens, estão vinculados a estes modelos de aprendizagem profunda. Porém, apesar da revolução trazida por modelos de aprendizagem profunda no âmbito da PLN, limitações destes modelos ainda representam grandes desafios. Entre os principais, está a falta de interpretabilidade destes modelos, gerando o chamado modelo "caixa preta", dificultando a interpretação e explicação destes modelos. Outra limitação importante é a necessidade de um grande volume de dados que estes modelos costumam utilizar para treinamento.

O desenvolvimento trazido pera terceira etapa de evolução da PLN abriu novos caminhos para a área, sobretudo em aplicações que buscam estudar relacionamentos semânticos em documentos. Um importante avanço neste sentido foi a utilização de *Word Embeddings* para extrair representações semânticas de palavras (MIKOLOV et al., 2013b; BENGIO et al., 2003) em um documento. Este tópico será especialmente coberto na próxima seção deste capítulo.

2.2.2 Bag of Words

Bag of Words (BoW) é um modelo de representação textual popularmente utilizado em atividades de Recuperação da Informação e classificação de textos (JOACHIMS, 1998; WANG et al., 2014). Neste modelo, um documento é representado por um vetor, onde o tamanho deste vetor é tamanho do vocabulário de todo o *corpus* utilizado para treinamento do modelo. Nesta representação, cada elemento do vetor representa, por exemplo, o número de ocorrências de uma dada palavra. A Tabela 2.1 apresenta um exemplo da utilização de BoW

para representação de documentos de texto.

Documento	**notícia**	**falsa**	**boato**	**real**
notícia falsa	1	1	0	0
boato	0	0	1	0
notícia real	1	0	0	1

Tabela 2.1: Representação de três documentos utilizando o modelo Bag of Words. Na tabela, os documentos são representados por vetores, onde cada posição do vetor corresponde à ocorrência (1) de uma palavra no documento, ou sua ausência (0) no mesmo. O tamanho do vetor corresponde ao tamanho do vocabulário presente nos documentos.

Esta técnica permite, de forma simples, representar documentos textuais de forma que estes possam ser utilizados por modelos preditivos, dado que estes, na maioria das vezes, exigem uma representação vetorial de tamanho fixo para treino e inferência. BoW tem sido um dos métodos de extração de *features* mais comum no escopo da PLN, onde cada palavra se torna uma *feature* que representa o documento.

Além de simples, a representação textual por meio de BoW também possui a característica de ser bastante flexível, pois permite a representação textual utilizando qualquer métrica que possa ser aplicada a nível de palavras ou termos. Por exemplo, a Tabela 2.1 considera a representação dos documentos utilizando o número de ocorrência dos termos presentes nos mesmos.

Representação BoW utilizando TFIDF

O *Term Frequency-Inverse Document Frequency* (TFIDF) consiste em uma medida estatística bastante comum, que permite destacar a relevância de uma palavra ou termo em um documento em relação a uma coleção de documentos. Por exemplo, para uma consulta, em um mecanismo de busca qualquer, em que se deseja retornar documentos por meio da seguinte consulta “notícias falsas de hoje”, é desejável que os termos “notícia”, “falsas” e “hoje” tenham mais relevância do que o termo “de”. O TFIDF permite que tal consulta seja relevante, pois além da frequência de um termo dentro do próprio documento, esta métrica também considera a frequência inversa deste termo em relação aos demais documentos. Em outras palavras, temos que:

- **TF** *Scoring* da frequência de um termo dentro de um documento.

- **IDF** *Scoring* de quão raro um termo é, em relação aos demais documentos.

Em outras palavras, utilizando o TFIDF, é possível distinguir termos que possam trazer informações mais úteis (i.e. palavras mais raras e relevantes) para um documento, em relação aos demais documentos. Formalmente, podemos definir TFIDF como:

$$w_{i,j} = tf_{i,j} * \log(\frac{N}{df_i}) \tag{2.1}$$

Onde temos que:

- $w_{i,j}$ peso TFIDF para o termo i no documento j.
- $tf_{i,j}$ número de ocorrências do termo i no documento j.
- df_i número de documentos que contém o termo i.
- N número total de documentos.

Assim como na representação BoW que utiliza a contagem de termos, exibida na Tabela 2.1, a representação utilizando TFIDF também tende a gerar vetores esparsos. Isto acontece pois apenas será atribuído um peso à posição do vetor que representar um termo presente no documento, de forma que a grande maioria dos demais termos, em não estando presentes no documento, terão o valor 0 atribuído a estes.

2.2.3 Word Embeddings

Em poucas palavras, *Word Embedding* (WE) (BENGIO et al., 2003) pode ser definido como uma representação vetorial densa, de n dimensões de uma dada palavra. Considerando-se uma base de dados suficientemente grande, esta representação favorece a captura de aspectos semânticos dos termos, permitindo, inclusive, o cálculo de similaridade semântica entre eles. Por exemplo, operações sobre os vetores “rei - homem + mulher” retornaria um vetor próximo ao representado por “rainha”.

A principal motivação, ao se criar uma representação como essa, é a de criar uma representação vetorial densa para palavras ou termos, que permita a captura de relacionamentos dentro do espaço vetorial. Estes relacionamentos podem ser de ordem de semântico, morfo-

lógico de contexto, ou qualquer outro relacionamento que possa ser apresentado dentro do corpus de criação da representação.

Atualmente, WE vêm sendo utilizado em diferentes cenários da PLN. Chen et al. (2013) avaliou algumas aplicações práticas de WE, destacando-se: (1) classificação de sentimentos; (2) identificação do gênero (masculino/feminino) de nomes próprios; (3) identificação de termos escritos no plural; (4) identificação de sinônimos e antônimos e (5) identificação de termos regionais entre Estados Unidos e Reino Unido. Estes exemplos apenas destacam algumas das possíveis aplicações de WE, porém, este conceito pode ser extrapolado para diversas áreas além do processamento de texto tradicional, como, por exemplo, em análise de DNA (LE et al., 2019) e recomendação de músicas (CHEN et al., 2016).

Dentre as diversas implementações, a mais popular é a *word2vec*, implementada por Mikolov et al. (2013a). Abordagens mais recentes, como BERT (DEVLIN et al., 2018) estão se tornando cada vez mais populares na literatura acadêmica. Apesar de diferentes estratégias algorítmicas para a criação de *embeddings*, todas têm como objetivo em comum: a criação de um espaço vetorial que permita a representação semântica de palavras, termos ou sentenças de um documento.

Neste trabalho, o *word2vec* será utilizado para a criação da representação vetorial dos termos textuais, utilizando o algoritmo Skip-Gram para a construção dos *embeddings*. Utilizando este algoritmo, a principal tarefa de uma rede neural utilizada para a geração de *embeddings* é prever as palavras ao redor do termo alvo, com base em uma janela pré-definida. Os pesos aprendidos pela rede neural constituem a representação vetorial que se deseja obter. A Figura 2.1 demonstra a ideia do algoritmo adotado.

O modelo recebe como entrada uma palavra, representada como um vetor, no formato de *one-hot encoding*. Esta forma de representação consiste em representar variáveis categóricas em formato de vetores binários, permitindo o uso destas variáveis para treinamento de modelos de aprendizagem profunda. O tamanho deste vetor é igual ao tamanho do vocabulário utilizado. Este vetor alimenta uma rede neural, onde esta retorna, por meio da camada de saída, as probabilidades das palavras próximas à palavra alvo. Porém, estas probabilidades serão desconsideradas, e apenas os pesos aprendidos a partir da camada escondida serão utilizados como representação vetorial.

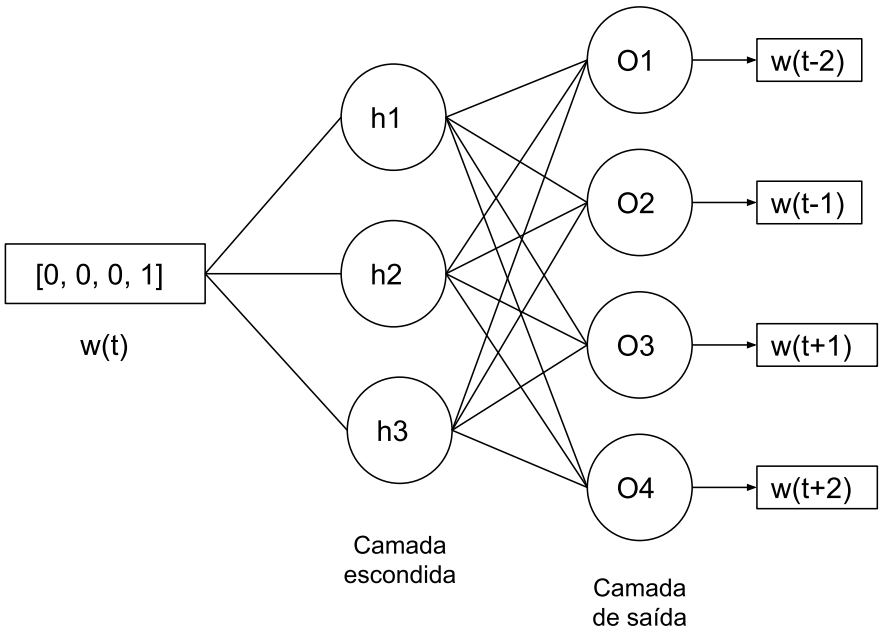

Figura 2.1: Arquitetura básica para geração de embeddings utilizando word2vec e Skip Gram. O objetivo é utilizar os pesos da camada escondida que a rede neural consegue aprender como embeddings para as palavras recebidas como entrada. As probabilidades retornadas na saída não são utilizadas para a geração do espaço vetorial.

2.2.4 Word Mover's Distance

Word Mover's Distance (WMD) é uma função que permite, por meio de WE, calcular uma "distância" entre dois documentos de texto (KUSNER et al., 2015). Esta métrica computa a distância entre dois documentos considerando a menor distância que as palavras de um documento, representadas em um espaço vetorial, precisam "viajar" para alcançar as palavras de um outro documento. Quanto menor a distância, mais parecidos são os documentos. O WMD permite, como consequência do uso de uma representação utilizando WE, computar esta distância considerando as possíveis nuances semânticas presentes nos documentos. Esta característica permite uma análise mais profunda sobre aspectos que estão vinculados à subjetividade de documentos, dado que a subjetividade está intimamente ligada à semântica do texto.

Formalmente, o WMD assume uma matriz de embeddings $X \in \mathbb{R}^{d \times n}$ para *n* palavras em um vocabulário onde $x_i \in \mathbb{R}^d$ é a representação de embedding da i^{th} palavra em um espaço dimensional *d*. O modelo também assume dois documentos *d* e *d'* representado como BoW normalizado. O WMD usa uma matriz de "fluxo" **T** para denotar o quanto uma palavra *i*

em um documento *d* precisa "viajar" para alcançar a palavra *j* no documento *d'*. A distância entre a palavra *i* e a palavra *j* se torna $\|x_i - x_j\|_2$. Logo, o método aprende a matriz de fluxo **T** com o objetivo de minimizar

$$\begin{aligned} \min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|x_i - x_j\|_2 \quad sujeito\ a, \\ \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i, \sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \ \ \forall i, j \end{aligned} \tag{2.2}$$

Como o WMD retorna uma métrica de distância entre dois documentos, quanto menor este valor, mais similar, dentro do espaço vetorial utilizado, seriam estes documentos. O WMD usualmente reporta um valor entre 0 e 1 como métrica de distância.

2.2.5 Valores SHAP

Buscando entender de forma mais aprofundada o contexto e explicação dos modelos de classificação gerados, é utilizado nesta pesquisa os valores reportados pelo SHAP (LUNDBERG; LEE, 2017). SHAP se baseia nos chamados "Valores SHAP", que são valores que representam a importância de cada feature para a predição de modelos de Aprendizado de Máquina. Por exemplo, features que ao terem seus valores modificados impactam significativamente na classificação de um modelo, são consideradas features mais relevantes. Esta análise permite a obtenção de um entendimento mais objetivo das decisões de classificação dos modelos implementados, gerando *insights* sobre o problema abordado. Com base nos Valores SHAP, é possível gerar visualizações que ajudam no entendimento dos modelos de classificação gerados nesta pesquisa.

Capítulo 3

Trabalhos Relacionados

Este capítulo descreve os principais trabalhos relacionados ao estudo e identificação de notícias falsas. O foco deste levantamento bibliográfico é em trabalhos que usam técnicas computacionais para a identificação destas notícias.

3.1 Identificação de Notícias Falsas

Estudos abordando a disseminação de notícias falsas são vastos. Porém, apenas recentemente, dado os avanços na NLP, na mineração de redes sociais e nos métodos de aprendizado de máquina, está sendo possível um entendimento mais profundo sobre as características das notícias falsas, e como os usuários interagem com com elas.

Shu et al. (2017) apresentam uma revisão da literatura sobre trabalhos relacionados à problemática de notícias falsas, dentro da perspectiva de mineração de dados, mas também considerando aspectos psicológicos e sociais envolvidos presentes no contexto. Os autores realizam uma ampla análise de trabalhos, sendo relevante, em especial, a descrição de possíveis pontos a serem considerados para futuras pesquisas, entre eles:

- Disponibilização de mais dados para *benchmark* de notícias falsas;
- Aprimoramentos para detecção de notícias falsas de forma precoce, antes que estas se disseminem em redes sociais;
- Realização de estudos quantitativos sobre aspectos psicológicos relacionados à notícias falsas;

- Condução de pesquisas com ênfase na análise de intenções por trás das notícias falsas, não apenas em detectá-las;
- Extração de características sobre notícias falsas precisa ser mais estudado (*word embedding e deep learning*);
- Criação de modelos mais complexos para tratamento de notícias falsas;
- Desenvolvimento de novas bases de dados de notícias falsas;
- Estudos sobre como conter a propagação de notícias falsas em redes sociais.

Além da utilização de texto como forma para detecção de notícias falsas (foco deste trabalho), o engajamento social, que consiste no nível de interação de usuários em redes sociais, também tem se mostrado importante para a detecção deste tipo de conteúdo. Janze e Risius (2017) investigam como notícias falsas disseminadas nas redes sociais podem ser identificadas por meio de aspectos cognitivos, visuais, afetivos e comportamentais presentes nas postagens de notícias falsas no Facebook. Para os aspectos cognitivos, os autores consideram características presentes diretamente nos textos, como:

- Quantidade de palavras em um documento.
- Polaridade (sentimento) do documento.
- O *loudness*, considerando termos capitalizados, bem como expressões contendo símbolos de exclamação, asterisco e *underline*.
- Cálculo do chamado *readability* de um documento, utilizando a métrica Flesch–Kincaid (KINCAID et al., 1975)

Para os aspectos visuais, os autores utilizam os níveis de brilho as imagens presentes nas postagens de notícias falsas e a presença ou não de rostos nas imagens. Para a consideração de aspectos afetivos, são utilizado marcadores reportados pelos leitores, que incluem: *like*, *love*, *wow*, *haha*, *sad*, *angry*. Para as *features* de comportamento, os autores se baseiam no número de compartilhamento das notícias, bem como nos comentários destas. Com as *features* descritas, os autores conseguem uma acurácia de 80%, utilizando SVM, para a detecção

de notícias falsas, em um cenário de dados balanceados. Abordagens como estas que consideram aspectos de postagens em redes sociais, apesar de reportarem bons resultados, têm a desvantagem de apenas conseguir identificar as notícias falsas após estas já terem iniciado seu ciclo de disseminação nas redes.

