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

Classificações de Notícias Falsas Baseadas em Similaridade Semântica a partir de Léxicos Automaticamente Construídos

Caio Libânio Melo Jerônimo

Campina Grande, Paraíba, Brasil

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

Classificações de Notícias Falsas Baseadas em Similaridade Semântica a partir de Léxicos Automaticamente Construídos

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

J56c Jerônimo, Caio Libânio Melo.
Classificações de notícias falsas baseadas em similaridade semântica a partir de léxicos automaticamente construídos / Caio Libânio Melo Jerônimo. – Campina Grande, 2022.
105 f. : il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2022.
"Orientação: Prof. Dr. Leandro Balby Marinho, Prof. Dr. Cláudio Elízio Calazans Campelo".
Referências.

1. Notícias Falsas. 2. Similiridade Semântica. 3. Classificação. I. Marinho, Leandro Balby. II. Campelo, Cláudio Elízio Calazans. III. Título.

CDU 025.4(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA MARIA ANTONIA DE SOUSA CRB 15/398

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

CAIO LIBÂNIO MELO JERÔNIMO

CLASSIFICAÇÕES DE NOTÍCIAS FALSAS BASEADAS EM SIMILARIDADE SEMÂNTICA A PARTIR DE LÉXICOS AUTOMATICAMENTE CONSTRUÍDOS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Doutor em Ciência da Computação.

Aprovada em: 07/02/2022

Prof. Dr. CLÁUDIO ELÍZIO CALAZANS CAMPELO, UFCG, Orientador

Prof. Dr. LEANDRO BALBY MARINHO, UFCG, Orientador

Prof. Dr. NAZARENO FERREIRA DE ANDRADE, UFCG, Examinador Interno

Prof. Dr. FÁBIO JORGE ALMEIDA MORAIS, UFCG, Examinador Interno

Prof. Dr. FABRÍCIO BENEVENUTO DE SOUZA, UFMG, Examinador Externo

Prof. Dr. RINALDO JOSÉ DE LIMA, UFRPE, Examinador Externo

Documento assinado eletronicamente por **NAZARENO FERREIRA DE ANDRADE**, **PROFESSOR 3 GRAU**, em 07/02/2022, às 17:31, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 07/02/2022, às 18:05, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **FABIO JORGE ALMEIDA MORAIS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 07/02/2022, às 19:04, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **LEANDRO BALBY MARINHO**, **PROFESSOR 3**

GRAU, em 08/02/2022, às 09:32, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **2094155** e o código CRC **802064D8**.

Referência: Processo nº 23096.001677/2022-62 SEI nº 2094155

Resumo

Métodos de detecção de notícias falsas baseados unicamente em características textuais permitem uma detecção precoce deste tipo de conteúdo. Esta estratégia de detecção não necessita de informações como o número de curtidas ou quantidade de compartilhamentos, informações disponíveis apenas quando a notícia já tem se disseminado nas redes sociais. Dentro deste escopo, o uso de léxicos como recurso para auxiliar na construção de *features* de classificação se destaca por ser um recurso capaz de agregar um conhecimento prévio ao processo de classificação. Porém, a construção deste tipo de recurso muitas vezes exige a participação de especialistas no processo, o que em muitos contextos, torna o processo muito custoso ou mesmo inviável. Nesta pesquisa, é proposto um método para a construção automática de léxicos voltados para a análise e classificação de notícias falsas. O método proposto utiliza documentos de notícias falsas e reais, onde são extraídos termos que auxiliam na diferenciação destes dois tipos de documentos. Também é proposta, a partir dos léxicos gerados, uma estratégia para a construção de *features* de classificação baseados em similaridade semântica. Nesta pesquisa, avaliamos e comparamos modelos treinados a partir dos léxicos gerados automaticamente com modelos treinados utilizando léxicos já presentes na literatura. Como principais resultados, foi possível verificar que os modelos que utilizaram os léxicos construídos nesta pesquisa se mostraram superiores em diferentes cenários, como também apresentaram, de forma sistemática, melhores resultados quando utilizados em conjunto com os léxicos já existentes na literatura. Por fim, é apresentada uma análise da explicabilidade dos modelos, permitindo revelar nuances das notícias falsas que só puderam ser observadas com o auxílio dos léxicos gerados nesta pesquisa.

Abstract

Fake news detection methods based on textual features allow early detection of this type of content. This detection strategy does not need information such as the number of likes or the number of shares, informations only available when the news has already been disseminated on social networks. Within this scope, the use of lexicons as a resource to assist in the construction of classification *features* stands out for being a resource capable of adding prior knowledge to the classification process. However, the construction of this type of resource often requires the participation of specialists in the process, which in many contexts makes the process very costly or even unfeasible. In this research, a method for the automatic construction of fake news lexicons is proposed. The proposed method uses false and real news documents, where terms that help to differentiate these two types of documents are extracted. It is also proposed, from the generated lexicons, a strategy for the construction of classification *features* based on semantic similarity. In this research, we evaluate and compare models trained with the constructed lexicons and compare them with models trained with lexicons already present in literature. As main results, it was possible to verify that the models that use the generated lexicons were superior in different scenarios, as well as presenting better results when used in conjunction with the lexicons that are present in literature. Finally, an explainable analysis of the models is presented, allowing to reveal nuances of fake news that could only be observed with the help of the lexicons generated in this research.

Agradecimentos

Agradeço aos meus pais, Magna e Neucimar (*in memoriam*), por terem criado, ao longo de suas vidas, todas as condições necessárias para que eu pudesse chegar até aqui.

Aos meus orientadores, Cláudio Elízio Calazans Campelo e Leandro Balby Marinho, agradeço pela paciência e por todo o conhecimento compartilhado ao longo desta longa jornada acadêmica.

Agradeço a todo o conhecimento que os colegas do Laboratório de Computação Inteligente Aplicada (LACINA) puderam compartilhar comigo. Certamente, todas as discussões que tivemos contribuíram para o desenvolvimento deste trabalho.

Sou grato aos professores e demais funcionários da Universidade Federal de Campina Grande, do Centro de Engenharia Elétrica e Informática, do Departamento de Sistemas e Computação e da Coordenação de Pós-Graduação em Ciência da Computação que apoiaram de forma direta ou indireta a realização deste trabalho.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Uma notícia enganosa pode ser 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). Porém, um conteúdo enganoso pode ser disseminado ainda por outros meios, como mensagens de aplicativos, áudios e imagens. Nos últimos tempos, este tipo de conteúdo veio se popularizando com o termo em inglês de *Fake News* ou mesmo notícia falsa. Esta nomenclatura é essencialmente imprecisa pois, a rigor, uma notícia sempre deveria ser verdadeira e verificada. Essas "notícias" usualmente tentam imitar o formato de uma notícia genuína. 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 título de simplificação e acessibilidade, esta tese irá usar a nomenclatura mais popular de "notícia falsa" para se referir a documentos de notícia que buscam, deliberadamente, enganar o leitor.

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 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

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

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 presentes 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. Mais recentemente, trabalhos que utilizam abordagens híbridas (REIS et al., 2019b) vêm ganhado destaque na literatura. Grande parte das pesquisas que investigam as características textuais de notícias falsas utilizam modelos de Aprendizado de Máquina considerando aspectos como frequência de ocorrência de determinadas palavras e tamanho dos documentos (AHMED; TRAORE; SAAD, 2017a; BOURGONJE; SCHNEIDER; REHM, 2017; HORNE; ADALI, 2017; WYNNE; WINT, 2019; FAUSTINI; COVõES, 2020). Tais modelos de classificação de notícias falsas reportam resultados expressivos, com acurácias atingindo cerca de 95% (KHAN et al., 2019). Mais recentemente, trabalhos voltados para classificação de notícias falsas baseados em *Deep Learning* têm representado a vanguarda com soluções baseadas em tecnologias indo desde Redes Neurais Recorrentes (RNN) à *Trasformers* (CHOUDHARY et al., 2021; WANI et al., 2021; MONTI et al., 2019).

Desde o fim do período eleitoral americano de 2016, onde houve uma grande demanda por pesquisas acerca de notícias falsas, muitos caminhos e possibilidades foram e continuam sendo amplamente exploradas na literatura. Porém, alguns pontos ainda continuam em aberto e pouco explorados. Um destes pontos é a construção automática de léxicos específicos que permitam, além da classificação de notícias falsas, entender melhor as nuances que possam estar ocultas nos documentos. Essencialmente, um léxico consiste em um conjunto definido de termos, ou *tokens* que possuem alguma característica específica. Neste contexto, pesquisas que utilizam modelos de Aprendizado de Máquina também chegam a utilizar léxicos criados manualmente, porém não especificamente para classificação deste tipo de conteúdo. Esta pesquisa apresenta um método para construção automática de léxicos

voltados para classificação de notícias falsas, bem como uma estratégia para, a partir dos léxicos gerados automaticamente, construir *features* baseadas em similaridade semântica que permitem classificar notícias falsas e reais. Também é apresentado um estudo sobre como estas *features* de classificação podem auxiliar na explicabilidade dos modelos construídos. Dessa forma, os léxicos construídos podem auxiliar tanto na construção de modelos preditivos mais eficazes, como também na compreensão de aspectos ainda pouco explorados acerca deste tipo de documento.

1.1 Motivação e Justificativa

A necessidade de modelos que permitam a detecção de notícias falsas de forma efetiva é premente, 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). O atual cenário de polarização 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 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].

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

Cenários como o descrito no parágrafo acima 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 deste tipo de conteúdo antes que ele possa se disseminar largamente pela rede.

Na literatura, grande parte das pesquisas que utilizam técnicas de Inteligência Artificial para a classificação de notícias utilizam algoritmos clássicos de Aprendizado de Máquina (CONROY; RUBIN; CHEN, 2015; VOLKOVA et al., 2017; PÉREZ-ROSAS et al., 2018) destacando-se abordagens que utilizam *Bag-of-Words* (BoW) como principais *features* de classificação. Mais recentemente, pesquisas utilizando *Deep Learning* (KUMAR et al., 2020; WANG, 2017; RUCHANSKY; SEO; LIU, 2017; VASWANI et al., 2017) vêm ganhando destaque, em especial, as abordagens utilizando *Transformers* (CHEN et al., 2021; VASWANI et al., 2017).

Um caminho ainda pouco explorado na literatura é a construção automática de léxicos para classificação de notícias falsas. É fato que, diversos trabalhos buscam formas de construção de léxicos, porém, o escopo predominante de tais trabalhos está voltado para a construção de léxicos de sentimentos (DARWICH et al., 2019; TAI; KAO, 2013).

O uso de léxicos construídos manualmente permite adicionar conhecimento especializado a soluções que envolvem o Processamento de Linguagem Natural (PLN) (GUTHRIE et al., 1996). Por exemplo, um léxico construído por um especialista em linguagem pode auxiliar na construção de modelos de predição de *reviews* de produtos. Apesar do uso e construção de léxicos ser uma escopo de estudo bastante explorado em áreas da PLN como análise de sentimentos e opiniões, no contexto de notícias falsas ainda é pouco explorado. Neste cenário, a construção automática de léxicos que atendam a um determinado escopo (e.g. notícias falsas) pode permite, por exemplo, a construção de léxicos sem, necessariamente, a colaboração de especialistas, bem como permitir o aprofundamento de estudos

acerca do problema em análise.

Nesta tese, é apresentado um arcabouço para a classificação de notícias falsas baseado em três eixos principais: (i) construção automática de léxicos voltados para a análise e classificação de notícias falsas; (ii) construção de novas *features* de classificação baseadas nos léxicos construídos; e (iii) uma análise explicativa acerca dos modelos gerados e como eles podem auxiliar no entendimento de nuances presentes nas notícias falsas.

No que tange a estratégia de construção dos léxicos (eixo i), nossa abordagem é inspirada em pesquisas realizadas por Reis et al. (2019b) onde a abordagem original consiste em um *framework* genérico de construção de léxicos baseado em uma estratégia gulosa. Essa estratégia foi primeiro usada pelo projeto Parlametria[6], onde nosso grupo de pesquisa atuou. No referido projeto, no entanto, são identificados termos capazes de prever a popularidade de postagens no Twitter[7] de parlamentares brasileiros. Nesta presente pesquisa, o algoritmo de construção de léxicos é adaptado para a construção de léxicos e *features* que podem ser usados em diversos problemas de classificação de documentos, entre eles, o de classificação de notícias falsas.