Detecção de notícias falsas baseando-se apenas em características textuais é ainda mais desafiador, dado ao fato de que estas notícias são escritas com o objetivo determinado de enganar os leitores, fazendo com que muitas destas notícias se pareçam com notícias reais. Esta característica acaba muitas fezes fazendo com que até mesmo profissionais da mídia sejam enganados por notícias falsas [1]. Muitas abordagens, que reportam resultados bastante expressivos, se baseiam em características lexicais dos textos. Abordagens utilizando BoW e outras características textuais simples como o tamanho dos documentos são comuns.

Ahmed, Traore e Saad (2017b) Propõem modelos para detecção de notícias falsas utilizando análises baseadas em n-gramas, que consistem em uma sequência contínua de n itens em um texto, como por exemplo, palavras ou um conjunto de palavras. Os autores comparam quais modelos apresentam um melhor desempenho dentro do escopo de notícias falsas. Os autores avaliaram diversos modelos, incluindo *Stochastic Gradient Descent, Support Vector Machines, Linear Support Vector Machines, K-Nearest Neighbour* e Arvores de Decisão. Foi considerado, como *features* para os modelos, vetores TFIDF, considerando uni-gramas, bi-gramas, tri-gramas e tetra-gramas. Os melhores resultados foram encontrados quando considerados o *Linear Support Vector Machines* utilizando uni-gramas, reportando uma acurácia de 92%.

Horne e Adali (2017) executam um estudo mostrando que notícias falsas são mais similares, em sua estrutura, com textos satíricos, e que estes textos são direcionados para usuários que se restringem a ler os títulos dos artigos, ao invés de ler o conteúdo na íntegra. Os autores se baseiam em *features* lexicais, como o número de ocorrência de palavras emotivas, características relacionadas à complexidade de leitura dos documentos (i.e. *readability*) e características de estilo textual, como por exemplo, a quantidade de verbos no documento. Os autores identificaram que é mais difícil classificar entre notícias falsas e sátiras. Também foi verificado que as features mais relevantes para a classificação de notícias falsas foram: número de substantivos, diversidade léxica (*Type-Token Ratio*), contagem de palavras e

[1] https://ijnet.org/en/story/how-journalists-can-avoid-being-manipulated-trolls-seeking-spread-disinformation

quantidade de citações. Os autores conseguiram uma acurácia de 71% para classificação de notícias falsas utilizando o corpo da notícia e 78% utilizando apenas os títulos. Os autores também destacam a necessidade de mais datasets para classificação de notícias falsas, bem como a construção de modelos não-supervisionados.

Pérez-Rosas et al. (2018) consideram características textuais como n-gramas, pontuação, termos denotando características psico-linguísticas, legibilidade (*readability*) e características sintáticas (gramáticas livres de contexto) para realizar a classificação de notícias falsas. Utilizando um total de 2.131 características de classificação, os autores conseguem uma acurácia de 74% para classificação de notícias falsas. Também é considerado um experimento realizando um cruzamento de bases de dados, onde o modelo utilizado (SVM) é treinado em uma base de noticias e testado em outra diferente, reportando acurácias acurácia média de 53%, demonstrando que o modelo não consegue generalizar bem para outros domínios de notícias falsas.

Utilizando notícias falsas em português, Monteiro et al. (2018) também executam uma classificação textual considerando um dataset composto por notícias de política, celebridades, notícias do cotidiano, tecnologia, economia e religião. Os autores utilizam *features* semelhantes às utilizadas nos trabalhos de Pérez-Rosas et al. (2018) e Horne e Adali (2017), considerando POS tags, BoW, e classes semânticas reportadas pelo *Linguistic Inquiry and Word Count* (LIWC[2]) (PENNEBAKER et al., 2015). Os autores utilizam também o SVM para classificação, reportando uma acurácia de 89%. Porém, esta acurácia apenas é atingida quando as *features* BoW são utilizadas em conjunto com as demais. Ao utilizar apenas BoW, os autores já reportam uma acurácia de 88%, demonstrando que estas *features* são as grandes responsáveis pelos resultados obtidos.

Rashkin et al. (2017) vai além de características lexicais e focam em aspectos estilísticos presentes na escrita dos documentos. Os autores comparam documentos de notícias reais com outras notícias de três categorias: propaganda, boatos e sátiras. O objetivo do trabalho é buscar o entendimento de notícias com conteúdo não confiável. Eles investigam a frequência de ocorrência de termos específicos, presentes em alguns léxicos já utilizados na literatura. Os léxicos utilizados são: *lying*, *subjective* ou *sentimental*, *hedging* e *intensity*. Entre os principais achados da pesquisa, estão que pronomes pessoais, bem como superlativos e ad-

[2]https://liwc.wpengine.com/

vérbios modais são utilizados com mais frequência em notícias falsas, quando comparadas com notícias reais. Os autores também tentam classificar as notícias entre reais, sátiras, boatos e propaganda utilizando a representação de BoW, considerando os léxicos utilizados. Nestas classificações, é reportado um F1 *score* de 65%.

No trabalho de Wang (2017), é apresentado um novo conjunto de dados para a detecção de notícias falsas, denominado de *LIAR*. Este dataset é composto por trechos falsos coletados do *PolitiFact*[3], que consiste em uma plataforma de checagem de fatos (i.e. *fact-checking*), sendo composto por trechos de notícias, bem como menções em televisão, rádio e redes sociais. Os dados são classificados como: *pants-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, e *true*. O autor ainda propõe um modelo híbrido, baseado em redes neurais convolucionais (CNN) para a classificação dos trechos avaliados. Como resultados utilizando o dataset proposto, o autor obtém, como melhor resultado, uma acurácia de apenas 27%, utilizando o modelo híbrido, baseado em CNN.

Ruchansky, Seo e Liu (2017) propõem um modelo que captura as três principais características utilizadas para o estudo de notícias falsas, que são: texto, a resposta dos usuários e a fonte. Os autores também utilizam abordagem baseada em *Deep Learning*, capturando a evolução temporal dos documentos. Em termos gerais, o modelo proposto, denominado de CSI, é baseado em três módulos: *Capture*, que é baseado em *Long-short Term Memory* (LSTM) para a captura de padrões temporais relacionados às atividades dos usuários sobre um dado artigo nas redes sociais, bem como características presentes no texto; *Score*, que aprende características baseadas no comportamento dos usuários na rede; e o módulo *Integrate*, que integra os dois primeiros módulos, para executar a classificação das notícias entre falsas e reais. Utilizando o modelo proposto, os autores conseguem obter acurácias na ordem de até 95% para classificação de notícias falsas.

3.2 Identificação de Notícias Falsas utilizando subjetividade

A análise de subjetividade em documentos de texto não é algo novo, porém, apenas recentemente, com os avanços na NLP, foi possível utilizar técnicas mais robustas para tal análise.

[3] https://www.politifact.com/

Aker et al. (2019) apresenta uma análise que compara a extração de *scores* de subjetividade automáticos, a nível de sentenças, com anotações manuais de subjetividade, por meio de voluntários, nos mesmos documentos. Para a análise automática, os autores consideram o cálculo de subjetividade por sentença, gerando assim, uma média da subjetividade por documento. Para tal análise, os autores utilizam a biblioteca *Pattern Web Mining Package*[4]. Para a análise manual, voluntários atribuem *scores* de subjetividade para todo o documento. Como resultado, os autores descrevem que, ao utilizar a metodologia de extração automática de subjetividade, foram atribuídos, em média, uma pontuação de 33,54 para notícias falsas e 30,73 para notícias reais. Estes resultados não apresentaram diferenças significativas. Porém, ao considerar a avaliação humana, foram obtidos resultados médios de 68,10 para notícias falsas e 41,24 para notícias reais. Segundo os autores, modelos que consideram o cálculo de subjetividade a nível de sentença, reportando uma média das sentenças para um documento tendem a perder informações quanto a sentenças específicas no documento, que contenham grande subjetividade. Estas mesmas sentenças, quando detectadas por humanos, produzem uma pontuação mais significativa de subjetividade. O estudo demonstra a dificuldade de se calcular, de forma automática, *scores* de subjetividade, destacando a necessidade de estudos que abordem, de forma eficiente, este problema.

No estudo realizado por Reis et al. (2019), os autores buscam analisar um grande conjunto de features, com o objetivo de identificar como possíveis combinações destas features podem contribuir para a detecção de notícias falsas. Para tal, os autores constroem aproximadamente 300.000 modelos, utilizando uma seleção randômica de features para cada um deles. As features utilizadas são compostas por: características textuais, estruturas de linguagens, características lexicais, características psicolinguísticas, estrutura semântica, subjetividade, viés em notícias, credibilidade da fonte, localização por meio de endereço IP, engajamento em redes sociais e padrões temporais. Apesar da vasta variabilidade das features utilizadas, apenas 2,2% dos modelos gerados obtiveram um desempenho aceitável (ROC-AUC >= 0,85), o que demonstra a real dificuldade de se identificar notícias falsas de forma automática. Esta dificuldade muitas vezes não se torna aparente, dado ao fato de que muitos modelos que reportam resultados expressivos apenas o reportam para um determinado conjunto de dados, não sendo generalizáveis o suficiente para efetivamente classificar notícias falsas. Adicional-

[4]https://github.com/pattern3/pattern

mente, um grande diferencial do trabalho é a exploração da explicabilidade dos modelos, a qual permite identificar a importância de determinadas features no processo de classificação.

Em seu trabalho, Volkova et al. (2017) constroem modelos preditivos para classificar 130 mil postagens no twitter como "suspeitos" ou "verificados", bem como classificar quatro sub-tipos de de notícias suspeitas, sendo estes: sátiras, boatos, *clickbait* e *propaganda*. Para adicionar o componente de subjetividade, os autores utilizam léxicos já presentes na literatura, e demonstram que tweets verificados possuem menos marcadores de subjetividade. Sátiras também parecem utilizar mais marcadores de subjetividade, ao serem comparadas com os outros tipos de documentos. Utilizando todas as features, que compreendem marcadores de subjetividade, viés, psico-linguísticos e fundamentos morais, bem como características derivadas da própria rede social, bem como do texto das postagens, os autores conseguem uma acurácia de 95% para classificação de postagens suspeitas e verificadas.

Wang et al. (2018) exploram granularidades mais refinadas de conteúdos possivelmente enganosos presentes no twitter, considerando uma nova taxonomia, desenvolvida a partir do sistema de classificação de conteúdo do Politifact (i.e. *"Truth-O-Meter"*) e da taxonomia já desenvolvida por Rashkin et al. (2017), denominada de SHPT. A Figura 3.1 exibe a nova taxonomia proposta. Os nomes originais das classificações foram mantidas, preservando a consistência nominal destas. No trabalho, além de features relacionadas ao conteúdo presente na rede social (e.g. como entidades nomeadas, texto do tweet que referencia a notícia), os autores também utilizam léxicos de sentimento e subjetividade, modelados como vetores TFIDF. Como resultado, os autores reportam uma acurácia máxima, utilizando regressão logística, de 98% utilizando a taxonomia SHPT, porém, ao utilizarem a nova taxonomia, o resultado máximo reportado foi de 30%. As features de subjetividade utilizadas isoladamente tiveram, respectivamente, 87% e 24% de acurácia.

Zhou et al. (2019) buscam avaliar diferentes nuances das notícias falsas, buscando aprimorar sua detecção com o objetivo de identificá-las antes de sua propagação pelas redes sociais (i.e. *Early Detection*), ou seja, detectá-las apenas pelo seu conteúdo textual. Para tal, os autores investigam as notícias sob diferentes aspectos, sendo estes: nível lexical, sintaxe, semântico e a nível de discurso. Para a análise, são utilizadas diversas características textuais, entre elas: BoW, POS tags, termos psicolinguísticos (LIWC), legibilidade (*readability*), polaridade de sentimentos, diversidade de palavras, valores quantitativos (caracteres, pala-

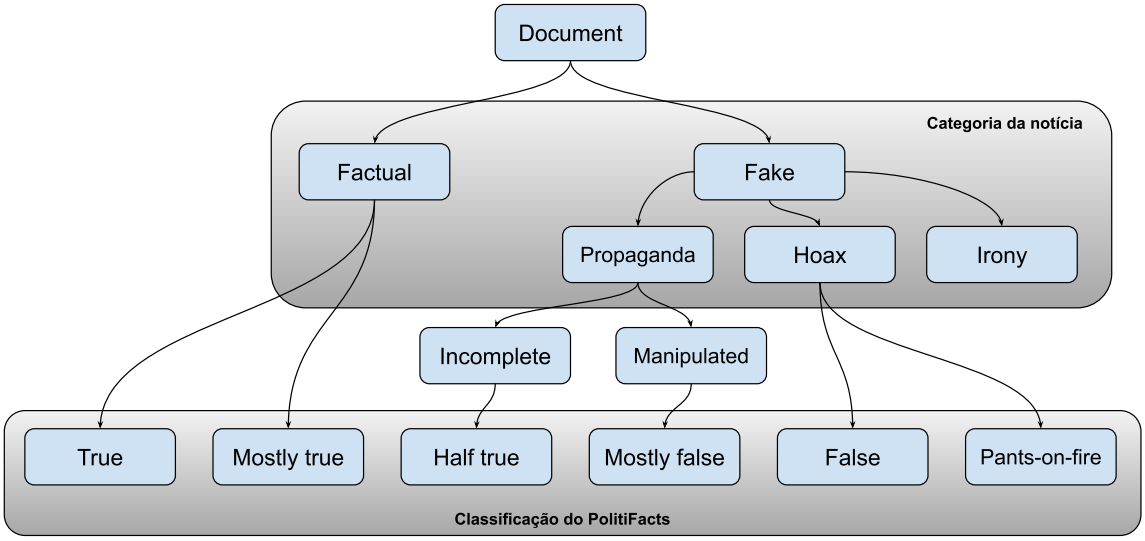

Figura 3.1: Taxonomia proposta por Wang et al. (2018), baseada nas categorias presentes no *"Truth-O-Meter"* do PolitiFacts e na taxonomia denominada SHPT (RASHKIN et al., 2017).

vras, sentenças e parágrafos), relacionamento retórico entre sentenças, ou *Rhetoric Structure Theory* (RST) e também subjetividade. O autores conseguem elencar a relevância destas características para a classificação de notícias falsas, onde, em primeiro lugar, estão as features que denotam a diversidade de palavras e valores quantitativos (i.e. caracteres, palavras, sentenças e parágrafos). Em segundo, estão features que denotam processos cognitivos e subjetividade. Em terceiro, estão features que expressam informalidade e sentimentos.

3.3 Posicionamento desta pesquisa em relação aos trabalhos relacionados

A Tabela 3.3 exibe as principais características dos trabalhos relacionados descritos neste capítulo. Também é apresentado, na última linha, a caracterização desta própria pesquisa.