De forma geral, o problema que esta tese busca investigar é: "Como gerar léxicos de notícias falsas de forma automática, e que estes possam ser usados para auxiliar na classificação e análise deste tipo de conteúdo enganoso?"

1.2 Questões de Pesquisa

Com o objetivo de guiar esta pesquisa dentro do escopo descrito anteriormente, e permitindo validar os pontos pontos fundamentais do que é proposto nesta tese, foram estabelecidas as seguintes questões de pesquisa. No restante desta pesquisa, os léxicos gerados utilizando a abordagem proposta serão referenciados, para fins de simplificação, apenas como "LG".

QP1 É possível encontrar diferenças significativas entre notícias falsas e reais utilizando as *features* construídas a partir dos LG?

H1-0 As *features* de classificação geradas a partir dos LG não conseguem diferenciar, de forma significativa, notícias falsas e reais.

[6]https://parlametria.org/home [7]https://twitter.com/

H1-1 As *features* de classificação geradas a partir dos LG conseguem diferenciar, de forma significativa, notícias falsas e reais.

Como descrito, esta pesquisa busca apresentar um arcabouço que contempla a construção automática de léxicos voltados para o escopo de notícias falsas, bem como a utilização destes léxicos para a construção de *features* de classificação. Estas *features* de classificação são baseadas em similaridade semântica e serão utilizadas para a construção de modelos preditivos. A questão Q1 busca, de início, analisar se as *features* geradas a partir dos LG permitem revelar, de forma significativa, diferenças entre notícias falsas e reais. Esta possível diferenciação será fundamental para a construção de modelos preditivos capazes de classificar notícias falsas e reais.

QP2 Modelos preditivos treinados a partir dos LG possuem desempenho superior a modelos treinados a partir de léxicos construídos manualmente?

H2-0 Modelos preditivos treinados a partir dos LG **não** possuem desempenho superior a modelos treinados a partir de léxicos construídos manualmente.

H2-1 Modelos preditivos treinados a partir dos LG **possuem** desempenho superior a modelos treinados a partir de léxicos construídos manualmente.

A questão Q2 já busca verificar se os modelos preditivos construídos utilizando as *features* construídas a partir dos LG apresentam um desempenho de classificação superior a modelos treinados tendo como base léxicos construídos manualmente e já usados na literatura dentro do escopo de notícias falsas. Para fins de simplificação, estes léxicos serão referenciados neste documento apenas como *baselines*. O objetivo central desta questão é de comparar o poder preditivo dos léxicos gerados automaticamente na pesquisa, com léxicos já existentes na literatura.

QP3 Modelos preditivos treinados com base nos LG + *baseline* podem obter um melhor desempenho de classificação de notícias falsas, quando comparados com os outros modelos gerados na pesquisa?

H3-0 Modelos preditivos treinados com base nos LG + *baseline* **não apresentam** um melhor desempenho de classificação de notícias falsas, quando comparados com os outros modelos gerados na pesquisa.

H3-1 Modelos preditivos treinados com base nos LG + *baseline* **apresentam** um melhor desempenho de classificação de notícias falsas, quando comparados com os outros modelos gerados na pesquisa.

A questão Q3 verifica a hipótese de uma possível melhoria geral no desempenho de modelos preditivos quando treinados usando as *features* obtidas a partir dos LG em conjunto com os léxicos usados como *baseline*. Em outras palavras, será avaliado o desempenho de modelos utilizando LG + *baseline*. Esta questão visa avaliar possíveis sobreposições semânticas presentes entres os léxicos. Por exemplo, caso os léxicos gerados nesta pesquisa apenas adicionem elementos já existentes nos *baselines*, seria esperado não haver grandes melhorias nas classificações, quando comparado com modelos treinados utilizando apenas os léxicos usados como *baseline*.

1.3 Objetivos

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

1.3.1 Objetivo Geral

Esta tese tem como objetivo central propor uma abordagem de classificação de notícias falsas baseada na construção automática de léxicos e extração de *features* de classificação.

1.3.2 Objetivos Específicos

Os seguintes objetivos específicos são considerados para atender o objetivo central desta tese:

- Adaptação de método usado em outras aplicações (e.g. análise de discursos presentes no Twitter) para a construção de léxicos de notícias falsas.
- Implementar método baseado em similaridade semântica para gerar *features* de classificação baseadas nos léxicos gerados na pesquisa.
- Realizar levantamento de léxicos presentes na literatura para uso como *baseline* para avaliação dos léxicos gerados na pesquisa.

- Avaliar o poder preditivo dos léxicos gerados automaticamente, comparando-os com outros construídos manualmente e já utilizados na literatura.

- Realizar estudo de explicabilidade dos modelos preditivos construídos nesta pesquisa.

1.4 Contribuições

O uso de léxicos para atividades de PLN já é amplamente difundido na literatura. Dentro do contexto de notícias falsas, léxicos construídos para diversos fins são utilizados para classificação de notícias falsas. Porém, poucos trabalhos buscam efetivamente construir léxicos voltados para este domínio.

Como principal contribuição desta tese, está a apresentação de uma abordagem voltada para a classificação de notícias falsas baseada em léxicos construídos automaticamente. Como já mencionado neste Capítulo, a abordagem apresentada é composta de três eixos centrais. No primeiro eixo, o de construção de léxicos, é apresentada uma estratégia para a construção automática de léxicos voltada para a classificação de documentos textuais. O método apresentado pode ser utilizado para a construção de léxicos de classificação textual em qualquer domínio. Porém, nesta pesquisa, a estratégia é utilizada e adaptada para a construção de léxicos voltados para a classificação de notícias falsas. No segundo eixo, o de construção de *features* de classificação, é proposto, a partir dos léxicos já construídos, um método para a construção de *features* de classificação baseadas em similaridade semântica entre documentos. Estas *features* buscam aproveitar o fato de que termos presentes em um léxico já carregam algum nível de similaridade entre si, o que facilita a construção de explicações mais intuitivas. No terceiro eixo desta pesquisa, é apresentada uma análise explicativa dos modelos preditivos construídos. Esta análise visa explorar como os léxicos e as *features* construídas na pesquisa podem auxiliar no entendimento de nuances presentes nas notícias falsas.

1.4.1 Contribuições Bibliográficas

O desenvolvimento desta pesquisa contribuiu com as seguintes publicações:

- Jeronimo, C. L. M., Marinho, L. B., Campelo, C. E., Veloso, A., & da Costa Melo,

A. S. (2019, December). Fake news classification based on subjective language. In Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (pp. 15-24).

- Vieira, L. L., Jeronimo, C. L. M., Campelo, C. E., & Marinho, L. B. (2020, November). Analysis of the Subjectivity Level in Fake News Fragments. In Proceedings of the Brazilian Symposium on Multimedia and the Web (pp. 233-240).

- Jeronimo, C. L., Campelo, C. E., Marinho, L. B., Sales, A., Veloso, A., & Viola, R. (2020, May). Computing with Subjectivity Lexicons. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 3272-3280).

- Vasconcelos, L., Campelo, C., & Jeronimo, C. (2020, May). Aspect Flow Representation and Audio Inspired Analysis for Texts. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 1469-1477).

- JERONIMO, C. L., MARINHO, L. B., CAMPELO, C. E., VELOSO, A., & MELO, A. S. D. C. (2020). CHARACTERIZATION OF FAKE NEWS BASED ON SUBJECTIVITY LEXICONS. Journal of Data Intelligence, 1(4), 419-441.

1.5 Estrutura do Documento

O restante do documento está organizado da seguinte maneira. No Capítulo 2, são apresentadas as Fundamentações Teóricas, que descrevem conceitos básicos que permeiam esta pesquisa, auxiliando assim, o entendimento de conceitos importantes. No Capítulo 3, é apresentada uma ampla revisão da literatura que abrange tópicos fundamentais desta pesquisa. No Capítulo 4, o método proposto é apresentado, bem como são descritas as metodologias de avaliação que serão executadas. No Capítulo 5, são apresentados os principais resultados encontrados nesta pesquisa. O Capítulo 6 apresenta as conclusões, limitações e trabalhos futuros.

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 consensual. 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 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 facilmente 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 facilmente 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, de forma maliciosa, 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 tende a se parecer com uma publicidade enganosa, porém, este último está mais direcionado à venda de produtos ou ganho financeiro.

A Figura 2.1 exibe um exemplo de notícia falsa publicada ainda no âmbito da eleição presidencial de 2018. Na imagem, é possível notar, em destaque, partes da notícia que apresentam algum indício de se tratar de um conteúdo duvidoso (ZHANG; GHORBANI, 2020). Sendo elas as seguintes:

1. **Título da notícia**: Logo no título, é possível observar elementos chamativos em caixa alta, como "A CASA CAIU". Tal característica busca chamar a atenção do leitor de imediato;

2. **Autoria da notícia**: É comum, em documentos de notícias falsas, não ser possível identificar o autor nem mesmo a data real de publicação da notícia;

3. **Fotos**: Em documentos de notícias falsas, é muito comum se observar fotos retiradas de outro contexto, sem que seja possível identificar, de fato, as pessoas que estão presentes na imagem;

4. **Corpo da notícia**: No corpo das notícias falsas, também é comum se observar a presença de termos e expressões de exagero, com o objetivo de impactar o leitor.

2.1.2 Objetividade Jornalística e 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 àqueles que, dominando a leitura e

Figura 2.1: Exemplo de notícia falsa supostamente publicada no âmbito das eleições de 2018. Na imagem, é possível observar, em destaque, partes da notícia que possuem pistas de que se trate de uma notícia falsa.

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, 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 (SALES; BALBY; VELOSO, 2019).

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.

2.1.3 Identificação de Notícias Falsas Baseada em Conteúdo Textual

A área de Processamento de Linguagem Natural (PLN) 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; NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). Atualmente, o problema de análise e identificação de notícias falsas representa um novo desafio para a PLN. Esse desafio vem demandando grandes esforços e investimentos em pesquisas que busquem construir ferramentas que possam tratar este problema (SHU et al., 2020; OSHIKAWA; QIAN; WANG, 2020).

A identificação de notícias falsas baseada em conteúdo textual consiste em um conjunto de estratégias que permitem, a partir do conteúdo textual das notícias, classificar se o documento é falso ou real. Esta estratégia considera que o conteúdo de notícias falsas e reais são, em alguma medida, diferentes entre si a nível puramente textual.

Sharma et al. (2019) apresentam algumas subdivisões para métodos envolvendo a identificação de notícias falsas baseadas em conteúdo. Dentre elas:

1. **Métodos baseados características lexicais:** Este conjunto de técnicas considera estratégias de classificação de notícias falsas baseadas em *features* como palavras-chave, uso de léxicos pré-definidos e quantidade de palavras/sentenças em um documento.

2. **Métodos baseados em análise linguística:** Técnicas presentes nesta categoria con-

sideram a identificação de notícias falsas baseadas em *features* extraídas a partir de representações no formato de *n-gramas*, que consistem em uma representação contínua de termos presentes em um texto. Outra estratégia de representação utilizada é a baseada em *Part-of-Speech*. Este tipo de representação considera um texto a partir de suas características sintáticas, como quantidade de verbos no documento. Também existem as representações textuais geradas a partir de Gramáticas Livre de Contexto. Nesta representação, é possível gerar, a partir de uma sentença, uma estrutura em formato de árvore, onde são representadas a estrutura sintática do texto.

3. **Métodos baseados em Aprendizagem Profunda:** Métodos baseados em Aprendizagem Profunda, ou *Deep Learning*, representam, de forma geral, o estado-da-arte na classificação de notícias falsas. Estes métodos dispensam processos custosos de engenharia de *features*, permitindo que as próprias redes neurais aprendam as características que melhor diferenciam notícias falsas e reais. As arquiteturas mais comuns para tratar este tipo de problema são as Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN). Mais recentemente, novas arquiteturas baseadas em *Transformers*, como o *Bidirectional Encoder Representations from Transformers* (BERT) (DEVLIN et al., 2018) vêm se destacando em problemas de classificação de noticias falsas.