	Features Textuais	Redes Sociais	Multilingual	Subjetividade	Explicabilidade	Deep Learning
Janze e Risius (2017)		x				
Ahmed, Traore e Saad (2017b)	x					
Horne e Adali (2017)	x					
Pérez-Rosas et al. (2018)	x					
Monteiro et al. (2018)	x					
Rashkin et al. (2017)	x					x
Wang (2017)	x					x
Ruchansky, Seo e Liu (2017)	x	x				x
Aker et al. (2019)	x			x		
Reis et al. (2019)	x	x		x	x	
Volkova et al. (2017)	x	x		x		x
Wang et al. (2018)	x	x		x		
Zhou et al. (2019)	x			x		
Proposta	**x**		**x**	**x**	**x**	**x**

As colunas consideradas para a construção da tabela são as que seguem:

- **Features Textuais:** Trabalhos que utilizam características textuais para a identificação/estudo de notícias falsas;
- **Redes Sociais:** Uso de características oriundas de redes sociais (e.g. curtidas e *likes*);
- **Multilingual:** Estudo que considera análise de noícias falsas em mais de um idioma;
- **Subjetividade:** Uso do subjetividade para caracterização e/ou classificação de notícias falsas;
- **Explicabilidade:** Uso de técnicas que permitam a explicabilidade de modelos preditivos no contexto de notícias falsas;
- **Deep Learning:** Uso de modelos e/ou técnicas de envolvam aprendizagem profunda no estudo de notícias falsas;

Esta pesquisa endereça cinco dos seis aspectos apresentados na Tabela 3.3. Neste trabalho, o foco consiste na análise de características textuais presentes nas notícias, abrindo mão de features provenientes de redes sociais. O foco unicamente em características textuais favorece a detecção precoce das notícias falsas, antes mesmo que estas possam se disseminar em redes sociais, gerando *likes* e compartilhamentos. Também é proposto o estudo de notícias falsas considerando as línguas portuguesa e inglesa. Esta abordagem favorece o entendimento das possíveis diferenças e dificuldades em se estudar o fenômeno de notícias falsas, considerando a ótica de duas línguas que apresentam estruturas diferentes. A quarta coluna da tabela, que consiste na análise de subjetividade, representa o foco principal desta pesquisa. O uso de subjetividade no âmbito de notícias falsas representa um problema difícil, que ainda necessita de grandes avanços. Porém, o estudo da subjetividade no âmbito de notícias falsas apresenta grande relevância, pois consiste em um dos grandes marcadores de autenticidade de um texto jornalístico de qualidade, onde estes tendem a ser mais objetivos.

No quesito de explicabilidade, é notável, ao se visualizar a própria tabela, a necessidade de um maior entendimento sobre como determinadas características exercem influência em modelos preditivos. Para tal, esta pesquisa também realiza uma análise mais profunda acerca

da explicabilidade dos modelos implementados. No contexto de *Deep Learning*, são utilizadas representações que derivam de redes neurais (*embeddings*) e, futuramente, modelos preditivos baseados em *Deep Learning*.

Capítulo 4

Análise de Subjetividade por Meio de Distâncias Semânticas

Neste capítulo, serão abordados aspectos relacionados à metodologia de pesquisa, como característica dos dados utilizados, design dos experimentos, e principais resultados encontrados. A primeira subseção deste capítulo irá tratar dos principais objetivos dos experimentos utilizados, apresentando um design geral destes. A segunda subseção irá detalhar as principais características das bases de dados utilizadas. A terceira subseção irá apresentar os principais resultados encontrados por meio dos experimentos executados.

4.1 Metodologia Geral de Experimentação

O principal objetivo dos experimentos executados é de verificar a viabilidade do uso de subjetividade extraída por meio de distâncias semânticas entre documentos e léxicos de subjetividade, para o estudo e classificação de notícias falsas.

Para realizar esta avaliação, as features de subjetividade aqui propostas serão empregadas em experimentos de classificação considerando notícias falsas e reais, utilizando dois cenários distintos de processamento destas features. Os métodos de processamento das features serão descritos nas subseções seguintes. Os modelos gerados serão comparados com um baseline utilizando features baseadas em TFIDF (BoW) dado que, dentre as técnicas que se baseiam em abordagens clássicas de Aprendizado de Máquina, estes modelos apresentam os melhores resultados para classificação de notícias falsas. Nesta etapa da pesquisa, foram

utilizados modelos clássicos de Aprendizado de Máquina devido à sua melhor interpretação e simplicidade, quando comparado a modelos de *Deep Learning*.

Adicionalmente, os experimentos serão executados para documentos escritos em Português e Inglês, trazendo assim, diferentes perspectivas que por vezes se mostram inerentes a uma linguagem específica.

4.2 Processamento das Características de Subjetividade

Esta seção descreve os dois métodos de processamento das características (i.e. *features*) de subjetividade utilizadas nesta pesquisa. Modelos de classificação serão construídos utilizando estas features para a classificação de notícias falsas e reais.

4.2.1 Média de Subjetividade por Documentos

Neste cenário de processamento de features, cada documento é representado pela média da distância semântica entre cada sentença do próprio documento x e o léxico de subjetividade utilizado. Para tal, esta pesquisa apresenta o conceito de segmentação de subjetividade por sentenças, onde é construído um vetor formado pelas distâncias semânticas entre cada sentença de um documento em relação a um léxico de subjetividade. Este vetor base é utilizado para o cálculo das médias de subjetividade para cada léxico. Desta forma, cada documento é representado por um vetor $v_n = \{v_1, v_2, v_3, ..., v_d\}$ considerando d diferentes léxicos para representar as nuances de subjetividade. Logo, cada distância entre um documento e um dado léxico l é calculado como:

$$\frac{1}{n}\sum_{i=1}^{n} WMD(x_i, l) \tag{4.1}$$

Onde para um dado documento x, calcula-se as distâncias semânticas utilizando o WMD para as n sentenças do documento, gerando uma média que representa a distância do documento x para o léxico l.

4.2.2 Vetores de Subjetividade por Sentenças

Neste cenário, as features são tratadas considerando a distância de cada uma das sentenças do documento em relação a um dado léxico l, porém, aqui não há uma extração de médias, mas os valores das distâncias são utilizados diretamente para cada sentença. Assim, um documento contendo n sentenças será representado por um vetor que contém todas as distâncias de cada uma das suas sentenças em relação ao léxico de subjetividade. Dada a variabilidade no tamanho dos documentos quanto ao número de sentenças, e a necessidade de uniformização dos vetores de features quanto às suas dimensões para a realização dos treinamentos dos modelos, os vetores gerados são "completados" com a média das distâncias. Este processo de preenchimento de um vetor com algum valor pré-definido é usualmente denominado de *padding*. Para tal, é estabelecido um limite máximo de 100 sentenças por documento para todos os experimentos. Desta forma, um documento contendo dez sentenças será representado por um vetor contendo dez distâncias semânticas e o restante do vetor será completado com a média destas dez distâncias.

4.3 Base de Dados

Esta seção descreve as bases de dados utilizadas nos experimentos realizados. Nesta subseção, essas bases são subdivididos de acordo com cada idioma utilizado, sendo estes o Português e o Inglês.

4.3.1 Base de Dados em Português

A Base de dados de notícias reais em Português foi coletado de dois grandes veículos de notícias do Brasil, sendo eles Estadão[1] e Folha de S. Paulo [2]. Como estes dois veículos de notícias são reconhecidos como grandes meios de comunicação e veiculação deste tipo de conteúdo, as notícias coletadas destes portais são consideradas como verdadeiras. Foram coletadas um total de 207.914 notícias destes dois portais relativas aos anos de 2014 a 2017. Esta faixa temporal foi utilizada por permitir a aquisição de um maior número de notícias. As notícias foram divididas entre quatro categorias: Política, Esportes, Cultura e Economia.

[1]<www.estadao.com.br> [2]<https://www.folha.uol.com.br/>

Domínio	Estadao	FolhaSP	Total	%
Política	24,638	30,765	55,403	26.6
Esportes	31,692	31,908	63,600	30.5
Economia	20,512	30,412	50,924	24.4
Cultura	15,456	22,531	37,987	18.2

Tabela 4.1: Distribuição das notícias reais coletadas para o Português.

A Tabela 4.1 exibe a distribuição dessas notícias.

A base de dados de notícias falsas para os experimentos em Português é composto por notícias falsas verificadas (fact-checking) que foram amplamente disseminadas no Brasil durante os anos de 2010 a 2017. As notícias foram coletadas manualmente de dois dos serviços mais populares de fact-checking no Brasil, o eFarsas[3] e Boatos[4]. Estes dois serviços rastreiam e verificam as notícias falsas mais populares em um dado momento, realizando uma investigação em cada documento, para atestar seu caráter de veracidade ou falsidade. Foram coletadas 121 notícias comprovadamente falsas destes dois veículos, totalizando mais de 40 fontes distintas das notícias. Esta base de dados, apesar de reduzida, tem a importante propriedade de ser formado por documentos que tiveram uma grande disseminação no Brasil (redes sociais e Web), significando que, de fato, elas enganaram muitos leitores. Esta base de dados também tem uma importante característica de ser advindo de diversas fontes diferentes, trazendo assim, uma boa representatividade no que tange às fontes das notícias. A temática das notícias falsas desta base de dados é predominantemente política.

4.3.2 Base de dados em Inglês

As notícias reais para o Inglês foram coletadas do popular conjunto de dados disponível publicamente no Kaggle, denominado "All the News"[5]. As notícias utilizadas foram publicadas durante os anos de 2016 e 2017. Desta base de dados, foram coletadas notícias do The Guardian (1.798 documentos), New York Times (1.598 documentos) e 2.598 documentos da CNN. Não há uma divisão específica quanto ao tema das notícias, porém, estas foram coletadas pelos autores a partir das notícias presentes nas capas dos referidos jornais, sendo estas, as principais notícias dos veículos.

[3]<http://www.e-farsas.com/> [4]<http://www.boatos.org/>
[5]<https://www.kaggle.com/snapcrack/all-the-news/version/4>

As notícias falsas para o Inglês foram compiladas do trabalho de Asr e Taboada (2019), correspondendo a notícias falsas de política oriundas do site de fact-checking Snopes[6] (103 documentos das categorias *fake e mostly-fake*), notícias falsas de política compiladas por Horne e Adali (2017) (75 documentos), notícias falsas coletadas pelo BuzzFeed[7] (41 documentos) e o BSDetector[8] (1425 documentos).

4.4 Léxicos de Subjetividade

Esta subseção descreve os agrupamentos de palavras (i.e. léxicos) que contém termos que representam algum nível de subjetividade, aqui denominados de léxicos de subjetividade. Nesta seção, a descrição dos léxicos está subdividida em léxicos em Português e em Inglês.

4.4.1 Léxicos em Português

Os léxicos em Português utilizados nesta pesquisa foram inicialmente propostos no trabalho de Amorim, Cançado e Veloso (2018b), onde os autores utilizam os léxicos para análise de subjetividade em correções de redações do ENEM. Os léxicos foram desenvolvidos por linguistas, tendo como base os trabalhos de Recasens, Danescu-Niculescu-Mizil e Jurafsky (2013) e Page (1967), permitindo o desenvolvimento de léxicos de subjetividade em Português. Este conjunto de léxicos compreende cinco diferentes dimensões de subjetividade, onde para cada um desses cinco léxicos, existem palavras que representam uma das cinco dimensões que segue:

- Argumentação: marcadores de discurso argumentativo. Este léxico também inclui algumas expressões compostas que expressam argumentação. Exemplos: "como consequência", "de certa forma", "em vez de";

- Pressuposição: termos que estão relacionados com a presunção/pressuposição prévia de que algo é verdadeiro, mesmo que de fato não seja. Exemplos: "demonstrar", "entender", "continuar a";

[6]<https://github.com/sfu-discourse-lab/Misinformation_detection>
[7]<https://github.com/BuzzFeedNews/2017-12-fake-news-top-50>
[8]<https://www.kaggle.com/mrisdal/fake-news>

- Modalização: marcadores que demonstram que o interlocutor possui uma posição definida sobre algo ou alguém. Exemplos: "acreditar", "recomendar";

- Sentimento: marcadores que indicam sentimento em um discurso. Exemplos: "amor", "aterrorizar", "felizmente";

- Valoração: marcadores que denotam intensidade no discurso. Exemplos: "completamente", "melhor", "mais".

4.4.2 Léxicos em Inglês

Para a execução de experimentos na língua Inglesa, é necessário a utilização de léxicos de subjetividade construídos especificamente para este idioma. Uma alternativa seria a tradução dos léxicos utilizados em Português, porém, tal tradução estaria sujeita a eventuais erros, o que poderia impactar nos resultados. Devido a essa possibilidade, foram utilizados léxicos de subjetividade já validados utilizados na literatura.

Para os experimentos nesse idioma, são utilizados três diferentes conjuntos de léxicos que expressam nuances de subjetividade. Estas nuances de subjetividade presentes nos léxicos consideram também aspectos de viés e sentimentos presentes em discursos na língua inglesa. O primeiro conjunto de léxicos foi compilado por Recasens, Danescu-Niculescu-Mizil e Jurafsky (2013). Este conjunto de léxicos contém seis diferentes aspectos que denotam algum tipo de subjetividade, sendo eles:

- Factive Verbs: pressupõem a verdade em uma oração. Ex: “realize”, “forget”, “exciting” (27 termos);

- Implicative Verbs: insere a ideia de implicação em uma oração. Ex: “succeed”, “fail”, “neglect” (32 termos);

- Assertive verbs: verbos que afirmam uma proposição. Ex: “believe”, “figure”, “affirm” (66 termos);

- Hedges usado para reduzir o compromisso com a verdade de uma proposição, evitando declarações assertivas. Ex: “apparently”, “could”, “estimate” (100 termos);

- Reporting Verbs: usado para reportar ações de pessoas ou de atividades. Ex: “accuse”, “assure”, “claim” (181 termos);

- Bias-inducing lemmas: denota uma posição previamente estabelecida, ou enviesada. Ex: “advocate”, “amazing”, “barbarian” (654 termos).

O segundo conjunto de léxicos é apresentado por Wilson, Wiebe e Hoffmann (2005). Este conjunto de léxico é parte do projeto Multi-Perpective Question Answering (MPQA) Subjectivity Lexicons[9] e é dividido em polaridades de sentimentos (positiva e negativa), classificados como sendo de forte ou fraca subjetividade. Para este léxico, foi considerado apenas os termos classificados como sendo de forte subjetividade. Após a filtragem considerando apenas os termos de forte subjetividade em ambas as polaridades, foi obtido um total de 3.078 léxicos para a polaridade negativa, e termos 1.482 para a polaridade positiva.

O terceiro conjunto de léxicos utilizados foi o apresentado por Choi, Deng e Wiebe (2014) e também representa polaridades de sentimentos (positiva e negativa). Os termos presentes nos léxicos foram extraídos de documentos com alto teor de subjetividade, como editorial de jornais e blogs. Este conjunto de léxicos contém 1.003 termos para polaridade negativa e 493 para a positiva, sendo os léxicos extraídos da porção que os autores denominaram de “gold standard”, que consiste em léxicos selecionados manualmente.

4.5 Setup dos Experimentos

Para a classificação das notícias falsas e reais, é utilizado dois modelos de Aprendizado de Máquina, XGBoost e Random Forest, que são conhecidos pelo seu grande poder de predição, atingindo um desempenho de estado da arte em diversas aplicações (OLSON et al., 2017).

Silverman (2016), em uma avaliação empírica, encontrou uma distribuição de aproximadamente 4 notícias reais para cada notícia falsa (4:1) em perfis de notícias reportados como de “esquerda” e “direita” no período das eleições americanas de 2016. Com o objetivo de empregar algum nível de desbalanceio nos dados, essa distribuição é aplicada como uma proporção padrão para os experimentos nesta pesquisa. Esta proporção vai de encontro com a maioria dos estudos que tratam de notícias falsas, que consideram o problema em questão

[9]<https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/>

como um problema balanceado, onde os autores empregam a mesma distribuição para notícias falsas e reais, fugindo assim, do senso comum de que exista mais notícias reais do que falsas em circulação.

Como a base de dados de notícias reais é maior que a de notícias falsas, em especial para o conjunto de notícias em Português, as execuções de treino/teste são randomizadas considerando 100 repetições (random sampling), respeitando a distribuição de 4:1 notícias reais e falsas. Desta forma, é possível gerar resultados médios que permitem avaliar o desempenho dos modelos utilizados.

Para calcular as distâncias semânticas utilizando o WMD, foi gerado um modelo implementando word embeddings a partir de um dump da Wikipedia para o Português. Para os experimentos em Inglês, foi utilizado o modelo pré-trainado denominado Google News embeddings[10].

Como métrica de avaliação, é utilizado as métricas Precision, Recall, F-measure e PR-AUC. A métrica PR-AUC representa a área da curva AUC considerando o Precision e Recall. Esta métrica tem como vantagem considerar o desempenho dos modelos ao variar seus limiares de predição (threshould), dando uma visão mais abrangente dos modelos. As clássicas métricas Precision, Recall, F-measure avaliam os modelos de forma pontual, utilizando seu limiar de classificação padrão, que em geral é de 0.5, onde acima deste limiar de classificação, o modelo tende a classificar uma amostra como sendo positiva, por exemplo. Nesta pesquisa, as quatro métricas de avaliação serão reportadas, porém, a métrica PR-AUC será empregada como métrica principal de avaliação.

4.6 Resultados

Esta seção irá descrever os resultados obtidos no âmbito de classificação e explicação dos modelos implementados para a classificação de notícias falsas e reais. Aqui, os resultados serão divididos em subseções para as notícias em Português e Inglês.

[10]<https://code.google.com/archive/p/word2vec/>

4.6.1 Resultados para as notícias em Português

Esta seção tem como objetivo descrever os principais resultados encontrados para os experimentos realizados. Os experimentos executados abrangem aspectos de análise estatística das distâncias semânticas por sentenças, classificação de notícias falsas e reais utilizando tais distâncias, e ensaios que buscam demonstrar a explicabilidade dos modelos por meio dos valores SHAP.

Análise das distâncias semânticas

Como descrito nas seções anteriores, as distâncias semânticas ente os léxicos e as sentenças dos documentos reportadas pelo algoritmo do WMD são utilizadas para a análise e classificação das notícias falsas e reais. As Tabela 4.2 e Tabela 4.3 exibem estatísticas descritivas destas distâncias, para as notícias falsas e reais, respectivamente. Em cada coluna, é apresentada as estatísticas das distâncias semânticas entre as sentenças dos documentos e cada um dos léxicos utilizados. Para os ensaios em Português, foram utilizados os cinco léxicos de subjetividade descritos em seções anteriores deste documento.

	Argumentação	**Pressuposição**	**Sentimento**	**Valoração**	**Modalização**
nº sentenças	1.262	1.262	1.262	1.262	1.262
média	0.872157	0.865573	0.866132	0.869331	0.874076
std	0.014562	0.011595	0.009140	0.011721	0.012182
min	0.811058	0.796779	0.816884	0.822704	0.808073
25%	0.863432	0.860071	0.860033	0.862856	0.867496
50%	0.873016	0.866392	0.865291	0.869716	0.875169
75%	0.881896	0.872832	0.872063	0.877078	0.881241
max	0.907376	0.894934	0.895981	0.900131	0.902114

Tabela 4.2: Estatísticas descritivas das distâncias semânticas para as notícias falsas em Português

A partir da Tabela 4.2 e Tabela 4.3, é possível observar que, para o caso das notícias reais e falsas em Português, os valores para as estatísticas apresentadas para as duas base de dados são bastante similares, havendo poucas diferenças para os dois conjuntos. Esta similaridade pode ser melhor visualizada nas Figuras 4.1 e Figura 4.2, que apresentam boxplots para os dois conjuntos.

A Tabela 4.4 exibe a análise estatística comparando os valores das distâncias semânticas entre as sentenças das notícias falsas e reais. Na análise, é executado o teste de hipótese de

	Argumentação	**Pressuposição**	**Sentimento**	**Valoração**	**Modalização**
nº sentenças	1.048.576	1.048.576	1.048.576	1.048.576	1.048.576
média	0.873015	0.865627	0.86633	0.869616	0.874199
std	0.0133645	0.0108445	0.00889849	0.0109894	0.0112677
min	0.775375	0.764377	0.809926	0.779943	0.776655
25%	0.864464	0.860394	0.860672	0.863438	0.867933
50%	0.87357	0.866169	0.86578	0.870071	0.874846
75%	0.882205	0.872328	0.871784	0.876617	0.881418
max	0.927491	0.916601	0.910878	0.919109	0.927155

Tabela 4.3: Estatísticas descritivas das distâncias semânticas para as notícias reais em Português

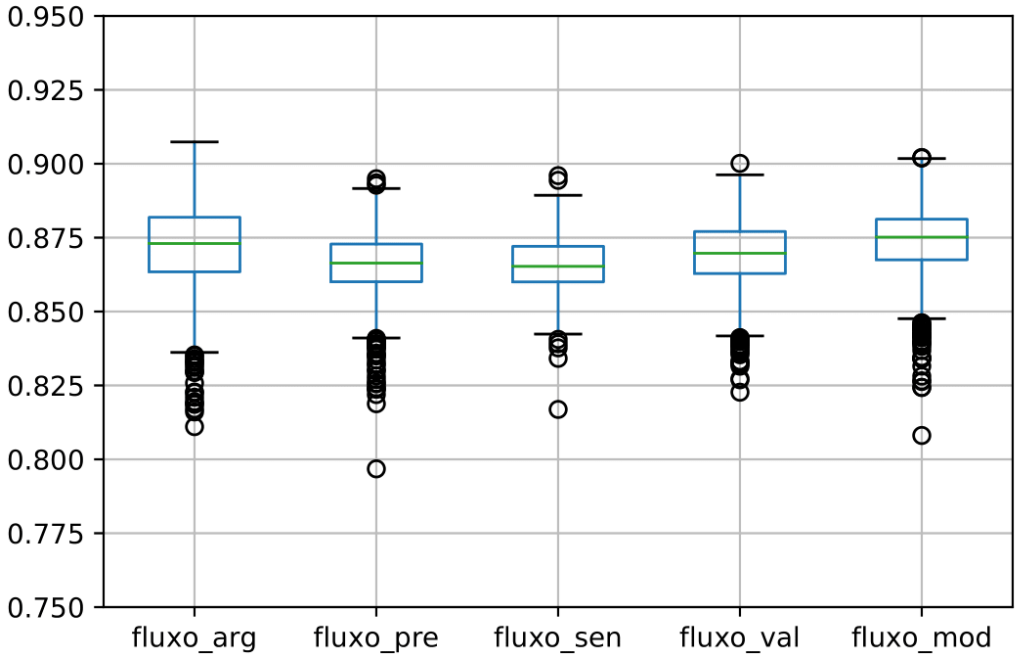

Figura 4.1: Boxplot para as distâncias semânticas das notícias falsas em Português.

Mann-Whitney, o qual consiste em um teste de hipótese não-paramétrico envolvendo duas amostras independentes, sendo robusto em cenários para amostras de tamanhos diferentes (MANN; WHITNEY, 1947). Ainda assim, devido à grande diferença de tamanho entre as base de dados de notícias falsas (1.262 sentenças) e reais (1.048.576 sentenças), para tal análise, foi empregada uma abordagem estratificada, onde foram executados 50 ensaios repetidos, onde em cada repetição, uma amostra de 1.262 distâncias de sentenças de notícias falsas e reais são randomicamente selecionadas para a execução do teste de hipótese. O

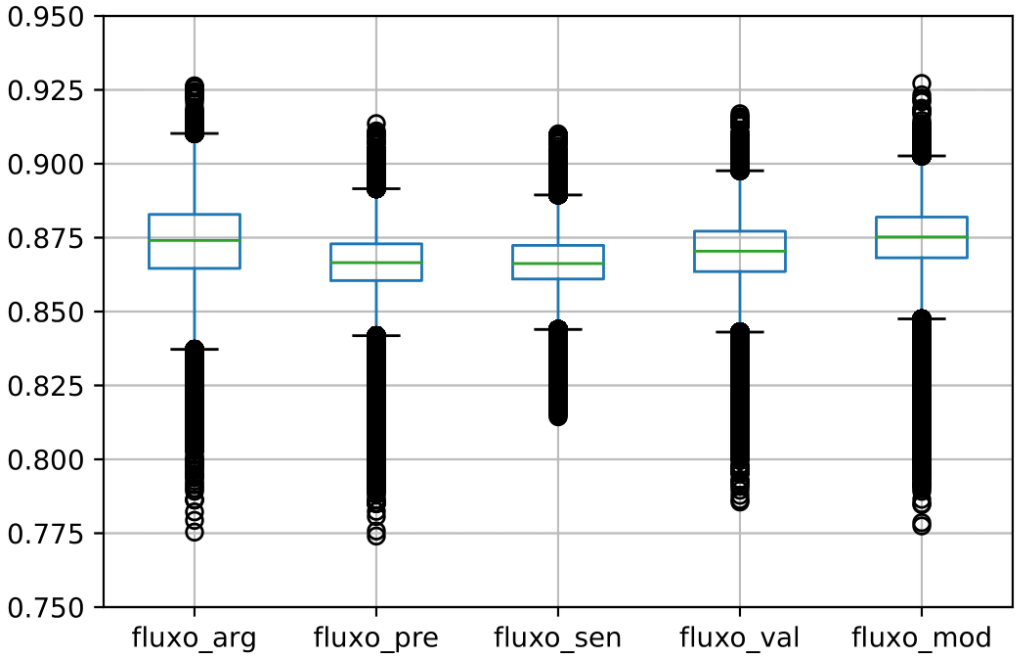

Figura 4.2: Boxplot para as distâncias semânticas das notícias reais em Português.

valor de 1.262 foi estabelecido por ser a quantidade de sentenças da base de dados de notícias falsas, permitindo assim, ensaios com a mesma quantidade de amostras. Desta forma, a primeira coluna da Tabela 4.4 apresenta a quantidade de ensaios, para cada léxico, onde a hipótese nula do teste é rejeitada (p-valor < 0.05). A hipótese nula assume a mesma distribuição para as distâncias das notícias falsas e reais. Nesta primeira coluna, é possível visualizar que o léxico de sentimentos apresentou a maior quantidade de rejeições da H0, obtendo 38 rejeições dentre os 50 ensaios realizados. Na segunda coluna, é apresentado um ensaio padrão, sem estratificação, onde as distâncias das sentenças das notícias falsas e reais são testadas de forma convencional, considerando que a hipótese alternativa representa que as duas amostras vêm de distribuições diferentes (*two-sided*). Nesta coluna é apresentado apenas os valores onde o p-valor é menor que 0.05, ou seja, onde este é significativo. Neste caso, pode-se perceber que o resultado demonstra que há diferenças significativas entre as sentenças das notícias falsas e reais para os léxicos de Argumentação (p-valor = 0,009) e Sentimento (p-valor = 0,001). Agora, sabendo que há diferenças significativas para dois léxicos, é imperativo saber, de forma significativa, qual o caráter destas diferenças. Por exemplo,

	# H0 rejeitada	H1 two-sided	H1 (Falsa >Real)	H1 (Falsa <Real)
Argumentação	22	0,009	-	0,004
Pressuposição	1	-	-	-
Sentimento	38	0,001	-	0,0005
Valoração	11	-	-	-
Modalização	3	-	-	-

Tabela 4.4: Testes de hipótese comparando as distâncias semânticas das sentenças para cada léxico, considerando as notícias falsas e reais em Português. Os resultados apresentam o número de ensaios em que foi reportado uma diferença significativa (p-value < 0,05) entre as distâncias semânticas das notícias falsa e reais em Português (coluna "# H0 rejeitada"). Na segunda coluna ("H1 two-sided") é apresentado os resultados (p-values significativos) de ensaio tradicional (execução simples do teste), comparando as distâncias das sentenças falsas e reais, onde a hipótese alternativa consiste em considerar que ambas as amostras pertencem a distribuições diferentes. Na terceira coluna ("H1 (Falsa >Real)") o mesmo ensaio é executado, porém agora a hipótese alternativa (H1) consiste em afirmar que os valores das distâncias semânticas das sentenças das notícias falsas são maiores (menos subjetivas) que os valores das sentenças das notícias reais. A quarta coluna ("H1 (Falsa <Real)") realiza a mesma testagem, porém, considerando que a hipótese alternativa considera que as distâncias semânticas das sentenças das notícias falsas são menores (mais subjetivas) que as distâncias das sentenças das notícias reais.

é importante saber se as sentenças das notícias falsas são mais subjetivas (menores) que as distâncias das sentenças de notícias reais.

Para tal análise, a terceira coluna da Tabela 4.4 apresenta a mesma testagem apresentada na segunda coluna, porém, a hipótese alternativa é modificada para considerar que as distâncias das sentenças das notícias falsas sejam maiores (menos subjetivas) que as distâncias das reais. Neste caso, vê-se que esta hipótese alternativa não pôde ser verificada para nenhum dos cinco léxicos. A quarta coluna apresenta o inverso, sendo a hipótese alternativa aceita quando as distâncias das notícias falsas forem menores que as das reais. Para este cenário, corroborando com os ensaios apresentados nas colunas anteriores da tabela, vê-se que novamente para os léxicos de Argumentação e Sentimentos, a hipótese nula pôde ser rejeitada, significando que as sentenças das notícias falsas são estatisticamente mais subjetivas que as das notícias reais. Em tese, estes resultados dão um indício de que as notícias falsas, em termos semânticos, parecem ser simultaneamente mais argumentativas e emotivas, quando comparadas com as notícias reais.

Resultados de Classificação

A Tabela 4.5 exibe os resultados médios em termos de PR-AUC, Precision, Recall e F1-Score para a classificação de notícias falsas utilizando as médias de subjetividade por documentos, considerando 100 repetições para cada modelo. Neste cenário, cada documento é representado como um vetor de cinco dimensões, onde cada dimensão representa a distância média das sentenças do documento em relação a cada um dos cinco léxicos de subjetividade utilizados para o Português. Para uniformização dos experimentos, documentos contendo até o máximo de 100 sentenças são utilizados.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0,27(±0,04)	0,13(±0.07)	0,30(±0.15)	0,08(±0,05)
RF	0,26(±0,04)	0,15(±0,06)	0,32(±0,13)	0,10(±0,05)

Tabela 4.5: Resultados médios da classificação de notícias falsas e reais utilizando as médias de subjetividade por documentos. Nesta representação, cada documento é representado por um vetor contendo cinco features, que consistem na média das distâncias reportadas pelo WMD em relação a cada um dos cinco léxicos de subjetividade utilizados para o Português.

A Tabela 4.6, mostra os resultados para a mesma classificação, porém, utilizando os Vetores de Subjetividade por Sentenças. Neste experimento, os documentos são representados por vetores de 100 dimensões, onde cada dimensão representa a distância de uma sentença relativa a um dos cinco léxicos utilizados. As demais dimensões do vetor são preenchidas (padding) com a média do vetor. Este vetor é então, concatenado com os vetores das outras quatro dimensões de subjetividade, obtendo assim, a representação de um único documento contendo 500 features, dado que para os experimentos em Português, são utilizados cinco léxicos de subjetividade.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0,30(±0.04)	0,13(±0.06)	0,36(±0.17)	0,08(±0,04)
RF	0,26(±0,04)	0,14(±0,07)	0,32(±0,13)	0,09(±0,05)

Tabela 4.6: Resultados médios da classificação de notícias falsas e reais utilizando os vetores subjetividade por sentenças. Nesta representação, cada documento é representado pelas distâncias semânticas de cada uma de suas sentenças relativa a um dos léxicos de subjetividade, considerando um tamanho máximo de 100 sentenças por documento.