2.2 Técnicas de Processamento de Linguagem Natural

Esta seção apresenta as principais abordagens de PLN e demais métodos utilizadas nesta pesquisa.

2.2.1 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 pode representar, por exemplo, a ocorrência ou

ausência de uma palavra no documento. 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.

2.2.2 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.

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 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.

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 semântica, morfológico de contexto, ou qualquer outro relacionamento que possa estar presente dentro do *corpus* de criação da representação. A ideia central é que palavras que sejam semanticamente similares, possuirão uma representação vetorial também similar. Isso permite, por

exemplo, a construção de representações textuais que tenham como base o campo semântico do documento, ao invés de apenas considerar a ocorrência pontual de termos ou palavras no documento, como ocorre em representações comuns baseadas em BoW.

Atualmente, WE vêm sendo utilizado em diferentes cenários da PLN. Chen et al. (2013b) 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. (2013). 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, é utilizado um conjunto de vetores (i.e. *embeddings* pré-treinados a partir de um banco de notícias do Google[1], utilizando a implementação *word2vec* e treinados utilizando o algoritmo *Continuous Bag of Words Model* (CBOW) (MIKOLOV et al., 2013). Utilizando este algoritmo, a principal tarefa de uma rede neural utilizada para a geração de *embeddings* é prever uma dada palavra, a partir dos termos que estão ao redor dela, com base em uma janela pré-definida. A Figura 2.2 apresenta um esquema básico desta arquitetura, onde, através de uma rede neural, a representação de um termo é "aprendida" a partir dos termos que o circundam, com base em uma janela pré-definida.

2.2.4 Similaridade Semântica via Tópicos Ocultos

Tipicamente, abordagens que buscam calcular a similaridade entre dois documentos enfrentam dois principais problemas. O primeiro, está relacionado com eventuais diferenças no tamanho dos documentos que deverão ser analisados. Abordagens como a Word Mover's

[1] https://code.google.com/archive/p/word2vec/

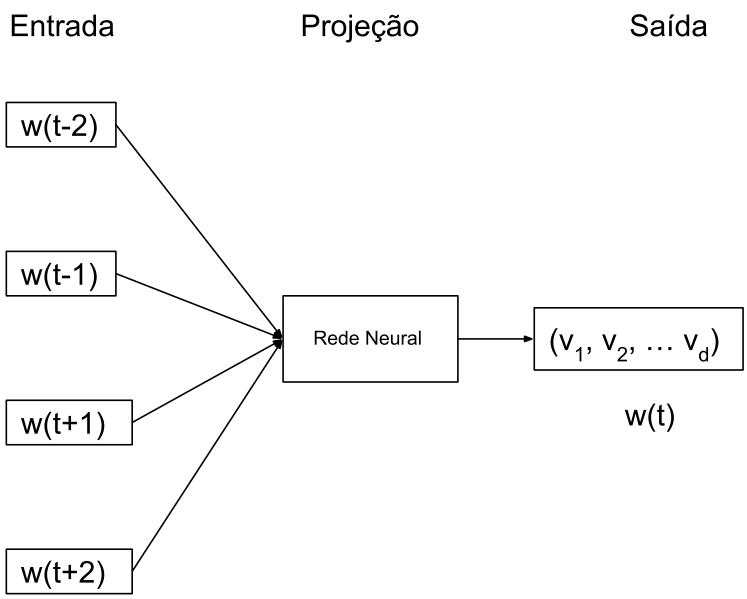

Figura 2.2: Arquitetura básica para geração de embeddings utilizando word2vec e CBOW. O objetivo é utilizar os pesos aprendidos por uma rede neural para representar uma palavra, dado os termos que a circundam, ou seja, seu contexto (MIKOLOV et al., 2013).

Distance (WMD) (KUSNER et al., 2015) sofrem diretamente este tipo de impacto, onde variações nos tamanhos dos documentos afetam, diretamente, os resultados obtidos, gerando possíveis vieses relacionados ao tamanho dos documentos em análise. O segundo problema está relacionado à invariável presença de termos de pouca relevância semântica nos documentos, gerando assim, ruídos nas análises. Abordagens como Do2Vec (LE; MIKOLOV, 2014) sofrem com este tipo de problema.

Buscando resolver os dois problemas descritos, o método apresentado por Gong et al. (2018), que é utilizado e referenciado nesta pesquisa como *Hidden Topics*, implementa uma abordagem para o cálculo de similaridade semântica entre documentos textuais considerando um espaço semântico pré-definido (i.e. *word embeddings*). Esta similaridade é calculada com base em tópicos ocultos presentes nos textos, tornando o método adequado para o cálculo de similaridade mesmo em situações onde o tamanho dos documentos é significativamente diferente. Basicamente, o algoritmo do *Hidden Topics* recebe como entrada dois documentos, onde um deles deve ser um documento contendo um conjunto estruturado de termos, onde estes termos formam diferentes tópicos (e.g. documentos de notícias). Os autores chamam este documento de "Documento Longo". O algoritmo busca encontrar um conjunto de tópicos representados por uma matriz $H = [h_1, ..., h_k]$ onde $h_k \in \mathbb{R}^d, d = 300$

em que seja possível reconstruir o documento com o menor erro possível. O erro é definido como:

$$E = \sum_{i=1}^{n} \|w_i - w_i'\|_2^2 \tag{2.2}$$

onde w_i representa o vetor associado a um determinado termo presente no documento e w_i' representa o vetor linearmente reconstruído para o termo w_i a partir dos vetores de tópicos ocultos presentes no documento. Desta forma, o algoritmo busca construir uma matriz de tópicos H' que minimize o erro E, ou seja:

$$H' = argmin \sum_{i=1}^{n} min \, \|w_i - w_i'\|_2^2 \tag{2.3}$$

O segundo documento que o algoritmo recebe como entrada, onde os autores chamam nos experimentos de "Documento Resumido" que representa um documento contendo termos com os quais será calculada a similaridade em relação ao primeiro documento passado como entrada, e que teve os tópicos ocultos extraídos. A ideia é que, com os tópicos extraídos a partir do primeiro documento, seja possível reconstruir também o segundo documento. Desta forma, baseando-se em similaridade de cossenos, o algoritmo retorna um valor entre 0 e 1, onde quanto maior o valor, maior a similaridade entre os tópicos extraídos em relação ao segundo documento passado como entrada. Adicionalmente, *stopwords* e preposições são removidas na etapa de pré-processamento do algoritmo.

A Figura 2.3 exibe o fluxo de execução do algoritmo apresentado pelos autores. Na imagem, é possível identificar os três principais módulos presentes na abordagem. Sendo o primeiro, o módulo de pré-processamento dos dados de entrada, onde como já mencionado, removem *stopwords* e preposições. Esta modificação visa manter apenas termos potencialmente relevantes para a etapa seguinte. No segundo módulo, os tópicos ocultos são extraídos dos documentos longos já pré-processados. O terceiro módulo presente na Figura 2.3 realiza o mapeamento entre tópicos gerados e os documentos resumidos. Nesta etapa, o algoritmo busca reconstruir o documento resumido a partir dos tópicos ocultos extraídos dos documentos longos. Quanto melhor esta reconstrução, maior será o valor de similaridade reportada pelo algoritmo.

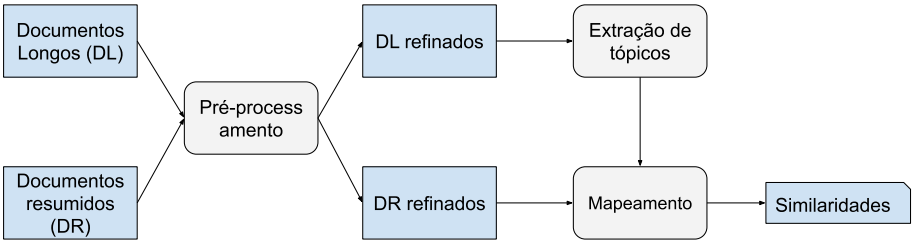

Figura 2.3: Passos do algoritmo para o cálculo de similaridade entre dois documentos por meio da identificação de tópicos ocultos apresentado por Gong et al. (2018). Nesta pesquisa, o algoritmo é referenciado apenas como *Hidden Topics*.

2.2.5 Valores SHAP

O desenvolvimento de diferentes algoritmos de classificação vêm contribuindo para o desenvolvimento de um vasto arsenal de ferramentas que possibilitam executar tarefas de PLN com grande efetividade. Porém, é comum que modelos complexos tomem decisões que muitas vezes não conseguem ser interpretáveis por humanos. Isso trás diversas implicações, pois em muitos cenários, a interpretabilidade de um modelo pode ser tão importante quanto as próprias predições executadas. Sem poder extrair explicações que sejam compreensíveis, problemas como confiabilidade dos modelos podem emergir, dada a falta de transparência acerca de suas decisões.

No contexto de notícias falsas, a construção de modelos preditivos que possam gerar explicações se faz ainda mais importante, pois estas explicações podem guiar profissionais em atividades de checagem de fatos. Usadas nesse contexto, as explicações geradas por um modelo preditivo podem contribuir, de forma mais ampla, para o entendimento de nuances e características ocultas em documentos de notícias falsas.

Com o objetivo de extrair explicações acerca dos modelos construídos nesta pesquisa, utilizamos o método apresentado por Lundberg e Lee (2017) que descreve os chamados *Shapley values*, ou valores SHAP (*SHapley Additive exPlanations*). O método apresentado pelos autores consiste em avaliar a influência que cada *feature* de classificação (ou grupos delas) exerce sobre as tomadas de decisão de um modelo. Dessa forma, os valores SHAP representam o grau de importância as *features* exercem sobre um modelo. A intuição por trás das explicações é que, ao se "perturbar" o valor de uma *feature* relevante para o modelo, esta perturbação irá, inevitavelmente, degradar a performance do modelo. Por outro lado, caso essas perturbações sejam aplicadas em uma *feature* pouco relevante, a degradação no

desempenho do modelo será menor. Esta análise permite a obtenção de um entendimento mais objetivo das decisões de classificação dos modelos implementados, gerando *insights* sobre o problema abordado, neste caso, o de classificação de notícias falsas. Com base nos Valores SHAP, é possível gerar visualizações que ajudam no entendimento dos modelos de classificação gerados nesta pesquisa.

2.3 Considerações Finais

Neste capítulo, foram apresentados referenciais teóricos que fornecem bases para o entendimento desta Tese. Forma considerados aspectos relacionados às definições e características presentes nos documentos de notícias falsas, como os diversos tipos de notícias definidas na literatura e como este tipo de conteúdo pode repercutir no dia a dia. Também foram apresentados os principais conceitos que estão presentes nesta Tese, auxiliando assim, o processo de entendimento técnico e teórico do que será apresentado neste trabalho.

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 PLN, 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;
- 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.
- A ênfase ou *loudness*, que considera termos capitalizados, bem como expressões contendo símbolos de exclamação, asterisco e *underline*.
- Cálculo da legibilidade ou *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 de reação 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, uma palavra 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 Árvores 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* 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 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.

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

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 o *Linguistic Inquiry and Word Count* (LIWC[2]) (PENNEBAKER et al., 2015). O LIWC consiste em uma base de dados de léxicos paga, construída manualmente por especialistas, que contém diferentes classes relacionadas a aspectos psicolinguísticos. 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.

Diversas pesquisas buscam analisar e classificar notícias falsas explorando nuances de subjetividade nas notícias. Este interesse deriva da hipótese de que notícias reais seriam mais objetivas (i.e. menos subjetivas) do que notícias falsas.

Aker et al. (2019) apresenta uma análise de notícias falsas no âmbito de subjetividade dos documentos. O trabalho 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*[3]. 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

[2]https://liwc.wpengine.com/ [3]https://github.com/pattern3/pattern

automática, *scores* de subjetividade, destacando a necessidade de estudos que abordem, de forma eficiente, este problema.

No estudo realizado por Reis et al. (2019b), 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. Adicionalmente, 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*. 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.