Nos resultados apresentados pela Tabela 4.6, pode-se observar que o uso dos vetores de subjetividade por sentença gera uma ligeira melhora no desempenho da classificação de

notícias falsas e reais em termos de PR-AUC, quando comparado com os resultados dos modelos que utilizam as médias por sentenças, reportados pela Tabela 4.5 considerando o modelo XGBoost. Este resultado se mostrou estatisticamente significativo, reportando um p-value = 0.000003, ao utilizar um Teste T de hipótese. Ao considerarmos o modelo Random Forest, a diferença entre os dois ensaios não se mostrou significativa (p-value = 0.46).

A Tabela 4.7 exibe os resultados para os modelos baseados em TFIDF. Nos resultados apresentados, pode-se notar um desempenho médio significativamente maior em relação aos modelos baseados em features de subjetividade. Devido à capacidade de representar termos específicos nos documentos, estes modelos conseguem gerar uma representação completamente baseada na ocorrência de palavras nos documentos, o que em certos casos, pode ser útil, como em problemas de recuperação da informação. Porém, quando utilizado em classificações textuais, estes modelos podem gerar representações enviesadas, por terem como base, a ocorrência de termos específicos. Para demonstrar esta hipótese, foram executados experimentos que aproveitam a variabilidade na base de dados de notícias reais, que é subdividido em quatro diferentes tópicos (i.e. Esporte, Política, Economia e Cultura). Para este cenário, os experimentos foram executados em um design denominado "domínio-cruzado", onde os modelos são treinados utilizando notícias reais de um determinado tópico, por exemplo Cultura. Para o conjunto de testes, é considerado notícias reais de um tópico distinto do utilizado para o treino, como por exemplo, Política. Deste modo, é possível avaliar a classificação de notícias reais e falsas considerando variações no domínio das notícias reais.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0,68(±0,07)	0,47(±0,09)	0,81(±0,09)	0,33(±0,08)
RF	0,51(±0,06)	0,27(±0,09)	0,76(±0,14)	0,17(±0,07)

Tabela 4.7: Resultados médios da classificação entre notícias falsas e reais utilizando a representação clássica baseada em TFIDF para as notícias em Português.

A Tabela 4.8 apresenta os resultados médios dos ensaios utilizando o design de domínio-cruzado para os modelos baseados nos vetores de subjetividade por sentenças. É possível observar nos resultados, que não há grandes variações observáveis em termos de PR-AUC, inclusive havendo melhoras significativas nos resultados em termos de F1-score e Recall para ambos os modelos, quando comparados com os resultados que não utilizam o design de domínio-cruzado (Tabela 4.6).

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.30(±0.05)	0.18(±0.08)	0.32(±0.12)	0.14(±0.07)
RF	0.25(±0.03)	0.17(±0.07)	0.29(±0.11)	0.13(±0.06)

Tabela 4.8: Resultado médio da classificação de notícias falsas e reais utilizando os vetores subjetividade por sentenças, considerando o design domínio-cruzado. Neste design, os tópicos das notícias reais são variados entre os conjuntos de traino e teste.

A Tabela 4.9 exibe o mesmo experimento, porém considerando modelos baseados em TFIDF. Nestes resultados, apesar de ainda serem superiores aos modelos baseados em subjetividade, pode-se notar que, o uso de features baseadas em BoW tem como consequência a menor generalização dos modelos quando submetidos a variações nos tópicos dos documentos. Nos resultados, quando comparados com o ensaio utilizando os mesmos modelos, porém, sem a utilização do design de domínio-cruzado (Tabela 4.7). Nos resultados apresentados, pode-se notar uma significativa redução no desempenho dos modelos. Em termos de PR-AUC, para o modelo XGBoost, houve uma redução de aproximadamente 32% na execução utilizando o domínio-cruzado. Para o Random Forest, a redução na PR-AUC foi de 25% no resultado médio. Estes resultados sugerem que os modelos clássicos de classificação de texto utilizados no cenário de notícias falsas podem sofrer de um forte viés presente na própria base de dados ao qual são avaliados na literatura. Na próxima seção, os modelos serão analisados sob a perspectiva das explicações de suas classificações, permitindo assim, um melhor entendimento sobre esta redução no desempenho da classificação dos modelos baseados em BoW e TFIDF.

Por fim, todos os modelos baseados em subjetividade apresentados para a classificação foram significativamente melhores, em termos de PR-AUC, do que modelos que baseiam suas classificações em escolhas aleatórias baseadas apenas na distribuição das classes. Como a distribuição das classes de notícias reais (classe 0) e notícias falsas (classe 1) foi definida como 4:1, tais modelos foram definidos para realizar escolhas aleatórias sabendo que a classe de notícias reais é mais provável de ocorrer nos dados. Estes modelos pseudo-aleatórios apresentaram, de forma consistente em todos os cenários, uma média de PR-AUC de 0,20±0,06, sendo significativamente menores do que todos os demais modelos baseados em subjetividade.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.46(±0.15)	0.42(±0.10)	0.40(±0.17)	0.57(±0.19)
RF	0.38(±0.13)	0.33(±0.12)	0.47(±0.22)	0.31(±0.13)

Tabela 4.9: Resultado médio da classificação de notícias falsas e reais utilizando os modelos baseados em features TFIDF, considerando o design de domínio-cruzado. Neste design, os tópicos das notícias reais são variados entre os conjuntos de treino e teste.

Explicação dos modelos

Para avaliar os modelos do ponto de vista de suas explicações, são utilizados nesta pesquisa os valores reportados pelo SHAP, que demonstram a relevância de cada feature para a predição ou classificação de um modelo. Este tipo de avaliação permite a observação de como nuances presentes nos dados influenciam a tomada de decisão de um modelo de classificação, por exemplo.

Para tal avaliação, foram considerados os melhores modelos, do ponto de vista de PR-AUC, que utilizam tanto as features de subjetividade propostas, como as baseadas em BoW e TFIDF. Para ambos os casos, foi selecionado o modelo de classificação XGBoost para esta análise. Este modelo foi escolhido por ter apresentado os melhores resultados de uma forma geral, ao compararmos com o Random Forest. Para o modelo treinado com as features de subjetividade propostas, foi escolhido o modelo que apresentou o melhor resultado nos ensaios utilizando as médias de subjetividade por documentos. Este modelo foi escolhido em detrimento dos modelos que utilizam os vetores de subjetividade por sentenças para manter a visualização mais simples e intuitiva, dado que este possui apenas cinco features. O melhor modelo utilizando estas features de subjetividade, apresentou um PR-AUC de 0,34. Para o modelo baseado em TFIDF, o que apresentou melhor desempenho obteve uma PR-AUC de 0,84. Em ambos os casos, foram escolhidos os melhores modelos treinados para a classificação convencional de notícias falsas e reais, ou seja, sem a adoção do design de domínio-cruzado.

A Figura 4.3 exibe a plotagem (*summary plot*) reportada pelo SHAP considerando a relevância das suas cinco features para a classificação do modelo. Na imagem, o eixo y apresenta as cinco features de subjetividade utilizadas em ordem de importância. Estas features representam a distância semântica média de cada documento para um léxico específico. No eixo x, tem-se a faixa de valores que consistem nos shap values, que expressam o peso que cada

feature exerce para determinar a classificação de uma amostra. O eixo positivo representa um maior peso para a classificação da classe alvo (classe 1), que neste caso, consiste nas notícias falsas. O eixo negativo representa uma maior tendência para a classificação da classe de notícias reais (classe 0). Logo, pontos deslocados mais a direita significam uma maior chance para a classificação de uma amostra como falsa, considerando uma determinada feature. Os pontos o gráfico representam uma amostra, no caso, uma notícia. E a cor do ponto representa o valor real de uma dada feature, onde quanto mais próximo da escala em vermelho, maior é o valor numérico de uma dada feature em relação a uma amostra específica, ou seja, um ponto.

Pode-se notar na Figura 4.3, que o léxico de sentimentos (sen_avg) foi o que apresentou uma maior relevância para as escolhas de classificação deste modelo em particular, dado que esta feature está no topo do eixo y. Adicionalmente, podemos observar que, ainda nesta feature, há uma prevalência de pontos azuis correlacionados positivamente com valores maiores que zero no eixo x. Isso significa que, distâncias semânticas menores relativas ao léxico de sentimentos estão correlacionadas com predições de notícias falsas. Isto ocorre pois o WMD reporta um valor de similaridade e, portanto, quanto menor a distância semântica entre dois documentos, mais similares ambos serão, dentro do espaço vetorial adotado. Para este caso, o inverso também pode ser observado, onde as predições de notícias reais estão claramente relacionadas com distâncias semânticas maiores para o léxico de sentimentos, demonstrado pelos pontos vermelhos correlacionados com os valores presentes na porção negativa do eixo x. Comportamento semelhante pode também ser observado para a feature de argumentação (arg_avg), mostrando uma tendência de que as notícias falsas parecem adotar uma linguagem simultaneamente mais emotiva e argumentativa, quando comparado com as notícias reais.

A Figura 4.4 apresenta a mesma análise, porém considerando o modelo treinado com as features baseadas em BoW e TFIDF. No eixo y, são apresentadas as vinte features mais relevantes para o modelo. Lembrando que, como este modelo se baseia em BoW, as features são pesos TFIDF associados a termos presentes no vocabulário dos documentos. Os termos exibidos estão reduzidos à sua raiz obtida através do *stemming* destes. Apesar do modelo reportar um desempenho significativamente maior que o modelo utilizando as métricas de subjetividade, pode-se notar que os termos considerados como mais relevantes não parecem possuir grandes relações com o problema de classificação de notícias falsas, sendo

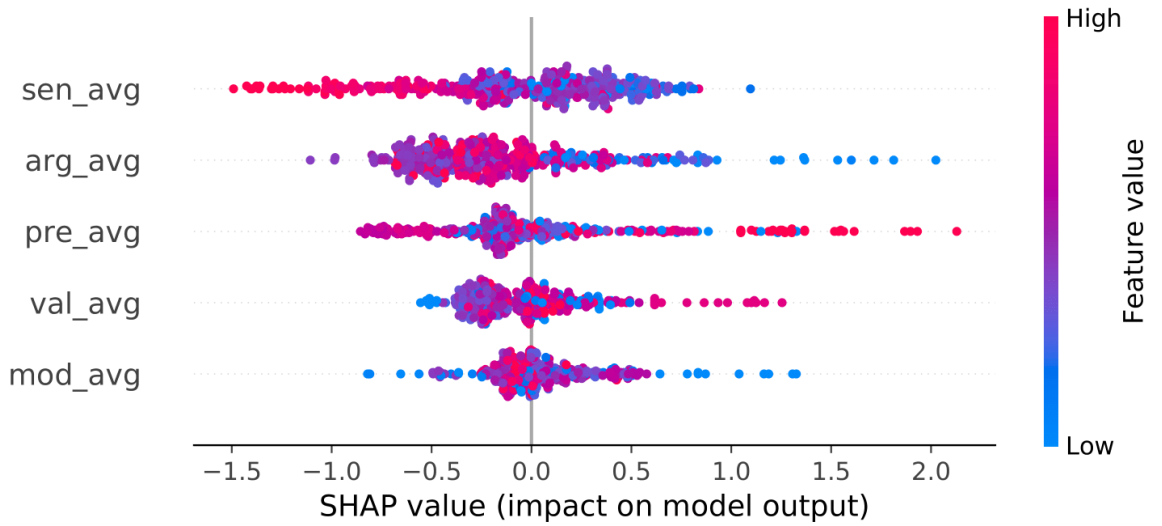

Figura 4.3: *Summary plot* gerado através do SHAP, exibindo o peso que as features exercem sobre a decisão de classificação do modelo. No eixo y, estão listadas as cinco features de subjetividade que formam a representação vetorial de um documento, considerando as médias de subjetividade por documento como features. No eixo x, estão os *shap values*, onde valores maiores que zero representam uma maior chance para a classificação da classe alvo (classe 1), que neste caso, são as notícias falsas. Valores negativos (menores que zero), representam uma maior chance para a classificação de notícias reais (classe 0).

estes termos muito genéricos e aparentemente pouco representativos dentro do contexto de notícias falsas. Possivelmente, este modelo conseguiu capturar nuances relativas ao estilo de escrita ou mesmo ao tópico dos documentos, ao invés de capturar características que, de fato, possam distinguir notícias falsas e reais. Por exemplo, temos o termo "equip" que representa a palavra "equipe", onde este apresenta valores maiores correlacionados com a porção negativa do eixo x, ou seja, este termo tem influência para a classificação de notícias reais. Na verdade, este termo é bastante frequente nas notícias reais de esportes, levando o modelo a aprender que este termo é importante para a classificação deste tipo de notícia. É importante destacar que, para este cenário, onde as features baseadas em TFIDF consistem em vetores esparsos, valores em vermelho representam termos que possuem um peso TFIDF maior que zero, significando que o termo ocorre em um dado documento. Comportamento similar ao apresentado pelo termo "equip" também pode ser observado para o termo "jog" que representa a palavra "jogo", estando este termo também vinculado às notícias reais de esportes.

Adicionalmente, dois termos se destacam dentre as features mais relevantes para este

modelo, sendo eles “lul” que representa o nome “lula” e “bolsonar” que representa o nome “bolsonaro”. Para o primeiro termo, vemos que sua ocorrência parece tendenciar a classificação do modelo para uma notícia falsa. Isto acontece pois, dada a faixa temporal em que as notícias falsas foram colhidas (entre 2010 e 2017), havia uma prevalência da palavra “Lula” entre as notícias falsas. Comportamento similar, porém com menor intensidade, pode ser observado para o termo que representa o nome “Bolsonaro”. Estes exemplos demonstram que, apesar do modelo baseado em BoW e TFIDF apresentar um ótimo desempenho, o mesmo parece capturar nuances que não estão diretamente ligadas à problemática da classificação de notícias falsas, mas sim com características mais vinculadas ao tópico ou estilo de escrita dos documentos. Esta hipótese se demonstra claramente na significativa queda de performance apresentada por este tipo de modelo nos experimentos que utilizam o domínio-cruzado, que podem ser comparados nas Tabelas 4.7 e Tabela 4.9.

4.6.2 Resultados para as notícias em Inglês

Seguindo o mesmo método de experimentos executados para os ensaios em Português, esta seção descreve os resultados obtidos para as notícias em língua inglesa.

Análise das distâncias semânticas

A Tabela 4.10 apresenta um sumário das estatísticas descritivas relativas às distâncias semânticas das sentenças para a base de dados de notícias falsas, totalizando 161.400 sentenças. De forma análoga, a Tabela 4.11 apresenta o sumário para as notícias reais em Inglês, totalizando 577.700 sentenças de notícias reais. Nestas tabela, ao compararmos as médias de cada um dos léxicos, podemos observar que as sentenças de notícias falsas tendem a apresentar valores de subjetividade sutilmente menores que as sentenças das notícias reais. Isso demonstra uma tendência de que as notícias falsas apresentam elementos semânticos mais fortes do que as notícias reais, denotando uma maior similaridade das notícias falsas com os léxicos de subjetividade. A única exceção é em relação ao léxico “hedges”, que apresentou maior semelhança semântica com as notícias reais. As Figuras 4.5 e Figura 4.6 exibem a distribuição dos dados por meio de boxplots. Ao analisar as medianas definidas nos boxplots, é possível observar as sutis diferenças em cada léxico, para ambos as base de dados, onde é

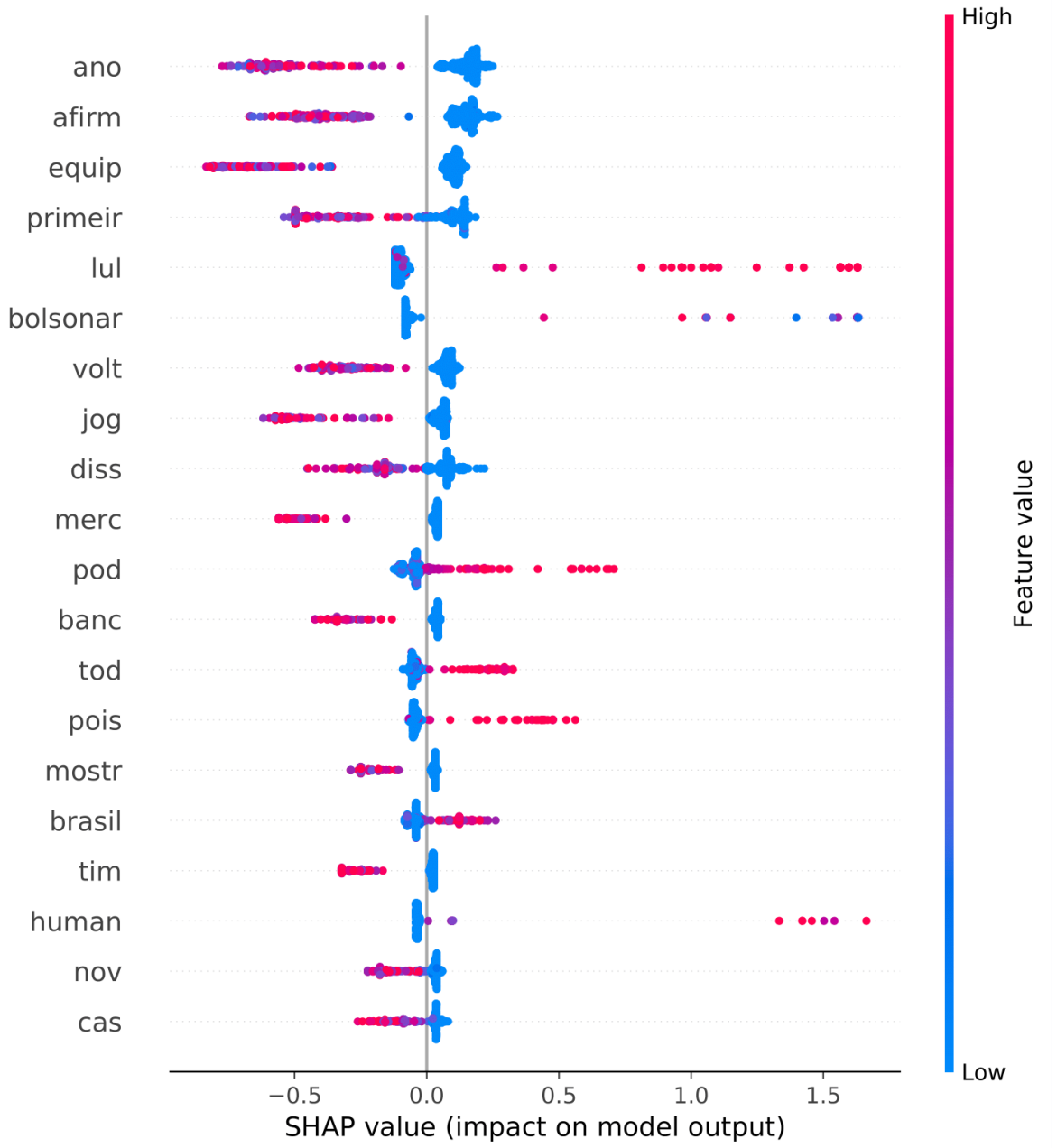

Figura 4.4: *Summary plot* gerado através do SHAP exibindo as features mais relevantes para a classificação do modelo baseado em BoW e TFIDF.

possível perceber que as medianas das notícias falsas tendem a ser menores que as medianas dos léxicos para as notícias reais, demonstrando que as sentenças das notícias falsas são semanticamente mais similares aos léxicos que expressam subjetividade.

A Tabela 4.12 apresenta os testes de hipóteses para as notícias em Inglês, seguindo o mesmo padrão dos testes de hipóteses executados para os ensaios em Português presentes na Tabela 4.4 da Seção 4.6.1. Nesta tabela, é possível verificar que, para todos os 50 ensaios

	assertives	factives	hedges	implicatives	negative	positive	report	bias	positive_gold	negative_gold
nº sentenças	161.400	161.400	161.400	161.400	161.400	161.400	161.400	161.400	161.400	161.400
média	0.848723	0.856772	0.840404	0.852624	0.795093	0.811316	0.814766	0.800111	0.801980	0.782325
std	0.017431	0.018439	0.017629	0.017137	0.021775	0.015803	0.013720	0.018127	0.043943	0.046436
min	0.743138	0.714145	0.730930	0.739973	0.704506	0.729462	0.732502	0.690459	0.630992	0.581031
25%	0.839179	0.847457	0.830128	0.843336	0.780995	0.802083	0.806709	0.789052	0.780129	0.761946
50%	0.848435	0.856724	0.840212	0.852669	0.796023	0.811782	0.813971	0.800781	0.808245	0.789917
75%	0.858291	0.867225	0.850557	0.862148	0.809183	0.820992	0.822361	0.811519	0.832919	0.815393
max	0.958710	0.966396	0.941242	0.950850	0.876647	0.879597	0.909514	0.887666	0.912833	0.881969

Tabela 4.10: Estatísticas descritivas para as distâncias semânticas obtidas para as sentenças presentes na base de dados de notícias falsas em Inglês.

	assertives	factives	hedges	implicatives	negative	positive	report	bias	positive_gold	negative_gold
nº sentenças	577.700	577.700	577.700	577.700	577.700	577.700	577.700	577.700	577.700	577.700
média	0.848964	0.857577	0.838923	0.853414	0.800162	0.813679	0.815159	0.80309	0.810614	0.792114
std	0.0164179	0.0175222	0.0170778	0.0160289	0.0209407	0.0160071	0.0134946	0.0164412	0.0389415	0.0402623
min	0.725402	0.697192	0.715651	0.726747	0.711466	0.731334	0.731048	0.709798	0.62922	0.578198
25%	0.840313	0.849224	0.830369	0.845632	0.788118	0.805208	0.807188	0.794083	0.791524	0.773882
50%	0.848829	0.857694	0.838786	0.853366	0.800783	0.813881	0.814977	0.802924	0.815124	0.796762
75%	0.857466	0.866482	0.847478	0.861681	0.813147	0.822689	0.822715	0.812097	0.836303	0.819765
max	0.958709	0.966396	0.941241	0.959407	0.89213	0.900637	0.909513	0.894564	0.920157	0.895628

Tabela 4.11: Estatísticas descritivas para as distâncias semânticas obtidas para as sentenças presentes na base de dados de notícias reais em Inglês.

randômicos executados, foi possível verificar diferenças significativas para todos os léxicos utilizados. Também é possível notar ao analisar a quarta coluna, que em praticametne todos os léxicos (com exceção do léxico de subjetividade "hedges"), as distâncias semânticas das sentenças falsas foram estatisticamente menores que as das notícias reais, contribuindo assim, para a hipótese de que notícias falsas parecem ser mais subjetivas que as reais, mesmo para o cenário de língua inglesa. Nesta tabela, como os valores para os p-valores foram muito pequenos, seus valores foram arredondados, exibindo apenas o valor 0.00 quando aplicável. Contribuiu para a grande significância dos testes, o fato de haver uma grande quantidade de amostras de sentenças falsas (161.400) e reais (577.700), aumentando assim, o poder do

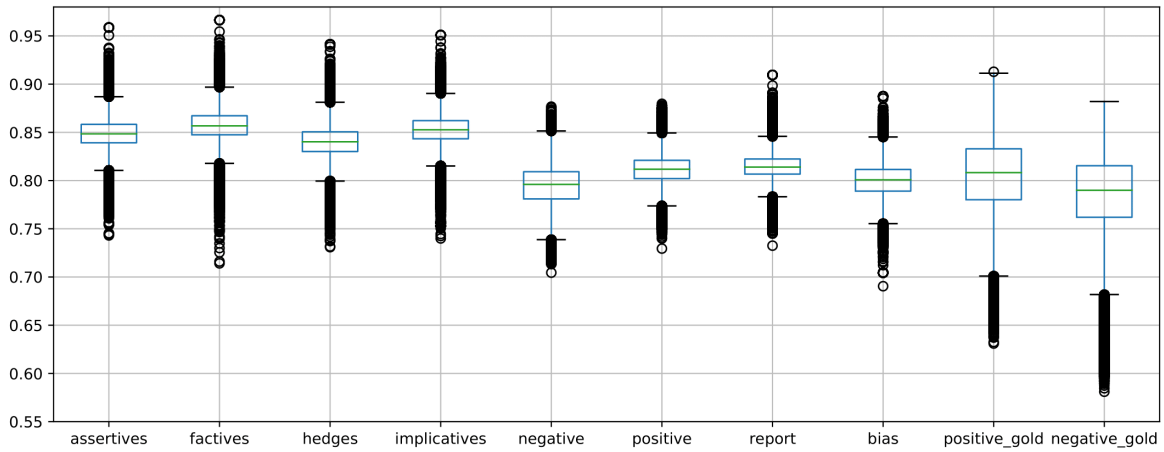

Figura 4.5: Boxplot para as distâncias semânticas das notícias falsas em Inglês.

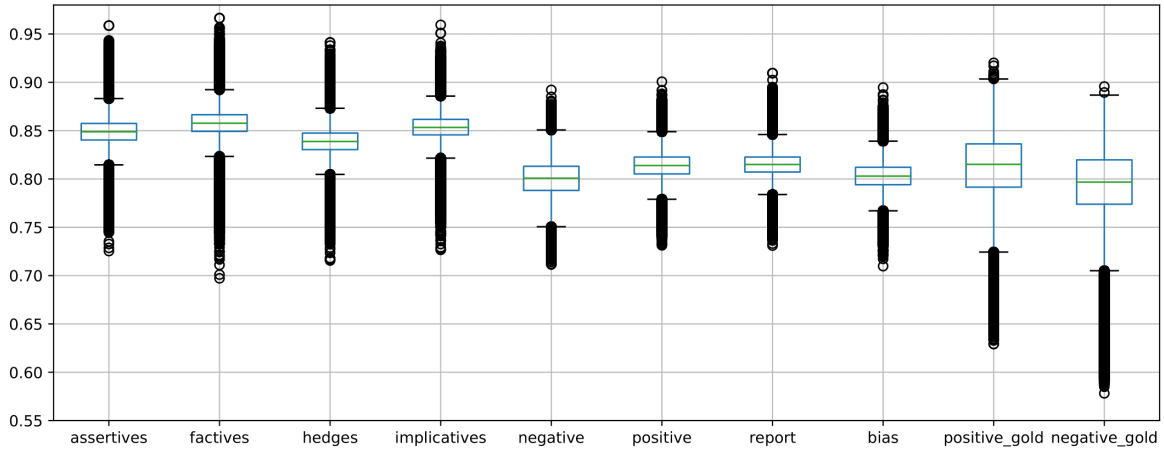

Figura 4.6: Boxplot para as distâncias semânticas das notícias reais em Inglês.

teste. Novamente, estas sutis diferenças, porém significativas, podem ser observadas nos boxplots apresentados, onde as medianas das notícias falsas (Figura 4.5) são menores que as medianas das notícias reais (Figura 4.6), demonstrando que as notícias falsas parecem ser mais subjetivas que as notícias reais.

Resultados de Classificação

A Tabela 4.13 apresenta os resultados médios da classificação de notícias falsas para a base de dados em Inglês, utilizando a média de subjetividade por documento. A princípio, é possível notar que os resultados apresentados são sistematicamente melhores quando comparados com os resultados reportados para as notícias em Português, apresentados, para este mesmo cenário, na Tabela 4.5. Este resultado indica que, possivelmente, tanto os léxicos utilizados para as notícias em Inglês quanto os embeddings utilizados, parecem se adequar melhor à classificação de notícias falsas. É importante notar que, os léxicos adotados nestes experimentos possuem o dobro de dimensões de subjetividade em relação aos léxicos em português, que possuem apenas cinco dimensões. Quanto aos embeddings, os vetores utilizados para os experimentos em inglês são oriundos de um conjunto de vetores derivados especificamente de notícias em inglês, enquanto que os embeddings utilizados em português são oriundos de artigos do Wikipédia. Esta diferença pode, em tese, tornar a classificação de notícias em inglês mais precisa, devida a melhor representatividade destas por parte dos vetores utilizados.

A Tabela 4.14 apresenta os resultados das classificações utilizando os vetores de sub-

	# H0 rejeitada	H1 two-sided	H1 (Falsa >Real)	H1 (Falsa <Real)
assertives	50	0.00	-	0.00
factives	50	0.00	-	0.00
hedges	50	0.00	0.00	-
implicatives	50	0.00	-	0.00
negative	50	0.00	-	0.00
positive	50	0.00	-	0.00
report	50	0.00	-	0.00
bias	50	0.00	-	0.00
positive_gold	50	0.00	-	0.00
negative_gold	50	0.00	-	0.00

Tabela 4.12: Testes de hipótese comparando as distâncias semânticas das sentenças para cada léxico, considerando as notícias falsas e reais em Inglês. Nos resultados, é possível observar que para todos os léxicos utilizados, houve diferenças significativas (p-value < 0,05). Para estes casos, a maioria os testes demonstrou que as notícias falsas apresentaram distâncias semânticas menores que as notícias reais, o que demonstra uma maior similaridade semântica das notícias falsas com os léxicos de subjetividade, o que denota uma maior subjetividade das notícias falsas.

jetividade por sentenças. Assim como observado para os resultados em Português, o uso dos vetores de subjetividade parece melhorar significativamente os resultados de classificação, em especial, para o modelo XGBoost, que apresenta, de forma sistemática, um melhor desempenho nos experimentos realizados. Isso demonstra que processor de *feature engeneering* mais complexos podem trazer bons resultados para este tipo de classificação. Neste cenário, os resultados também foram significativamente melhores quando comparados com os resultados apresentados para as notícias em Português (4.8).

A Tabela 4.15 exibe os resultados para o modelo de classificação baseado em BoW e TFIDF. Novamente, este cenário é o que apresenta os melhores resultados. Porém, é importante destacar que estes modelos também sofrem dos mesmos problemas que puderam ser observados na Seção 4.6.1, onde as features mais relevantes do modelo são acessadas por meio do SHAP, exibindo assim, explicações para as classificações para o modelo treinado com notícias em Português. Detalhes relativos à explicação deste modelo treinado com notícias em Inglês será discutido na próxima seção deste capítulo.

Como as notícias reais em Inglês não possuem a mesma divisão por tópicos, semelhante às notícias em Português, os experimentos utilizando o design de domínio-cruzado não foram executados para o cenário de notícias em Inglês. Assim como ocorrido para o cenário de experimentos em Português, todos os modelos gerados para os experimentos em Inglês

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.43(±0,06)	0,33(±0,08)	0,58(±0,11)	0,23(±0,07)
RF	0,40(±0,06)	0,32(±0,08)	0,57(±0,12)	0,23(±0,07)

Tabela 4.13: Resultado médio para a classificação de notícias falsas e reais em Inglês, considerando as médias de subjetividade por documento como features.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0,50(±0,07)	0,36(±0,10)	0,63(±0,12)	0,25(±0,09)
RF	0,38(±0,06)	0,32(±0,09)	0,55(±0,12)	0,23(±0,08)

Tabela 4.14: Resultado médio para a classificação de notícias falsas e reais em Inglês, considerando os vetores de subjetividade por sentenças como features.

foram significativamente melhores do que modelos de classificação randômica, que baseiam suas classificações em escolhas aleatórias. Os modelos randômicos também reportam, assim como para o cenário em Português, uma média de PR-AUC de 0,20±0,06, sendo significativamente menores que os modelos baseados em subjetividade.