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,

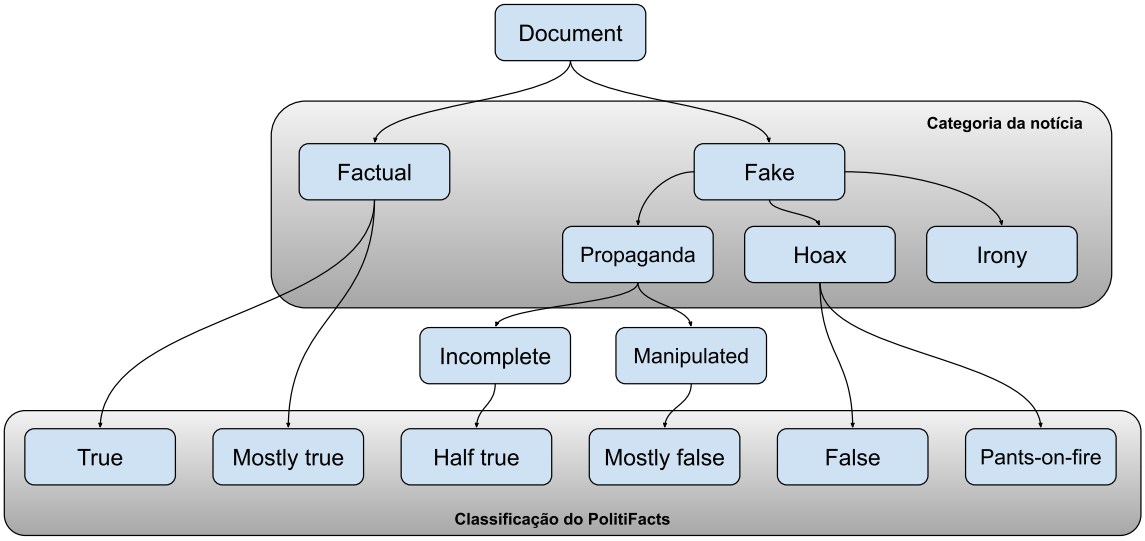

Figura 3.1: Taxonomia proposta por Wang et al. (2018), baseada nas categorias presentes no *"Truth-O-Meter"* do PolitiFacts e na taxonomia denominada SHPT (RASHKIN et al., 2017).

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, palavras, 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.

3.1.1 Classificação de Notícias Falsas Baseada em Léxicos

Castelo et al. (2019) apresentam uma abordagem para classificação de notícias falsas que é agnóstica a tópicos. Esta característica é importante pois grande parte dos trabalhos presentes na literatura utilizam como *features* vetores extraídos a partir de *Bag of Words*. Este tipo de representação textual permite alcançar ótimos valores de acurácia, porém, especialmente em problemas de classificação com limitações na quantidade de amostras rotuladas (e.g.

classificação de notícias falsas), representações baseadas em BoW tendem a sofrer problemas de *overfitting* e redução na capacidade de generalização (CHEN et al., 2013a). Visando suprir este problema, os autores desenvolvem uma estratégia de classificação baseada em *features* presentes em páginas web como também características linguísticas presentes em notícias falsas. Como parte das *features* empregadas nos modelos construídos, os autores usam o LIWC e outras *features* que permitem extrair características específicas das notícias falsas.

Mertoğlu e Genç (2020) apresentam uma proposta para construção de léxicos para classificação de notícias falsas para a língua Turca. Esta pesquisa destaca o fato de não existirem trabalhos que proponham a construção de léxicos voltados, especificamente, para notícias falsas. Optando por construir uma metodologia de construção de léxicos voltados para a língua Turca, os autores utilizam um *dataset* de notícias e exploram a característica da língua turca de ser um idioma aglutinativo, onde sufixos são concatenados no final dos termos para mudar seu sentido. Com isso, os autores constroem modelos de classificação usando como léxicos os termos extraídos das notícias considerando as seguintes classes: "*raw word*", os termos originais acrescido de sua classificação POS (*Part of Speech*) "raw+POS", termo considerando sua raiz sintática "root/stem of word" e os sufixos das palavras "suffix". Os autores constroem os léxicos dessas quatro categorias baseados em um score que considera a frequência de ocorrência dos termos tanto dentro do *dataset* de notícias falsas quando no de notícias reais. Apenas usando os léxicos construídos como *features*, os autores conseguem obter valores de F1-Score de mais de 80% nos resultados reportados. Até o melhor do nosso conhecimento, esta é a única pesquisa de que fato propõe uma abordagem para construção de léxicos para classificação de notícias falsas.

Pérez-Rosas et al. (2018) consideram características textuais como n-gramas, pontuação, termos denotando características psicolinguísticas oriundas do LIWC. Os autores também usam *features* de 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.

Rashkin et al. (2017) vão 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*, *sentimental*, *hedging* e *intensity*. Entre os principais achados da pesquisa, estão que pronomes pessoais, bem como superlativos e advérbios 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%.

Carvalho et al. (2020) propõem um método manual para a construção de uma versão em Português do léxico baseado na Teoria de Fundamentos Morais (i.e. *Moral Foundations Theory* (GRAHAM et al., 2013) para a análise e classificação de notícias falsas para a língua Portuguesa. Os autores descrevem os dez passos utilizados para a construção do léxico, que consideram passos de avaliação manual por especialistas. Como melhores resultados, os autores reportam 85% de F1-score para a classificação de notícias falsas.

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.

3.1.2 Classificação de Notícias Falsas Utilizando Aprendizagem Profunda

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*[4], 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.

Trabalhos que permeiam o estado da arte na identificação de notícias falsas seguem explorando novas arquiteturas de aprendizagem profunda. Nessa direção, trabalhos vêm explorando as chamadas *Generative Adversarial Networks* (GANs), que consiste em uma arquitetura de aprendizado formada por um modelo gerador e um discriminador.

Zellers et al. (2019) desenvolveram o modelo GROVER, que consiste em uma GAN que permite, a partir de uma sentença ou fragmento de texto, a construção de uma notícia falsa completa. Os autores evidenciam que os modelos já existentes falham em classificar as notí-

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

cias falsas geradas pelo GROVER como falsas. Nos experimentos, apenas o descriminador do GROVER era capaz de classificar as notícias falsas geradas com mais de 90% de acurácia. A pesquisa demonstra, fundamentalmente, a necessidade da construção de modelos de classificação que sejam menos susceptíveis a "ataques" por meio de notícias falsas geradas automaticamente.

Le, Wang e Lee (2020) desenvolvem o modelo MALCOM, que consiste em um gerador de textos falsos, que utilizando a arquitetura das GANs, também consegue "enganar" modelos estado da arte na classificação de conteúdo falso. No trabalho, os autores também destacam a necessidade da construção de modelos de classificação de notícias falsas que sejam mais resilientes a conteúdos gerados por meio de GANs.

Outra arquitetura de aprendizado de máquina que vem ganhando destaque em problemas de processamento de linguagem natural são os *Transformers*. Vaswani et al. (2017) Descrevem com detalhes a arquitetura dos *Transformers*, que se baseiam unicamente no mecanismo de atenção, que consiste em adicionar pesos a determinadas *features*, aumentando assim, a sua relevância no processo de treinamento da rede neural. Em termos gerais, a arquitetura é composta com um *Encoder* e um *Decoder*, onde o *Encoder* mapeia a sequência de entrada, por exemplo, um texto, para uma representação vetorial de n dimensões. Esta representação vetorial alimenta o *Decoder*, que gera como saída outra sequência, que pode ser, por exemplo, em um outro idioma, para modelos treinados para realizar traduções textuais.

Chen et al. (2021) apresentam um modelo de linguagem baseado em *Transformer* onde os autores realizam um treinamento, ou *tunning* no modelo para que este possa classificar notícias falsas. Os autores também utilizam para o treinamento notícias criadas por GANs com o objetivo de tornar o modelo mais robusto. Como principal resultado, os autores conseguem obter mais de 99% de F1-score em seu melhor cenário de avaliação, considerando um *dataset* de notícias falsas sobre COVID19.

O trabalho apresentado por Li et al. (2021) também utiliza um *dataset* relacionado a COVID19 para avaliar um modelo baseado em *Transformers*, onde os autores propõem um *ensemble* de modelos que considera diferentes modelos de linguagem pré-treinados, como BERT (*Bidirectional Encoder Representations from Transformers*) e Roberta. Nos resultados, os autores chegam a obter resultados acima de 98% de F1-score para a classificação de notícias falsas.

Schwarz, Theóphilo e Rocha (2020) propõem o EMET, um *framework* baseado em aprendizagem profunda com o objetivo de classificar postagens falsas em redes sociais. O método faz uso de *embeddings* extraídos de um *encoder*, bem como também utilizam as reações dos leitores ao ler uma determinada postagem. Os autores chegam a obter um F1-score de 93%.

3.2 Construção de Léxicos

Diversos trabalhos buscam construir léxicos que podem ser utilizados em diversas aplicações, como classificação de documentos. Porém, a maioria dos trabalhos buscam a construção de léxicos voltados para o problema de análise de sentimentos. Choi e Wiebe (2014) apresentam um método para a construção de léxicos para a análise de opiniões considerando polaridades as polaridades positiva e negativa. Os autores se baseiam na ideia de "*sense lexicons*", que consiste em termos que determinam uma opinião positiva ou negativa sobre uma entidade. Os autores utulizam uma abordagem supervisionada em conjunto com relações presentes no WordNet[5] para construir os léxicos.

Wang e Xia (2017) utilizam uma arquitetura baseada em aprendizagem profunda para a construção de *embeddings* no contexto de análise de sentimentos. Estes *embeddings* permitem a construção de léxicos de melhor qualidade. Já Huang, Niu e Shi (2014) apresentam uma abordagem para a construção automática de léxicos de sentimentos para domínios específicos baseada na propagação de rótulos por meio de um léxico pré-existente.

Fast, Chen e Bernstein (2016) demonstram uma ferramenta denominada de *Empath*, que permite, a partir de um pequeno conjunto de termos iniciais, ou *seeds*, construir e validar diferentes categorias de léxicos. A abordagem utiliza modelos de aprendizagem profunda e *embeddings* para gerar léxicos que possuam um contexto similar aos termos utilizados como entrada. Deng et al. (2019) também utilizam uma abordagem baseada em aprendizagem profunda para a construção de léxicos. Os autores focam na construção de léxicos de sentimentos baseados em LSTM e no mecanismo de atenção, permitindo construir léxicos que sejam relevantes para a classificação de sentimentos.

[5]WordNet 3.0 - http://wordnet.princeton.edu/

3.3 Posicionamento desta pesquisa em relação aos trabalhos relacionados

A Tabela 3.4 exibe as principais características dos trabalhos relacionados descritos neste capítulo. Também é apresentado, na última linha, a caracterização desta pesquisa. 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*);
- **Léxicos:** Uso de léxicos para a caracterização de notícias falsas;
- **Criação de léxicos específicos:** Proposição de métodos para a construção de léxicos para a caracterização 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 atende quatro dos seis aspectos apresentados na Tabela 3.4. Neste trabalho, o foco consiste na proposta de um método para a construção automática de léxicos voltados para a análise e classificação de notícias falsas. O método proposto contempla a própria construção dos léxicos, como também o desenvolvimento de *features* baseadas nos léxicos gerados. Estas *features* serão usadas para a classificação das notícias. Na tabela, é possível verificar que apenas duas pesquisas propõem, de fato, métodos para a construção de léxicos voltados para este tipo de conteúdo. No melhor do nosso conhecimento, esta pesquisa é a primeira a propor a construção de léxicos voltados para a classificação deste tipo de notícias para a língua inglesa. Uma possível causa para esta lacuna parece estar relacionada à grande disponibilidade de léxicos já existentes, porém, construídos para outros domínios. Mesmo apresentando bons resultados, os léxicos existentes na literatura não foram, de fato, construídos especificamente para a caracterização de notícias falsas. Adicionalmente, existe uma

grande dificuldade para a manutenção de léxicos construídos manualmente, especialmente no cenário de notícias falsas, onde diferentes domínios e assuntos surgem constantemente.