Explicação dos modelos

De forma análoga à apresentada na Seção 4.6.1 para as notícias em Português, a Figura 4.7 exibe as features mais relevantes para o melhor modelo gerado nas classificações usando as médias de subjetividade por documento. Este modelo apresentou uma PR-AUC de 0,52. Pode-se notar que a feature mais relevante para este modelo em específico a relacionada ao léxico "hedges", o qual seus valores mais reduzidos (em azul) parecem exercer influência para a classificação de notícias reais (eixo x negativo). Esta observação parece concordar com a semântica destes termos pois, em suma, termos "hedges" tendem a denotar um menor grau de asserção a uma afirmação, como por exemplo, termos como "apparently" e "could". Estes termos podem indicar, possivelmente, a tentativa dos escritores das notícias reais em manter um grau de imparcialidade sobre determinados fatos, o que na maioria das vezes, não acontece em notícias falsas, onde os criadores destes conteúdos tendem a demonstrar e a con-

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0,81(±0,04)	0,70(±0,05)	0,82(±0,06)	0,61(±0,07)
RF	0,60(±0,07)	0,46(±0,10)	0,84(±0,10)	0,32(±0,09)

Tabela 4.15: Resultados médios da classificação entre notícias falsas e reais utilizando a representação clássica baseada em TFIDF para as notícias em Inglês.

vencer o leitor da veracidade de um fato inverídico. Também é possível notar que as notícias reais parecem ser mais positivas, considerando o léxico "positive_gold_avg" e as falsas com tendência à negatividade, sendo isto observado no léxico "negative_avg". Também é interessante observar que as notícias falsas apresentam um grau de enviesamento maior, quando comparadas com as notícias reais, sendo notado pelo léxico "bias_avg". Este indício de viés é fundamental para a compreensão das nuances apresentadas pelas notícias falsas, dado que, na maioria das vezes, estas tendem a ser direcionadas para suportar fatos inverídicos ou mesmo atacar personalidades, e para tal, fazem uso de termos que demonstram um enviesamento por parte do autor da notícia. Porém, na mesma figura, é possível notar algumas incongruências, como por exemplo, o léxico sétimo léxico em relevância "negative_gold_avg" parece discordar do terceiro léxico mais relevante "negative_avg". Provavelmente, estas discordâncias se dão devido a diferenças na forma em que os léxicos foram gerados (léxicos gerados manualmente ou automaticamente). Em todo caso, para este exemplo, como o léxico "negative_avg" apresenta uma maior relevância para a classificação, sendo o terceiro mais relevante, este parece ser mais confiável, ao menos para este cenário específico.

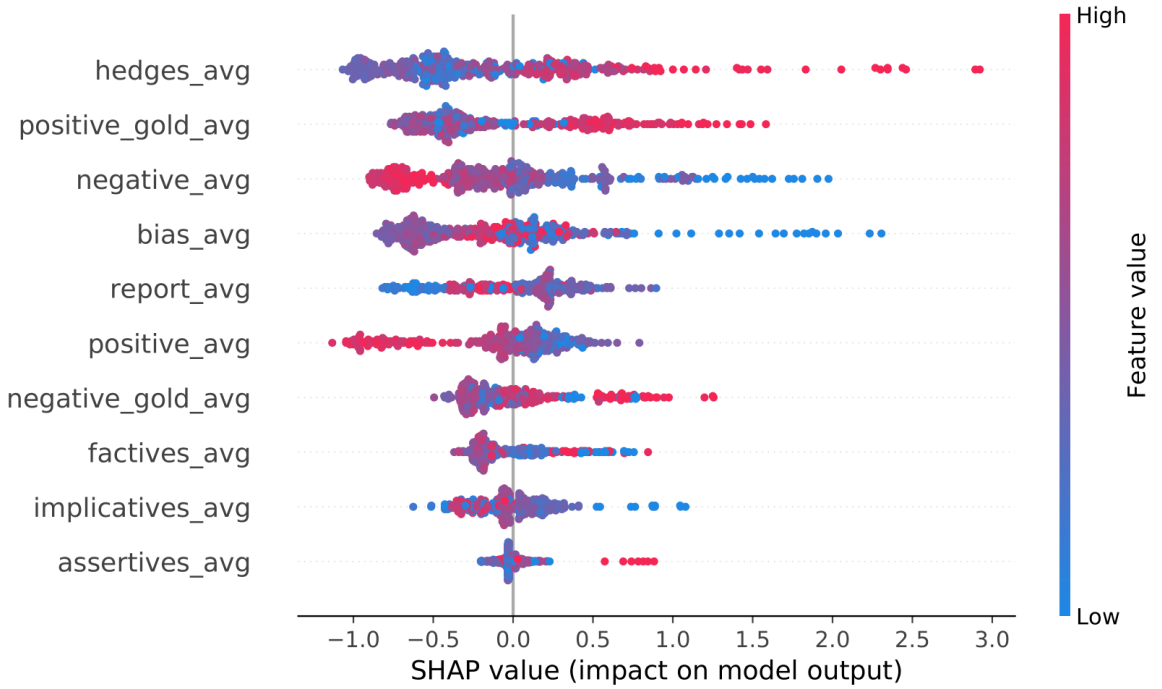

Figura 4.7: *Summary plot* gerado através do SHAP, exibindo o peso que as features exercem sobre a decisão de classificação do modelo utilizando as médias de subjetividade por documento como features para a classificação das notícias em Inglês.

A Figura 4.8 apresenta as explicações para o melhor modelo baseado em TFIDF. Este modelo reportou uma PR-AUC de 0,83. Na figura, é possível observar que o termo "donald", referente a Donald Trump é o mais decisivo para a decisão do modelo em classificar uma notícia como falsa. Este resultado demonstra um claro viés aprendido pelo modelo, onde documentos que tenham este termo tenderão a ser classificados como falso. Em um claro viés contrário, o segundo termo mais relevante é "cnn", onde todas as ocorrências fazem o modelo tender a classificar um documento como sendo real. Outro exemplo é o termo "hillari", que assim como o termo "donald", faz o modelo tender a classificar um documento como sendo falso. Estes exemplos, assim como os foi observado para as notícias em Português, demonstram um claro viés morfológico presente nos modelos clássicos para classificação de notícias falsas, onde os mesmos ao invés de apreenderem nuances factíveis relacionadas à problemática de notícias falsas, parecem aprender características que estão mais relacionadas à forma do texto, e não se o mesmo é realmente falso ou real. Este é um problema relevante pois, atualmente, os modelos clássicos de machine learning empregados para a classificação de notícias falsas, que reportam os melhores resultados, utilizam features baseadas em BoW e TFIDF (KHAN et al., 2019). Os resultados das explicações parecem demonstrar que estes modelos podem, na verdade, estar sofrendo de um enviesamento provocado pelas próprias base de dados de treino, onde os modelos reportam resultados a nível de estado da arte, mas podem não estar, de fato, aprendendo as reais nuances das notícias falsas.

4.7 Discussão dos Resultados

Os principais resultados apresentados são animadores no tocante à possíveis diferenças de subjetividade entre as notícias falsas e reais. Estas diferenças podem, em tese, ajudar na compreensão de características que de fato diferenciem estes dois tipos de notícias no âmbito textual.

Em relação à **QP1**, definida como: "A utilização de métodos de extração de subjetividade baseados em semântica permitem revelar diferenças significativas entre notícias falsas e reais?"; pode-se entender que a hipótese nula desta questão pôde ser rejeitada, dado que, nos experimentos realizados tanto para o Português quanto para o Inglês, foi possível encontrar diferenças significativas entre as distâncias semânticas reportadas pelo WMD para as senten-

ças das notícias falsas e reais. Estas diferenças foram significativamente mais evidentes para as notícias em Inglês (Tabela 4.12) quando comparadas com os testes de hipótese reportados para as notícias em Português (Tabela 4.4). Estas diferenças, consequentemente, refletiram no melhor desempenho dos modelos de classificação gerados para a classificação de notícias em Inglês. Esta melhora, como descrito em seções anteriores, pode estar relacionada com o uso de Word Embeddings construídos para representar especificamente o cenário de notícias, em contrapartida com os embeddings construídos a partir de um dump de artigos da Wikipédia, que foram utilizados para representação vetorial das notícias em Português. Outro ponto importante é que, no cenário de notícias em Inglês, foram utilizados léxicos já amplamente utilizados na literatura, bem como também foi possível utilizar mais léxicos de subjetividade (10 léxicos) em comparação com os experimentos em Português, em que foram utilizados 5 léxicos de subjetividade.

Para a **QP2**, que pergunta: "É possível determinar, de forma significativa, que notícias falsas são mais subjetivas que notícias reais?"; em quase todos os cenários observados, os léxicos de subjetividade indicaram que as notícias falsas apresentam níveis de subjetividade maiores que as notícias reais, para os casos onde houve diferenças significativas. Esse nível de subjetividade maior é representado por distâncias semânticas menores entre as sentenças das notícias falsas em relação aos léxicos de subjetividade. Ou seja, quanto menor esta distância, mais subjetivo seria o texto analisado. A exceção encontrada foi relacionada ao léxico "edges" nos experimentos em inglês. Este léxico representa termos que indicam um menor comprometimento do autor do texto com os fatos apresentados, sendo isso exemplificado pelos termos como: "apparently" e "could". Para este léxico, as notícias reais foram mais similares semanticamente, reportando distâncias significativamente menores em relação às sentenças falsas. Porém, uma hipótese para este caso é que, possivelmente, escritores e jornalistas de grandes veículos de imprensa, ao tentarem manter um certo nível de imparcialidade nos documentos, façam uso destes termos com o objetivo de reduzir o nível de comprometimento com fatos ainda não totalmente esclarecidos. Por outro lado, os autores de notícias falsas tenderiam a reportar fatos inverídicos como sendo totalmente verdadeiros, abrindo mão, em geral, do uso deste recurso linguístico.

Por fim, a questão de pesquisa **QP3** pergunta: "É possível, com uso das *features* de subjetividade propostas, construir modelos de classificação mais generalizáveis em relação

a modelos baseados em BoW?". Para esta questão, por meio dos experimentos utilizando o design de domínio-cruzado para as notícias em Português, foi possível observar que os modelos baseados nas features de subjetividade utilizando os Vetores de Subjetividade por Sentenças não apresentara perda de desempenho neste cenário, considerando o modelo XGBoost, quando comparado com o cenário sem o uso do domínio-cruzado. Porém, para todos os modelos avaliados, ao utilizarmos as features baseadas em BoW, houve perdas que chegaram a 30% em termos de PR-AUC, demonstrando que este tipo de modelo acaba tendo seu desempenho prejudicado em cenários onde o domínio das notícias varia entre os conjuntos de treino e teste.

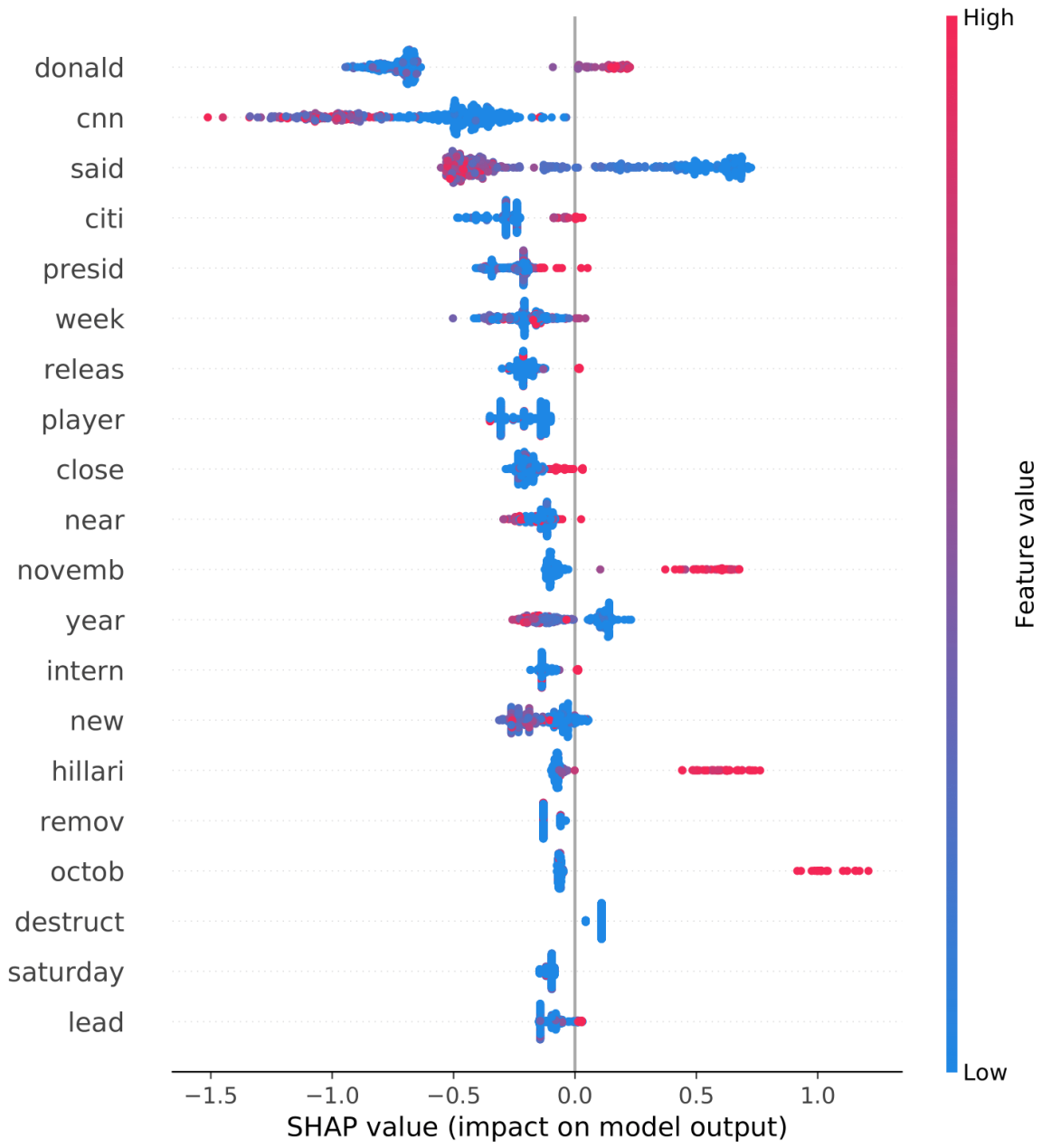

Figura 4.8: *Summary plot* gerado através do SHAP exibindo as features mais relevantes para a classificação do modelo baseado em BoW e TFIDF para as notícias em Inglês.

Capítulo 5

Conclusões e Trabalhos Futuros

Este capítulo se dedica a apresentar as principais conclusões desta pesquisa até o momento, bem como descreve possíveis trabalhos futuros que deverão ser executados em etapas futuras. Também é apresentado, ao final deste capítulo, um cronograma de atividades a serem executadas.

5.1 Conclusões

Este trabalho tem como foco principal, estudar como a subjetividade textual pode auxiliar no entendimento e também na identificação de notícias falsas de forma automática.

Para tal análise, é proposta uma abordagem baseada na segmentação de subjetividade por sentenças dos documentos, onde é gerado um vetor base que representa as distâncias semânticas de cada sentença de uma notícia em relação a um léxico de subjetividade. Esses vetores são utilizados para a geração de features de subjetividade que são utilizadas como base para a análise e classificação de notícias falsas e reais. Como principais resultados, foi possível observar diferenças significativas nos níveis de subjetividade entre estes dois conjuntos de notícias, tanto para experimentos utilizando documentos em Português, quanto para em Inglês. Apesar de apresentarem um desempenho de classificação significativamente inferior ao estado da arte, os modelos baseados em subjetividade ainda possuem uma ampla margem para melhorias, podendo, inclusive, serem empregados com outras features além da própria subjetividade.

Por fim, um dos grandes achados desta pesquisa foi a observação de que modelos clás-

sicos construídos para a classificação de notícias falsas podem, na verdade, estar sofrendo de vieses presentes na própria base de dados de treino. Esses vieses acabam permitindo a obtenção de excelentes resultados de classificação, porém, estes resultados parecem estar muito mais relacionados a tópicos e termos presentes nos documentos, do que com a captura, por parte dos modelos, de nuances que realmente estejam vinculadas ao contexto de notícias falsas. Estes resultados puderam ser observados ao utilizarmos modelos de classificação baseados em BoW e TFIDF, onde o desempenho destes modelos cai substancialmente quando se alterna os tópicos presentes nas bases de dados de treino e teste (design de domínio-cruzado). Para os modelos baseados em subjetividade, esta queda foi substancialmente menor, havendo casos onde não foi notada qualquer queda de desempenho, demonstrando uma possível maior generalização destes modelos. Estes achados puderam ser melhor compreendidos ao observar as explicações dos modelos geradas pelo SHAP, onde para os modelos clássicos baseados em BoW e TFIDF, fica evidente que estes acabam se tornando enviesados a determinados termos presentes nos documentos. Este enviesamento resulta em resultados expressivos de classificação, porém, pouco ajudam a compreender as reais nuances que permeiam a problemática de notícias falsas.

5.2 Trabalhos Futuros

Como trabalhos futuros, que devem dar continuidade a esta pesquisa, estão as seguintes atividades:

1. Revisar a literatura para identificar modelos e técnicas de identificação de notícias falsas baseadas em Deep Learning;
2. Estudo de diferentes métodos e técnicas para a extração de subjetividade em texto;
3. Estudar métodos de similaridade semântica mais eficientes que o WMD;
4. Incorporar diferentes base de dados de notícias falsas aos experimentos já executados;
5. Aplicar as features de subjetividade já desenvolvidas em modelos baseados em Deep Learning;

		Atividades							
Ano	**Mês**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2020	**OUT**	x							
2020	**NOV**	x	x						
2020	**DEZ**		x	x					
2021	**JAN**				x				
2021	**FEV**					x			
2021	**MAR**					x	x		
2021	**ABR**							x	
2021	**MAI**							x	x
2021	**JUN**								x
2021	**JUL**								x
2021	**AGO**								x

Tabela 5.1: Cronograma inicial para a execução de atividades futuras desta pesquisa.

6. Incorporar o uso de subjetividade a modelos “estado da arte” para verificar possíveis incrementos no desempenho destes modelos;

7. Produção de artigo científico para relatar os novos avanços na pesquisa;

8. Escrita da tese final de doutorado;

A Tabela 5.1 apresenta um cronograma inicial para a execução das atividades futuras.

Bibliografia

AHMED, H.; TRAORE, I.; SAAD, S. Detection of online fake news using n-gram analysis and machine learning techniques. In: TRAORE, I.; WOUNGANG, I.; AWAD, A. (Ed.). **Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments**. Cham: Springer International Publishing, 2017. p. 127–138. ISBN 978-3-319-69155-8.

AHMED, H.; TRAORE, I.; SAAD, S. Detection of online fake news using n-gram analysis and machine learning techniques. In: TRAORE, I.; WOUNGANG, I.; AWAD, A. (Ed.). **Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments**. Cham: Springer International Publishing, 2017. p. 127–138. ISBN 978-3-319-69155-8.

AKER, A.; GRAVENKAMP, H.; MAYER, S. J.; HAMACHER, M.; SMETS, A.; NTI, A.; ERDMANN, J.; SERONG, J.; WELPINGHUS, A.; MARCHI, F. Corpus of news articles annotated with article level subjectivity. In: **Workshop on Reducing Online Misinformation Exposure-ROME**. [S.l.: s.n.], 2019.

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. **Journal of Economic Perspectives**, v. 31, n. 2, p. 211–36, May 2017. Disponível em: <http://www.aeaweb.org/articles?id=10.1257/jep.31.2.211>.

ALLCOTT, H.; GENTZKOW, M. **Social Media and Fake News in the 2016 Election**. [S.l.], 2017. (Working Paper Series, 23089). Disponível em: <http://www.nber.org/papers/w23089>.

AMORIM, E.; CANÇADO, M.; VELOSO, A. Automated essay scoring in the presence of biased ratings. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 229–237. Disponível em: <https://www.aclweb.org/anthology/N18-1021>.

AMORIM, E.; CANÇADO, M.; VELOSO, A. Automated essay scoring in the presence of biased ratings. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 229–237. Disponível em: <https://www.aclweb.org/anthology/N18-1021>.

ASR, F. T.; TABOADA, M. Big data and quality data for fake news and misinformation detection. **Big Data & Society**, SAGE Publications Sage UK: London, England, v. 6, n. 1, p. 2053951719843310, 2019.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C. A neural probabilistic language model. **Journal of machine learning research**, v. 3, n. Feb, p. 1137–1155, 2003.

BOURGONJE, P.; SCHNEIDER, J. M.; REHM, G. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In: **Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 84–89. Disponível em: <https://www.aclweb.org/anthology/W17-4215>.

BURKHARDT, J. M. History of fake news. **Library Technology Reports**, v. 53, n. 8, p. 5–9, 2017. Disponível em: <https://journals.ala.org/index.php/ltr/article/viewFile/6497/8631>.

CHEN, C.-M.; TSAI, M.-F.; LIN, Y.-C.; YANG, Y.-H. Query-based music recommendations via preference embedding. In: **Proceedings of the 10th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2016. (RecSys '16), p. 79–82. ISBN 9781450340359. Disponível em: <https://doi.org/10.1145/2959100.2959169>.

CHEN, Y.; PEROZZI, B.; AL-RFOU, R.; SKIENA, S. The expressive power of word embeddings. **arXiv preprint arXiv:1301.3226**, 2013.

CHOI, Y.; DENG, L.; WIEBE, J. Lexical acquisition for opinion inference: A sense-level lexicon of benefactive and malefactive events. In: **Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 107–112. Disponível em: <https://www.aclweb.org/anthology/W14-2618>.

CHURCH, K. W.; MERCER, R. L. Introduction to the special issue on computational linguistics using large corpora. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 19, n. 1, p. 1–24, mar. 1993. ISSN 0891-2017.

DENG, L.; LIU, Y. A joint introduction to natural language processing and to deep learning. In: ____. **Deep Learning in Natural Language Processing**. Singapore: Springer Singapore, 2018. p. 1–22. ISBN 978-981-10-5209-5. Disponível em: <https://doi.org/10.1007/978-981-10-5209-5_1>.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

FERRARA, E.; VAROL, O.; DAVIS, C.; MENCZER, F.; FLAMMINI, A. The rise of social bots. **Commun. ACM**, ACM, New York, NY, USA, v. 59, n. 7, p. 96–104, jun. 2016. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2818717>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

HENRIQUES, R. P. O conceito de objetividade jornalística em luiz amaral e wilson gomes. **14º Encontro Nacional de Pesquisadores em Jornalismo**, 2016. Disponível em: <http://sbpjor.org.br/congresso/index.php/sbpjor/sbpjor2016/paper/viewFile/284/113>.

HORNE, B.; ADALI, S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: . [s.n.], 2017. Disponível em: <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15772/14898>.

JANZE, C.; RISIUS, M. Automatic detection of fake news on social media platforms. In: **PACIS**. [S.l.: s.n.], 2017. p. 261.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: NÉDELLEC, C.; ROUVEIROL, C. (Ed.). **Machine Learning: ECML-98**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 137–142. ISBN 978-3-540-69781-7.

JR, E. C. T.; LIM, Z. W.; LING, R. Defining "fake news" a typology of scholarly definitions. **Digital journalism**, Taylor & Francis, v. 6, n. 2, p. 137–153, 2018.

KHAN, J. Y.; KHONDAKER, M.; ISLAM, T.; IQBAL, A.; AFROZ, S. A benchmark study on machine learning methods for fake news detection. **arXiv preprint arXiv:1905.04749**, 2019.

KINCAID, J. P.; JR, R. P. F.; ROGERS, R. L.; CHISSOM, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida, 1975.

KUSNER, M. J.; SUN, Y.; KOLKIN, N. I.; WEINBERGER, K. Q. From word embeddings to document distances. In: **Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37**. JMLR.org, 2015. (ICML'15), p. 957–966. Disponível em: <http://dl.acm.org/citation.cfm?id=3045118.3045221>.

LAZER, D. M.; BAUM, M. A.; BENKLER, Y.; BERINSKY, A. J.; GREENHILL, K. M.; MENCZER, F.; METZGER, M. J.; NYHAN, B.; PENNYCOOK, G.; ROTHSCHILD, D. et al. The science of fake news. **Science**, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018.

LE, N. Q. K.; YAPP, E. K. Y.; HO, Q.-T.; NAGASUNDARAM, N.; OU, Y.-Y.; YEH, H.-Y. ienhancer-5step: Identifying enhancers using hidden information of dna sequences via chou's 5-step rule and word embedding. **Analytical Biochemistry**, v. 571, p. 53 – 61, 2019. ISSN 0003-2697. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0003269719300788>.

LEE, C.; SHIN, J.; HONG, A. Does social media use really make people politically polarized? direct and indirect effects of social media use on political polarization in south korea. **Telematics and Informatics**, v. 35, n. 1, p. 245 – 254, 2018. ISSN 0736-5853. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0736585317305208>.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information**

Processing Systems 30. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Ann. Math. Statist.**, The Institute of Mathematical Statistics, v. 18, n. 1, p. 50–60, 03 1947. Disponível em: <https://doi.org/10.1214/aoms/1177730491>.

MARCHI, R. With facebook, blogs, and fake news, teens reject journalistic "objectivity". **Journal of Communication Inquiry**, SAGE Publications Sage CA: Los Angeles, CA, v. 36, n. 3, p. 246–262, 2012.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: BURGES, C. J. C.; BOTTOU, L.; WELLING, M.; GHAHRAMANI, Z.; WEINBERGER, K. Q. (Ed.). **Advances in Neural Information Processing Systems 26**. Curran Associates, Inc., 2013. p. 3111–3119. Disponível em: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.

MONTEIRO, R. A.; SANTOS, R. L.; PARDO, T. A.; ALMEIDA, T. A. de; RUIZ, E. E.; VALE, O. A. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 324–334.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–551, 09 2011. ISSN 1067-5027. Disponível em: <https://doi.org/10.1136/amiajnl-2011-000464>.

OLSON, R. S.; CAVA, W. L.; MUSTAHSAN, Z.; VARIK, A.; MOORE, J. H. Data-driven advice for applying machine learning to bioinformatics problems. **arXiv preprint arXiv:1708.05070**, World Scientific, 2017.

PAGE, E. Grading essays by computer: progress report. In: **Proceedings of the Invitational Conference on Testing Problems**. [S.l.: s.n.], 1967. p. 87–100.

PENNEBAKER, J. W.; BOYD, R. L.; JORDAN, K.; BLACKBURN, K. **The development and psychometric properties of LIWC2015**. [S.l.], 2015.

PÉREZ-ROSAS, V.; KLEINBERG, B.; LEFEVRE, A.; MIHALCEA, R. Automatic detection of fake news. In: **Proceedings of the 27th International Conference on Computational Linguistics**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 3391–3401. Disponível em: <https://www.aclweb.org/anthology/C18-1287>.

RASHKIN, H.; CHOI, E.; JANG, J. Y.; VOLKOVA, S.; CHOI, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In: **Proceedings of the 2017**

Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 2931–2937. Disponível em: <https://www.aclweb.org/anthology/D17-1317>.

RECASENS, M.; DANESCU-NICULESCU-MIZIL, C.; JURAFSKY, D. Linguistic models for analyzing and detecting biased language. In: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2013. p. 1650–1659.

REIS, J. C. S.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. Explainable machine learning for fake news detection. In: **Proceedings of the 10th ACM Conference on Web Science**. New York, NY, USA: Association for Computing Machinery, 2019. (WebSci '19), p. 17–26. ISBN 9781450362023. Disponível em: <https://doi.org/10.1145/3292522.3326027>.

RUCHANSKY, N.; SEO, S.; LIU, Y. Csi: A hybrid deep model for fake news detection. In: **Proceedings of the 2017 ACM on Conference on Information and Knowledge Management**. [S.l.: s.n.], 2017. p. 797–806.

SCHUDSON, M. The objectivity norm in american journalism*. **Journalism**, v. 2, n. 2, p. 149–170, 2001. Disponível em: <https://doi.org/10.1177/146488490100200201>.

SHU, K.; SLIVA, A.; WANG, S.; TANG, J.; LIU, H. Fake news detection on social media: A data mining perspective. **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 19, n. 1, p. 22–36, set. 2017. ISSN 1931-0145. Disponível em: <https://doi.org/10.1145/3137597.3137600>.

SILVERMAN, C. This analysis shows how viral fake election news stories outperformed real news on facebook. **BuzzFeed news**, v. 16, 2016.

TUCHMAN, G. A objectividade como ritual estratégico: uma análise das noções de objectividade dos jornalistas. **Jornalismo: questões, teorias e "estórias". Lisboa: Vega**, v. 2, p. 74–90, 1993.

VOLKOVA, S.; SHAFFER, K.; JANG, J. Y.; HODAS, N. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. [S.l.: s.n.], 2017. p. 647–653.

Wang, L.; Wang, Y.; de Melo, G.; Weikum, G. Five shades of untruth: Finer-grained classification of fake news. In: **2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. [S.l.: s.n.], 2018. p. 593–594.

WANG, M.; CAO, D.; LI, L.; LI, S.; JI, R. Microblog sentiment analysis based on cross-media bag-of-words model. In: **Proceedings of International Conference on Internet Multimedia Computing and Service**. New York, NY, USA: Association for Computing Machinery, 2014. (ICIMCS '14), p. 76–80. ISBN 9781450328104. Disponível em: <https://doi.org/10.1145/2632856.2632912>.

WANG, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection.

arXiv preprint arXiv:1705.00648, 2017.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing contextual polarity in phrase-level sentiment analysis. In: **Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing**. USA: Association for Computational Linguistics, 2005. (HLT '05), p. 347–354. Disponível em: <https://doi.org/10.3115/1220575.1220619>.

ZHOU, X.; JAIN, A.; PHOHA, V. V.; ZAFARANI, R. Fake news early detection: A theory-driven model. **arXiv preprint arXiv:1904.11679**, 2019.

ZHOU, X.; ZAFARANI, R. Fake news: A survey of research, detection methods, and opportunities. **arXiv preprint arXiv:1812.00315**, 2018.