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 construídos. De modo geral, esta pesquisa busca suprir as duas principais lacunas apresentadas na Tabela 3.4, que consiste em considerar métodos de construção de léxicos voltados para notícias falsas, bem como a explicabilidade dos modelos construídos.

3.4 Considerações Finais

Nesta seção, foram apresentados alguns dos principais trabalhos relacionados ao tema desta tese. O objetivo do levantamento bibliográfico foi de traçar pesquisas relevantes que abarcassem aspectos importantes no âmbito de classificação de notícias falsas. Neste sentido, esta seção foi subdividida em trabalhos de classificação de notícias falsas mais generalizados, utilizando diversas características textuais para a realização da classificação; trabalhos que utilizaram léxicos como *features* de classificação; e pesquisas que implementaram técnicas de Aprendizagem Profunda para a detecção de notícias falsas. Por fim, foi apresentado um posicionamento desta pesquisa em relação aos trabalhos catalogados.

	Feat. Textuais	Redes Sociais	Léxicos	Criação de léxicos	Explicabilidade	Deep Learning
Ahmed, Traore e Saad (2017b)	x					
Horne e Adali (2017)	x		x			
Pérez-Rosas et al. (2018)	x		x			
Monteiro et al. (2018)	x		x			
Rashkin et al. (2017)	x		x			x
Wang (2017)	x					x
Ruchansky, Seo e Liu (2017)	x	x				x
Aker et al. (2019)	x					
Reis et al. (2019b)	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			
Castelo et al. (2019)	x		x			
Mertoğlu e Genç (2020)	x		x	x		
Carvalho et al. (2020)	x		x	x		
Zellers et al. (2019)	x					x
Le, Wang e Lee (2020)	x					x
Chen et al. (2021)	x					x
Li et al. (2021)	x					x
Schwarz, Theóphilo e Rocha (2020)	x					x
Proposta	**x**		**x**	**x**	**x**	

Capítulo 4

Construção Automática de Léxicos Baseados em Documentos de Notícias

Neste capítulo, é apresentada uma proposta de solução para o problema de construção automática de léxicos para classificação e análise de notícias falsas. Em seguida, são apresentados detalhes da avaliação experimental, onde são descritas as metodologias de avaliação bem como uma descrição das bases de dados utilizadas e resultados obtidos. Em seguida, é apresentado um estudo acerca da capacidade de explicação dos modelos de classificação gerados.

4.1 Construção Automática de Léxicos para Identificação de Notícias Falsas

O uso de léxicos nas atividades de processamento de linguagem natural tem como principal vantagem incorporar um conhecimento prévio em atividades que envolvem o processamento de texto. Por exemplo, um léxico de sentimentos (composto de termos que apresentam polaridades positiva e negativa) criado por especialistas, pode ser usado, com bastante eficácia, em atividades de classificação de sentimentos em texto.

Porém, apesar de permitir embarcar conhecimento especializado, os léxicos tendem a requerer um alto custo para sua criação, demandando grande esforço manual para sua construção. Em léxicos de sentimentos, por exemplo, há uma relativa facilidade na identificação

manual, por exemplo, de termos que pertencem a polaridades positiva e negativa. Porém, para cenários mais específicos, como o de notícias falsas, a disponibilidade de especialistas para gerar tais léxicos se torna um fator limitante. Neste contexto, esta pesquisa tem como objetivo apresentar um método para a construção automático de léxicos baseados em documentos de notícias falsas, os quais podem auxiliar, tanto na classificação, como na exposição de nuances presentes neste tipo de documento textual. O método apresentado tem como característica a extração de léxicos baseadas em documentos de notícias falsas recebidos como entrada. Em resumo, o método proposto extrai termos que melhor classificam os documentos passados como entrada, gerando assim, um conjunto de léxicos extraídos a partir desses documentos. Estes léxicos são utilizados para construir *features* baseados em similaridade semântica, que serão utilizadas para treinar modelos preditivos.

4.1.1 Descrição do Método de Construção de Léxicos

O método de construção de léxicos proposto necessita, inicialmente, de dois conjuntos de dados (documentos textuais) que possam ser classificados de forma binária. No contexto desta pesquisa, foram utilizados documentos de notícias falsas e reais, onde considera-se que uma notícia não pode ser falsa e real simultaneamente.

A solução proposta consiste em uma abordagem gulosa para encontrar um léxico $L = \{w_1, w_2, w_3, ..., w_n\}$ que maximize uma métrica de classificação (MC) utilizada para classificar dois conjuntos distintos de documentos D_1 e D_2. Cada termo do léxico L deve derivar de um vocabulário inicial (V) criado a partir de termos mais relevantes presentes nos documentos. A partir destes termos presentes no vocabulário inicial, serão selecionados, por meio de classificações sucessivas, um subconjunto de (V) que melhor classifica os documentos D_1 e D_2 passados como entrada. A ideia central é extrair um conjunto de termos que, a princípio, sejam informativos para a classificação destes documentos. Esses termos podem carregar características como estilo de escrita, informalidade textual, características semânticas ocultas dentre outras. Neste processo, são descartados *stopwords* e nomes próprios, como nomes de pessoas e lugares. Isso permite atenuar um eventual enviesamento dos léxicos que são construídos.

Nas próximas subseções, as etapas de construção dos léxicos serão descritas em mais detalhes.

Construção do Vocabulário Inicial

Como descrito anteriormente, o léxico final deriva de um vocabulário inicial (V), que é gerado a partir dos termos presentes nos documentos D_1 e D_2. Este vocabulário é construído com base nos termos com maior peso TFIDF associado aos documentos. Um *threshold* mínimo é definido manualmente, onde apenas termos que tenham um peso TFIDF acima deste *threshold* são selecionados para constituírem (V). Dessa forma, temos que $\{c \in V | peso\_tfidf(w) >= threshould\}$, onde w representa um termo candidato a compor o vocabulário inicial, e a função $peso\_tfidf$ retorna o peso TFIDF deste termo.

Função para Seleção de Termos Candidatos

A função de construção dos léxicos é essencialmente uma função gulosa. Algoritmos gulosos necessitam de uma função de seleção para avaliar os candidatos a solução local, que neste caso, são palavras, que deverão compor a solução final, ou seja, o léxico final (L). Para esta função, a abordagem proposta utiliza um classificador, que é treinado com cada candidato presente em (V), utilizando uma representação vetorial simples dos termos, considerando a quantidade de ocorrências desses termos no *dataset*. Inicialmente, um termo aleatório é selecionado em (V) e usado como *feature* para classificação dos documentos D_1 e D_2. Sempre que um termo resulta em uma melhora na classificação, em termos da métrica de classificação (MC), este novo termo é adicionado a (L) e um novo termo é selecionado para avaliação. Desta forma, à medida que novos termos vão sendo adicionados em (L), o desempenho das classificações tende a melhorar, até atingir um platô de desempenho de classificação.

O Algoritmo 1 exibe um pseudo-código que descreve e formaliza o processo de construção de léxicos proposto. O Algoritmo recebe como entrada um *dataset* D de notícias falsas e reais e um *threshold* manualmente definido. Na linha 1, a função "filtra_tfidf" recebe como entrada o *threshold* definido, o *dataset* de notícias falsas e reais D. Esta função irá calcular os pesos TFIDF para os termos presentes no *dataset* e irá retornar apenas os termos com pesos maiores que o *threshold* definido. Esta função irá calcular os pesos TFIDF para as notícias presentes em D e apenas os termos presentes nas notícias falsas (definido pelo marcador "falsas") que sejam maiores ou iguais ao *threshold* será retornado. Na linha 2, o

conjunto de termos que irá compor o léxico construído ao final do processo é instanciado como um conjunto vazio. Na linha 3, uma métrica de classificação MC é instanciada. A partir da linha 4, o algoritmo itera na lista de candidatos definidos nas linhas anteriores, de forma a permitir que as classificações sucessivas selecionem os melhores termos para compor o léxico final. A linha 5 adiciona um termo candidato ao conjunto de possíveis léxicos finais. A linha 6 vetoriza o *dataset* D considerando a contagem de ocorrências dos termos presentes na lista "lista_final" (i.e. vocabulário). Esta representação vetorial baseada em BoW será utilizada para avaliação por meio de validação cruzada na linha 7, por meio da função "avalia" que recebe a representação vetorizada dos documentos de notícias. Na linha 8, o resultado da classificação é comparado com o melhor resultado encontrado para a métrica de classificação MC. Caso o resultado seja superior ao resultado já presente em MC, o novo termo candidato será adicionado definitivamente ao conjunto "lexico_final" e irá compor o resultado final, ou seja, o léxico final construído.

Algorithm 1 Pseudo-código descrevendo o processo de criação de léxicos a partir de classificações de notícias falsas e reais.

```
1: candidatos ← filtra_tfidf(threshold, D, "falsas")
2: lexico_final ←
3: MC ← 0
4: for termo in candidatos do
5:     lexico_final.add(termo)
6:     D_vetorizado ← CountVectorizer(lexico_final, D)
7:     resultado ← avalia(D_vetorizado)
8:     if resultado ≤ MC then
9:         lexico_final.remove(termo)
10:    end if
11: end for
12: resultado ← lexico_final
```

4.2 Construção de *Features* de Classificação Baseadas em Similaridade Semântica

Os léxicos construídos na etapa anterior representam um conjunto de termos que permitem, a princípio, classificar os documentos de notícias passados como para o algoritmo. É comum na literatura, construir representações de documentos baseadas em representações vetoriais

baseadas em pesos TFIDF ou mesmo simples frequência de ocorrência de termos presentes em um vocabulário. Porém, tais representações são limitadas, deixando de representar aspectos e nuances semânticas mais profundas presentes nos documentos. Esta pesquisa propõe o uso de similaridade semântica como estratégia para a construção de *features* de classificação baseadas no léxicos obtidos através do algoritmo descrito.

Para a construção das *features* de classificação, esta pesquisa propõe o uso de similaridade semântica para representar documentos textuais. Este tipo de representação considera que cada termo presente em um documento é representado por um vetor dentro de um espaço semântico (i.e. *word embeddings*). A principal vantagem desta abordagem é que, ao invés de considerar a ocorrência pontal de um termo no documento, como ocorre com abordagens baseadas em BoW, aqui cada palavra é representada por seu componente semântico presente nos *embeddings* utilizados. Para extrair tais *features*, são calculadas similaridades semânticas entre os léxicos gerados e um documento de notícias. O resultado é um vetor $v = (v_1, v_2, v_3, ..., v_d)$ onde cada valor representa a similaridade semântica entre um dado léxico e o documento de notícias, considerando d a quantidade de léxicos distintos usados para representar o documento. A estratégia apresentada nesta pesquisa representa uma evolução em relação à estratégia apresentada por Jeronimo et al. (2019, 2020) pois, nos referidos trabalhos, foram utilizados como método para o cálculo das similaridades, o algoritmo WMD (KUSNER et al., 2015). Esta estratégia possui imprecisões no que tange o cálculo das similaridades entre documentos textuais de tamanhos significativamente diferentes. Nesta presente pesquisa, foi adotado o método *Hidden Topics* (GONG et al., 2018). Este método emprega uma abordagem onde o tamanho dos documentos não exerce influência nos resultados de similaridade, sendo então, mais apropriada para o cálculo de similaridade entre léxicos de tamanho arbitrário e documentos de notícias.

4.3 Base de Dados de Notícias

Os *datasets* utilizados para as avaliações são compostos de notícias falsas e reais na língua inglesa. É importante destacar que nos dados utilizados estão excluídas postagens em aplicativos de mensagens instantâneas, bem como fragmentos textuais presentes em imagens.

Foram considerados três *datasets* para as avaliações executadas nesta pesquisa. O pri-

meiro, denominado de aqui de BSDetector[1] (SHU et al., 2020) consiste em um compilado de notícias falsas publicadas em 2016 e que compuseram um *plugin* para navegador, emitindo alertas sempre que o usuário navegava por páginas classificadas como suspeitas de propagarem notícias falsas. A lista de notícias que compõem este *dataset* foi manualmente construída por especialistas em conteúdo duvidoso, e dividida em diferentes categorias de notícias, de acordo com seu conteúdo. Das categorias presentes no *dataset*, foram utilizadas as seguintes:

- Bias: Fontes que disseminam propaganda política e distorções grosseiras de fatos (356 documentos);
- Conspiracy: Fontes que são conhecidas disseminadoras de teorias da conspiração (328 documentos);
- Hate: Fontes que promovem ativamente o racismo, misoginia, homofobia e outras formas de discriminação (239 documentos);
- Junk-Science: Fontes que promovem a pseudociência e outras afirmações cientificamente duvidosas (102 documentos);
- Satire: Fontes que promovem notícias satíricas de humor em forma de notícia (99 documentos);
- State: Notícias falsas relacionadas a países repressores/ditaduras (118 documentos).

Como o *dataset* do BSDetector não possui notícias reais, foram coletadas notícias reais do popular conjunto de dados disponível publicamente no Kaggle, denominado "All the News"[2]. As notícias utilizadas foram publicadas durante os anos de 2016 e 2017. Dessa 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 a partir das notícias presentes nas capas dos referidos jornais, sendo estas, as principais notícias dos veículos. Para simplificação, os dados do BSDetector e "All the News" serão referenciados nesta pesquisa apenas como BSDetector.

[1] <https://github.com/selfagency/bs-detector> [2] https://www.kaggle.com/snapcrack/ all-the-news/version/4

O segundo *dataset* de notícias utilizado é *Celebrity* (PÉREZ-ROSAS et al., 2018), que contém notícias falsas e reais sobre celebridades, colhidas a partir de portais de entretenimento. O *dataset* contém 250 notícias falsas e 249 notícias reais, sendo portanto, balanceado. O terceiro *dataset* de notícias falsas utilizado nesta pesquisa é o apresentado por Koirala (2021), e consiste em notícias falsas e reais sobre COVID-19. A base de dados é composta por 1034 notícias falsas e 2013 notícias reais.

4.4 Léxicos Construídos Manualmente - *Baselines*

Para avaliar a capacidade de classificação dos léxicos gerados na pesquisa, definidos em seções anteriores como LG, são utilizados três diferentes conjuntos de léxicos já presentes na literatura e que foram construídos com a intervenção de especialistas. Para a escolha dos léxicos adotados como *baseline*, foram escolhidos léxicos que pudessem, de alguma forma, representar um aspecto que seguramente pudesse diferenciar notícias falsas e reais. O aspecto escolhido foi a subjetividade dos documentos, tendo a premissa que, idealmente, notícias falsas sejam mais subjetivas que notícias reais. Dentro do escopo que abrange nuances de subjetividade, os léxicos selecionados para compor o *baseline* consideram, além da própria subjetividade, aspectos de sentimentos e viés. O primeiro conjunto de léxicos foi compilado por Recasens, Danescu-Niculescu-Mizil e Jurafsky (2013) e denominado aqui que *Bias-inducing terms*[3]. Este conjunto apresenta seis diferentes léxicos que expressam diferentes nuances de viés e subjetividade textual. Estes léxicos são utilizados nos trabalhos apresentados por Volkova et al. (2017), Rashkin et al. (2017) para o estudo e identificação de notícias falsas. Os léxicos são apresentados a seguir, mantendo a nomenclatura original na língua inglesa.

- *Factive Verbs*: pressupõem fatos em uma oração (27 termos);
- *Implicative Verbs*: insere a ideia de implicação em uma oração (32 termos);
- *Assertive verbs*: verbos que afirmam uma proposição (66 termos);
- *Hedges*: usado para reduzir o compromisso com a verdade de uma proposição, evitando declarações assertivas (100 termos);

[3] http://zissou.infosci.cornell.edu/data/npov/bias_related_lexicons.zip

- *Reporting Verbs*: usado para reportar ações de pessoas ou de atividades (181 termos);
- *Bias-inducing lemmas*: denota uma posição previamente estabelecida, ou enviesada (654 termos).

A Tabela 4.1 apresenta exemplos dos léxicos presentes no conjunto *Bias-inducing terms*.

Léxicos	Termos
Factive Verbs	learn, note, notice, observe, perceive, recall, remember, reveal, see, resent, amuse, suffice, bother, make, sense, care
Implicative Verbs	manage, remember, bother, get, dare, care, venture, condescend, happen, fit, careful, misfortune, sense, succeed, deign
Assertive verbs	think, believe, suppose, expect, imagine, guess, seem, appear, figure, acknowledge, admit, affirm, allege, answer, argue
Hedges	about, almost, apparent, apparently, appear, appeared, appears, approximately, around, assume, assumed, essentially, estimate, estimated
Reporting Verbs	accuse, acknowledge, add, admit, advise, agree, blame, boast, caution, charge, demonstrate, deny, describe, determine, disagree, disclose
Bias-inducing lemmas	abortion, abuse, abusive, accept, break, bring, collapse, colony, come, crime, criminal, critic, deal, death, essential, excellent

Tabela 4.1: Exemplos de termos presentes no conjunto de léxicos denominado nesta pesquisa de *Bias-inducing terms*

O segundo conjunto de léxicos é apresentado por Wilson, Wiebe e Hoffmann (2005). Este conjunto de léxicos é parte do projeto *Multi-Perpective Question Answering Subjectivity Lexicons* (MPQA)[4] e é dividido em polaridades de sentimentos (positiva e negativa), classificados como sendo de forte ou fraca subjetividade. Para estes léxicos, 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 1.482 para a polaridade positiva. Por possuírem a especificação de termos que apresentam altos níveis de subjetividade, estes léxicos são frequentemente utilizados em abordagens para a classificação de notícias falsas (WANG et al., 2019; Wang et al., 2018). A Tabela 4.2 exibe exemplos de termos presentes neste conjunto de léxicos.

A relação entre detecção de sentimentos e notícias falsas já é demonstrada na literatura, onde este tipo de característica textual contribui para a classificação dessas notícias (FAUSTINI; COVOES, 2020; BHUTANI et al., 2019). O terceiro conjunto de léxicos utilizados é o apresentado por Choi e Wiebe (2014) e também representa polaridades de sentimentos (positiva e negativa). No trabalho, alguns termos (*seeds*) foram manualmente anotados por

[4]https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

Léxicos	Termos
Positive	generosity, joke, repent, celebration, alluring, moving, commend, upliftment, endurable, trendy, altruist, justify, novel, spellbound, insistently
Negative	deprave, indecisively, incompetence, cheap, decadence, presumptuous, ambivalence, catastrophe, terror, neglect, condemnable, flabbergast, unimaginable, enormous, dubiously

Tabela 4.2: Exemplos de termos presentes no conjunto de léxicos oriundos do projeto *Multi-Perpective Question Answering* (MPQA)

especialistas a partir de textos subjetivos, onde os autores executam uma expansão automática dos termos utilizando a estrutura presente no WordNet[5]. Este conjunto de léxicos contém 1.003 termos para polaridade negativa e 493 para a positiva. A Tabela 4.3 apresenta alguns exemplos do conjunto de léxicos utilizado na pesquisa dos referidos autores. Nesta pesquisa, estes conjunto de léxicos será referenciado como "Wiebe". Estes léxicos também podem ser obtidos através do projeto MPQA[6].

Léxicos	Termos
Positive	ascertain, overcome, believe, better, free, beautify, seek, supercharge, answer, recognize, pander, adhere, luxuriate, revolutionise
Negative	obstruct, skin, lose, deplore, bedevil, congest, confuse, embarrass, scourge, charge, offend, suspect

Tabela 4.3: Exemplos de termos presentes no conjunto de léxicos oriundos do trabalho de Choi e Wiebe (2014).

4.5 Configurações da Etapa de Construção dos Léxicos

Nesta seção, serão descritos todos os modelos e métodos necessários para a execução e replicação dos experimentos executados nesta pesquisa. Inicialmente, como descrito na seção 4.1.1 deste capítulo, o algoritmo de construção de léxicos se baseia nas classificações de um modelo em que, sempre que uma métrica de avaliação MC aumenta com um novo termo candidato, este termo é incluído na lista de léxicos que será a solução final. Neste processo, é utilizado o algoritmo de aprendizagem LightGBM[7] (KE et al., 2017) utilizando uma configuração padrão de hiper-parâmetros. Este algoritmo foi escolhido por se destacar em termos de velocidade de treinamento e eficiência no uso de memória, além de apresentar uma alta acurácia em diversos problemas de classificação. Esses requisitos são necessários para compor

[5]https://wordnet.princeton.edu/ [6]http://mpqa.cs.pitt.edu/lexicons/effect_lexicon/
[7]https://github.com/microsoft/LightGBM

a estratégia de construção de léxicos proposta, a qual realiza classificações sucessivas para conseguir encontrar a melhor combinação de léxicos e reportá-los como saída do algoritmo.

Ainda no que tange a execução do algoritmo que constrói os léxicos baseado nos documentos de notícias, é descrito na seção 4.1.1 a construção de um vocabulário inicial V baseado nos pesos TFIDF dos termos dos documentos. Este vocabulário inicial tem como objetivo filtrar termos com um peso TFIDF mínimo, descartando assim, termos que aparecem com bastante frequência dos documentos, como também demais termos irrelevantes para a caracterização das notícias. Este peso TFIDF mínimo foi empiricamente ajustado para 0,05, onde apenas termos com peso acima deste valor são considerados para a construção do vocabulário inicial.

Como descrito em seções anteriores, após a construção dos léxicos pelo algoritmo já descrito, os mesmos são usados como base para o cálculo das *features* baseadas em similaridade semântica através do algoritmo *Hidden Topics*. Este algoritmo recebe como entrada um conjunto de léxicos e outro conjunto de documentos de texto (i.e. notícias falsas e reais), com o qual serão extraídas as similaridades semânticas entre eles. Estes valores serão utilizados, de fato, como *features* de classificação de modelos preditivos. O *Hidden Topics* utiliza diversos parâmetros, dentre os principais, está uma camada de *word embeddings* que nesta pesquisa foi utilizado o modelo pré-treinado denominado Google News embeddings[8]. Este modelo pré-treinado foi adotado por ser construído a partir de um grande volume de documentos de notícias publicadas nas plataformas do Google, contendo vetores considerando um vocabulário de aproximadamente 3 milhões de palavras. Outro atributo interessante relacionado ao *Hidden Topics* é a quantidade de tópicos utilizados, que consiste na quantidade de tópicos ocultos que são extraídos para calcular a similaridade semântica entre os textos. Para os experimentos realizados, este valor está definido como 15 tópicos, valor este que se adéqua à maioria dos casos (GONG et al., 2018).

A Figura 4.1 exibe um diagrama descrevendo os passos básicos para a construção de um léxico a partir de documentos de notícias. Inicialmente, o Algoritmo 1 de construção de léxicos descrito recebe como entrada um conjunto de notícias falsas de um determinado tipo. Neste caso, notícias falsas do tipo "Viés". O algoritmo também recebe um conjunto de notícias reais, de forma que o conjunto de notícias falsas e reais esteja balanceado. Como já

[8]https://code.google.com/archive/p/ word2vec/

descrito, o algoritmo irá realizar classificações utilizando os termos mais relevantes presentes no conjunto de dados de notícias falsas para compor um léxico que permita as melhores classificações possíveis. Como estratégia para abreviar a finalização do algoritmo de construção de léxicos, foi utilizada uma abordagem de *early stopping* baseado no tempo em que a métrica de avaliação é modificada. Basicamente, caso a métrica não fosse atualizada depois de 48 horas, o algoritmo finaliza sua execução reportando os léxicos obtidos. Ao fim do processo, um conjunto de termos representando o léxico extraído a partir dos documentos de notícias falsas de Viés será retornado.

A Figura 4.2 apresenta o processo de extração de *features* a partir dos léxicos gerados, e que serão utilizadas para, de fato, treinar modelos preditivos visando a classificação de notícias falsas. Na Figura, o módulo de extração de *features* recebe como entrada um conjunto de léxicos já existentes, como também o conjunto de notícias falsas e reais que será utilizado para avaliação. Nesta etapa, são calculadas as similaridades semânticas entre cada notícia e cada um dos léxicos gerados. Desta forma, caso sejam utilizados, por exemplo, seis léxicos, cada notícia será representada por um vetor de dimensão igual a 6. Desta forma, como saída, é retornada uma matriz M de dimensões $n \times d$ para o conjunto de notícias falsas e outra para o conjunto de notícias reais, onde n é a quantidade de amostras e d é a quantidade de léxicos utilizados. Após o processo de vetorização das notícias, estas são, enfim, utilizadas para treinamento e avaliação de modelos preditivos.

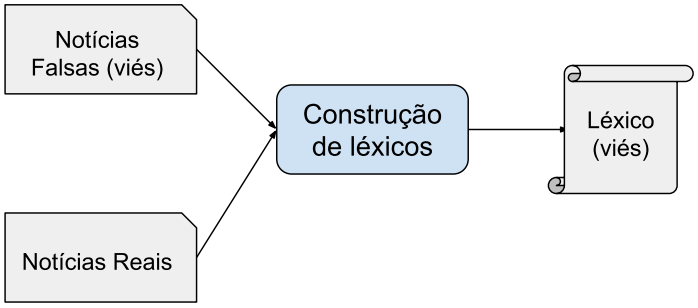

Figura 4.1: Diagrama exibindo os passos básicos para a construção dos léxicos. Basicamente, o algoritmo de construção recebe documentos de notícias falsas e reais, e extrai, a partir dos termos mais relevantes presentes no conjunto de dados de notícias falsas, o termos que irão compor um léxico que melhor classifica os dois *datasets.*

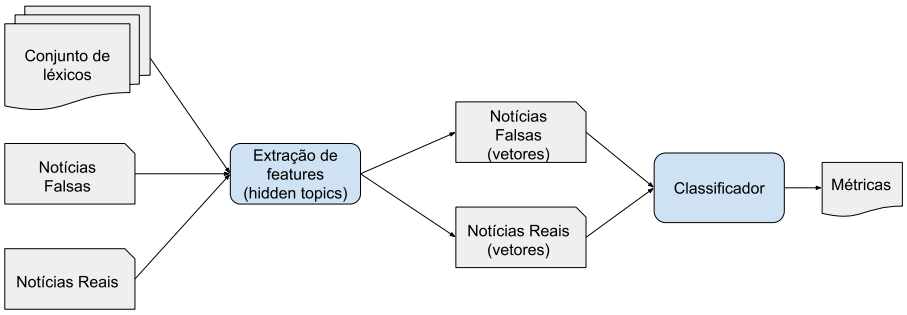

Figura 4.2: Diagrama exibindo os passos básicos para a extração das *features* de classificação baseadas em similaridade semântica a partir dos léxicos recebidos como entrada. Em termos gerais, as notícias falsas e reais são vetorizadas utilizando as similaridades semânticas entre cada notícia e os léxicos passados como entrada. A matriz de vetores para resultante é utilizada para o treinamento e avaliação de modelos preditivos.

4.5.1 Extração de Sentenças Iniciais para Construção dos Léxicos de Notícias Falsas

Durante as execuções iniciais do algoritmo de construção de léxicos apresentado na Seção 4.1.1, foi investigada a possibilidade de uma melhor performance para a classificação de notícias falsas ao considerar as primeiras sentenças das notícias, ao invés do corpo completo dos documentos. Esta hipótese é sustentada por trabalhos que apontam que os elementos iniciais das notícias seriam mais apelativos para o leitor (HORNE; ADALI, 2017; DOGO; DEEPAK; JUREK-LOUGHREY, 2020), o qual muitas vezes, compartilha este tipo de desinformação apenas lendo os títulos ou o começo das notícias, ao invés de ler o conteúdo na íntegra. Dogo, Deepak e Jurek-Loughrey (2020) demonstram que as sentenças iniciais das notícias falsas parecem divergir semanticamente do restante da notícia com mais frequência do que notícias reais. Apesar dos autores não conseguirem demonstrar as razões para tal divergência, hipóteses de que o início das notícias falsas seriam úteis para "fisgar"o leitor podem ser válidas, usando elementos de linguagem mais emotivos e subjetivos. Para avaliar tal hipótese nesta pesquisa, experimentos avaliaram o desempenho de classificação no momento em que os léxicos são criados, portanto, durante a execução do algoritmo que constrói o léxico por meio das classificações entre as notícias falsas e reais.

Durante a etapa de construção dos léxicos, os melhores resultados foram observados quando os léxicos eram construídos, de fato, a partir das três primeiras sentenças das notícias oriundas do *dataset* BSDetector e “All the News”. Este *dataset* foi escolhido para

esta avaliação preliminar por possuir uma grande variedade de tópicos e fontes de notícias distintos. Os resultados deste experimento preliminar foram de 75% de ROC-AUC quando gerados a partir das três primeiras sentenças. Porém, quando gerados com todo o corpo das notícias, os resultados foram de 55% de ROC-AUC. Com base nestes resultados, o uso das três primeiras sentenças foi adotado como padrão para a construção de léxicos de notícias falsas, e integrou o arcabouço de construção de léxicos proposto nesta pesquisa. Esta escolha também permite melhorar o desempenho do algoritmo de construção de léxicos por reduzir o espaço de termos possíveis para a construção dos léxicos.

4.6 Considerações Finais

Esta seção descreveu a abordagem utilizada para a construção automática dos léxicos, bem como o processo envolvido na extração das *features* de classificação. Também foi apresentado os conjuntos de dados de notícias utilizados, bem como os léxicos que foram empregados nos experimentos.

Capítulo 5

Metodologia de Avaliação e Resultados

Este capítulo descreve a metodologia de avaliação e os principais resultados encontrados para a classificação e análise de notícias falsas e reais, utilizando os léxicos construídos na pesquisa. Estes resultados são comparados com modelos treinados a partir de léxicos construídos manualmente e já utilizados na literatura.

5.1 Metodologia Geral de Experimentação

O principal objetivo dos experimentos executados é de verificar quão efetivo os Léxicos Gerados (LG) são, quando comparados com léxicos já presentes na literatura e construídos manualmente por especialistas (*baselines*). As avaliações serão executadas com foco nas análises estatísticas e classificação de notícias falsas, permitindo avaliar se os LG são competitivos quando comparados com os *baselines*.

Para a execução dos experimentos, tanto os LG quanto os léxicos utilizados como *baseline* serão submetidos ao mesmo processo de construção de *features* de similaridade semântica usando o método *Hidden Topics*. Com as *features* geradas, estas são utilizadas para a realização de análises estatísticas e classificações em três diferentes *datasets* de notícias falsas. Ainda no contexto de experimentação dos modelos de classificação construídos, serão analisadas a capacidade de explicação dos modelos, e como eles podem ajudar no entendimento de nuances presentes nos documentos de notícias falsas.

5.2 Construção de Modelos de Classificação para Avaliação

Para avaliar os LG, bem como o poder de classificação das *features* extraídas a partir deles, é utilizado o algoritmo de classificação XGBoost (*Extreme Gradient Boosting*). Este algoritmo é utilizado em trabalhos de classificação de notícias falsas mais recentes (REIS et al., 2019b; ZHOU et al., 2020; HAUMAHU; PERMANA; YADDARABULLAH, 2021; REIS et al., 2019a), apresentando também um bom desempenho de classificação em diversas aplicações (OLSON et al., 2017).

Como descrito na Seção 4.3, o *dataset* de notícias BSDetector possui seis classes de notícias falsas (i.e. *Conspiracy* (conspirações), *Bias* (viés), *Hate* (ódio), *Satire* (sátiras), *Junk-Science* (pseudo-ciência) e *State* (notícias falsas de estados sob regimes repressores)). Dada a grande diversidade de notícias e temas presentes neste *dataset*, bem como sua divisão entre diferentes categorias de notícias falsas, ele foi escolhido como sendo base para a construção dos léxicos que serão usados para a avaliação do método proposto. Basicamente, os léxicos são construídos seguindo a abordagem proposta na Seção 4.1.1, considerando os documentos de notícias falsas de cada uma destas classes do BSDetector. Com isso, serão gerados seis léxicos de notícias falsas, uma para cada uma das seis categorias de notícias falsas presentes no *dataset*. Estes léxicos serão utilizados na etapa de avaliação desta pesquisa. Para manutenção das nomenclaturas das seis classes de notícias falsas presentes no *dataset*, estas serão referenciadas por seus nomes em inglês.

Como já descrito anteriormente, a abordagem proposta nesta pesquisa elege termos para compor o léxico final a partir de sucessivas classificações na etapa de construção dos léxicos. Dada esta característica da abordagem, é necessário um conjunto de notícias reais para a construção dos léxicos. Para tal, foram selecionados 90% dos dados de notícias falsas do BSDetector e 90% das notícias reais presentes no *dataset* "All the News" para a etapa de construção dos léxicos.

Nesta etapa de construção dos léxicos, os dados de notícias falsas e reais são utilizados de forma balanceada para cada uma das seis classes de notícias falsas presentes no BSDetector. Por exemplo, para gerar o léxico relativo à classe de notícias falsas de Viés (i.e. *Bias*) que possui 297 notícias falsas no *dataset* de construção do léxicos (90% do total de notícias de

Bias), são extraídas, de forma aleatória, 297 notícias reais para compor o conjunto de dados que irá ser usado na etapa de construção desse léxico. Os 10% restantes dos dados são usados na etapa de avaliação dos modelos.

Os seis léxicos gerados a partir dos dados do BSDetector e "All the News" serão também avaliados em dois conjuntos de dados que não foram usados na etapa de construção dos léxicos, que são os conjuntos de notícias COVID19 e *Celebrity*.

Para avaliar os modelos construídos, utilizamos a ROC-AUC, que consiste em avaliar o *trade-off* entre a sensibilidade (i.e. taxa de verdadeiros positivos) e a especificidade (i.e. taxa de verdadeiros negativos). Esta métrica tem como vantagem considerar todos os *thresholds* possíveis para o modelo de classificação. Como forma de complementação dos resultados, as métricas *Precision*, *Recall* e *F1 Score* também serão reportadas.

Como descrito no Capítulo anterior, foi observado um melhor desempenho na etapa de construção dos léxicos quando os mesmos foram construídos a partir das sentenças iniciais das notícias. Porém, para realizarmos uma avaliação mais abrangente dos léxicos construídos, todos os modelos preditivos gerados serão avaliados/testados considerando tanto as primeiras sentenças das notícias, como também o corpo completo dos documentos.

Uma análise descritiva das *features* obtidas a partir dos léxicos será apresentada na próxima seção.

5.3 Análise Descritiva dos Léxicos Gerados na Pesquisa

A Tabela 5.1 exibe exemplos presentes em cada um dos seis léxicos construídos por meio do método apresentado nesta pesquisa, bem como tamanho total de cada um deles. É possível perceber na tabela, a disparidade no total de termos presentes nos LG quando comparados com os léxicos usados como *baseline* e descritos na Seção 4.4. A título de comparação, os seis LG possuem um total de 208 termos. Por outro lado, os três léxicos usados como *baseline* possuem 1.060 ("Bias-inducing terms), 4.560 (MPQA) e 1.496 ((CHOI; WIEBE, 2014)). A listagem completa dos seis léxicos construídos nesta pesquisa pode ser verificada no Apêndice A.

Léxicos	Termos	Num. Termos
Bias	investigation, season, wants, big, republicans, enough, campaigning, another, little, economic, believe, present, sessions, fraud, black	44
Conspiracy	voter, daily, white, however, head, sexual, infowars, legal, actually, executive, wikileaks, attacks, left, fall, described, choice, reasons, muslim	57
Hate	popular, men, jewish, statement, college, police, kings, hit, court, threat, vote, south, radical	36
Satire	story, published, seen, told, posted, pretty, election, first, government, always, work	19
Junk-science	scientific, today, help, chemical, effective, real, remedies, treat, system, protests	24
State	tv, reuters, bulletin, politics, file, army, national, foreign, capital, fighters, issues, control, emergency, protest, movement	28

Tabela 5.1: Exemplos de termos presentes em cada um dos seis léxicos construídos por meio do método proposto nesta pesquisa.

5.3.1 Análise Preliminar dos Dados

Como já descrito, as *features* de classificação usadas para avaliação são resultantes do cálculo de similaridade semântica entre os léxicos de notícias falsas construídos e o *dataset* de notícias. Desta forma, cada notícia tem sua representação reduzida a um vetor de *features* de dimensão igual à quantidade de léxicos utilizados para representá-la. Como descrito na seção anterior, para validar o método de construção automática de léxicos, serão gerados seis léxicos para representar as notícias falsas (i.e. *Conspiracy*, *Bias*, *Hate*, *Satire*, *Junk-Science* e *State*). Nesta seção, as *features* extraídas a partir dos LG serão analisadas para cada um dos três *datasets* utilizados para avaliação.

Análise de Dados para o *Dataset* BSDetector

A Tabela 5.2 exibe a distribuição de termos para o *dataset* de notícias falsas BSDetector e notícias reais "All the News". As informações apresentadas são a média de termos e sentenças presentes nos dados, bem como a quantidade total de termos. Na referida tabela, é possível visualizar um comportamento comum, onde em geral, as notícias reais tendem a ter mais termos e sentenças por documento, quando comparado com notícias falsas. A Figura 5.1 exibe uma nuvem de palavras para as notícias falsas presentes no *dataset*. Na figura, é possível observar termos que estão vinculados ao fragmento temporal do qual as notícias foram colhidas, como por exemplo "Trump" e "Clinton".

A Figura 5.2 exibe a distribuição dos valores das *features* resultantes do cálculo de similaridade semântica para os documentos de notícias falsas e reais. Na imagem, é exibido os boxplots para as notícias falsas e reais. Na imagem, a Figura 5.2(a) exibe a distribuição

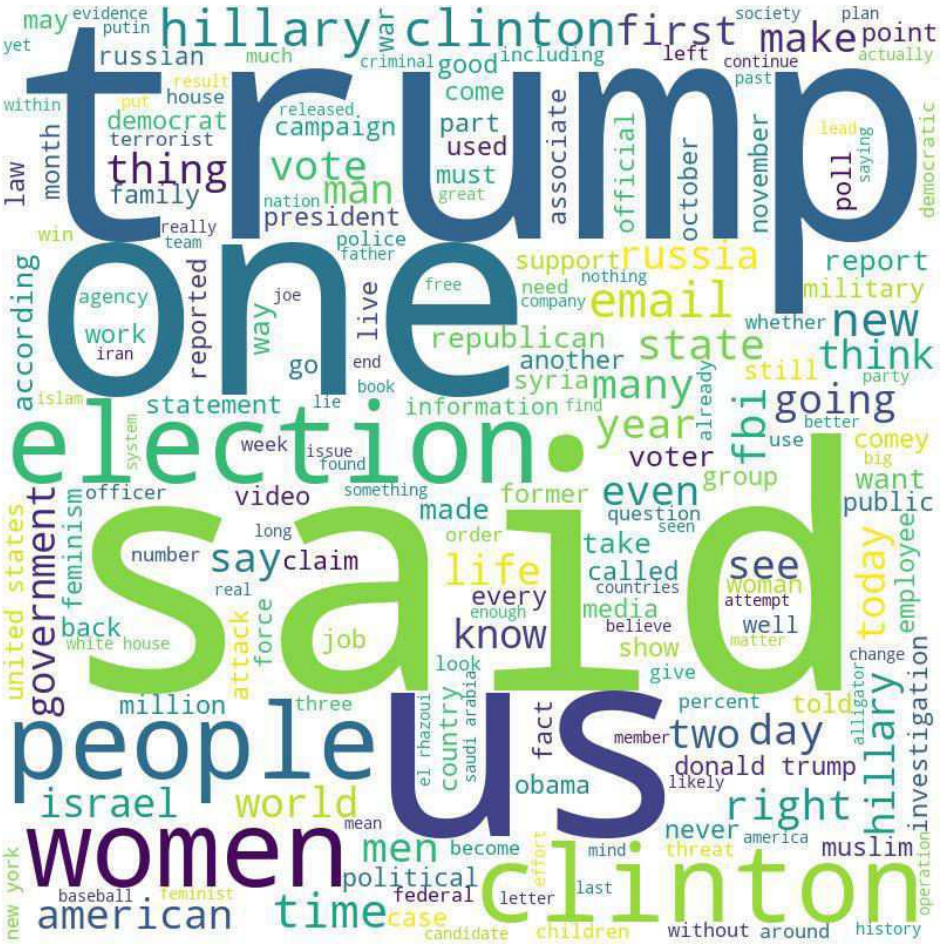

Figura 5.1: Nuvem de palavras contendo termos mais frequentes presentes nas notícias falsas do BSDetector.

Dataset	Termos por documento	Sentenças por documento	Termos únicos
BSDetector	295	20	33186
All the News	558	40	76485

Tabela 5.2: Distribuição de termos para os conjunto de notícias falsas BSdetector e o de notícias reais ‘ “All the News”.

das similaridades entre cada um dos LG (eixo x) em relação às notícias falsas (vermelho) e as notícias reais (azul), considerando as três primeiras sentenças das notícias. É possível notar que, para as notícias falsas, existe uma maior similaridade das notícias em relação aos LG (eixo x), o que é representado por valores maiores de similaridade, quando comparado com as notícias reais. A diferença na distribuição das similaridades é notável para o léxico de sátiras. Valores maiores de similaridade para as notícias falsas sugerem que os léxicos conseguiram captar o contexto semântico presente nas notícias falsas, de forma a se tornarem mais “similares” quando comparado às notícias reais. A princípio, este comportamento

seria o esperado, dado que os LG são gerados considerando os termos presentes nas notícias falsas.

Por sua vez, a Figura 5.2(b) apresenta a distribuição das similaridades considerando o corpo completo das notícias. Neste cenário em particular, as notícias reais apresentaram maiores similaridades com os LG, quando comparado com as próprias notícias falsas. Este comportamento pode estar relacionado à forma como o *dataset* de notícias reais foi obtido pelos autores, onde as notícias das páginas principais dos portais eram colhidas, incluindo textos mais subjetivos como editoriais e artigos de opinião.

A Figura 5.3 exibe os intervalos de confiança dos dados considerando as três primeiras sentenças das notícias (Figura 5.3(a)) e para o corpo completo das notícias (Figura 5.3(b)). Intervalos de confiança com pouca ou nenhuma sobreposição denotam distribuições significativamente distintas entre as notícias falsas e reais.

A Tabela 5.3 exibe a análise estatística comparando os valores das similaridades semânticas entre as notícias falsas e reais, para cada um dos seis léxicos gerados, considerando também as três primeiras sentenças das notícias, bem como o corpo inteiro dos documentos. Na análise, é executado o teste de hipótese de Mann-Whitney, o qual consiste em um teste de hipótese não-paramétrico envolvendo duas amostras independentes, sendo também robusto em cenários para amostras de tamanhos diferentes (MANN; WHITNEY, 1947). Valores de p-valor <= 0,05 representam resultados significativamente distintos. A título de arredondamento, resultados de p-valores menores que $1 * 10^{-5}$ serão exibidos nas tabelas apenas como 0,00. Os resultados demonstram que as similaridades semânticas foram significativamente diferentes entre as notícias falsas e reais para os LG de *Bias*, *Conspiracy*, *Hate* e *Satire* considerando apenas as três primeiras sentenças das notícias. As similaridades para o léxico de sátiras foram as que apresentaram diferenças mais significativas quando considerado apenas as sentenças iniciais dos documentos. Por sua vez, quando considerado o corpo inteiro, o léxico de *Hate* gerou *features* significativamente mais distintas, apresentando um p-valor de aproximadamente $3 * 10^{-17}$. Este resultado demonstra que os léxicos construídos podem, de fato, ser usados como boas *features* para classificação de notícias falsas.

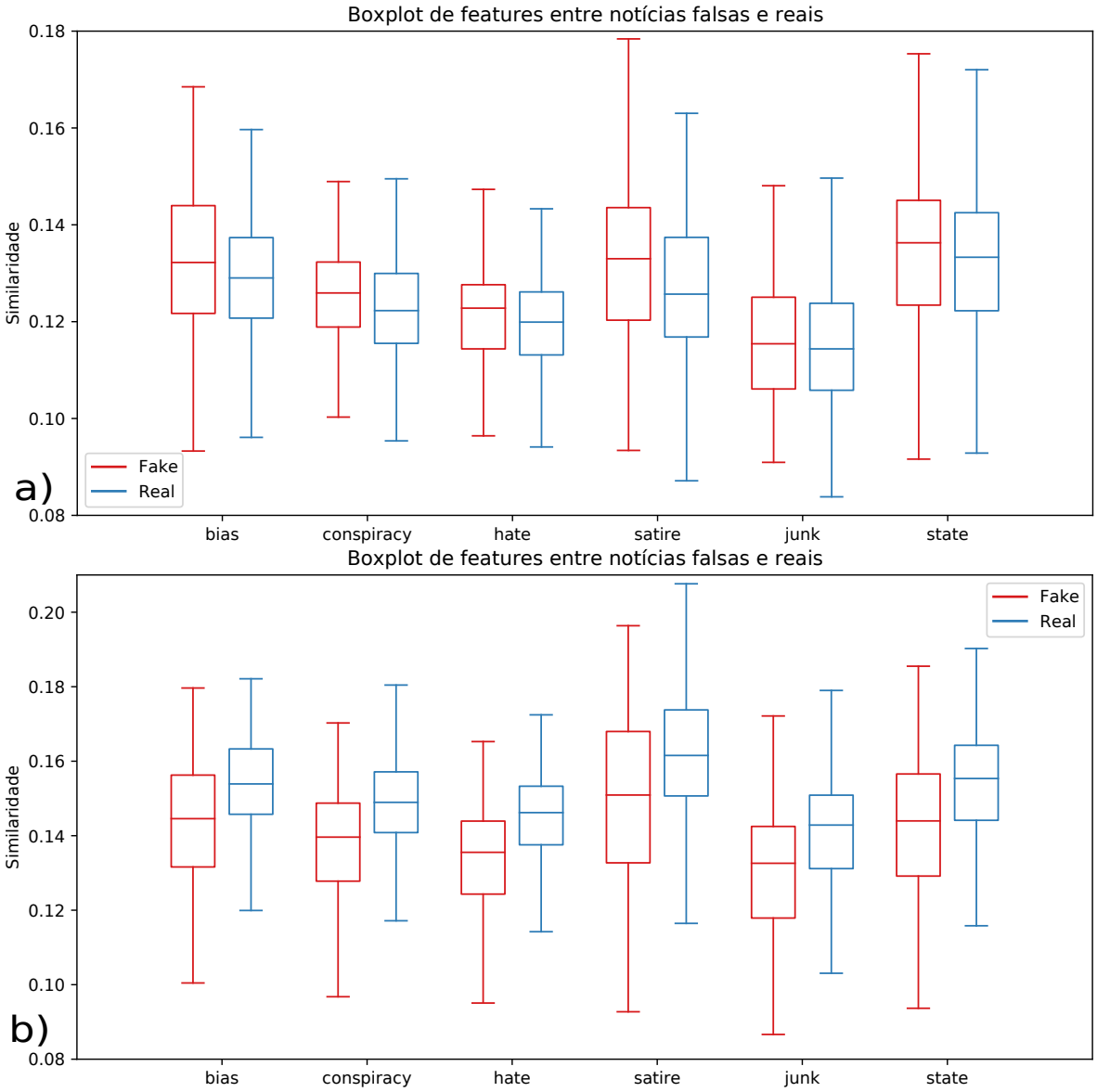

Figura 5.2: Boxplot exibindo a distribuição das *features* baseadas em similaridades semânticas para as notícias falsas e reais. A figura mais acima, exibe a distribuição para as três primeiras sentenças das notícias, enquanto a figura abaixo, exibe a distribuição considerando todo o corpo da notícia.

Análise de Dados para o *Dataset* COVID19

A Tabela 5.4 exibe a distribuição de termos para o *dataset* de notícias relacionadas a COVID19. Este *dataset* possui a peculiaridade de ter notícias falsas maiores que as notícias reais. A Figura 5.4 apresenta a nuvem de palavras para as notícias falsas presente no *dataset* COVID19. É possível notar na imagem, termos fortemente ligados à tópicos relacionados à disseminação da COVID19.

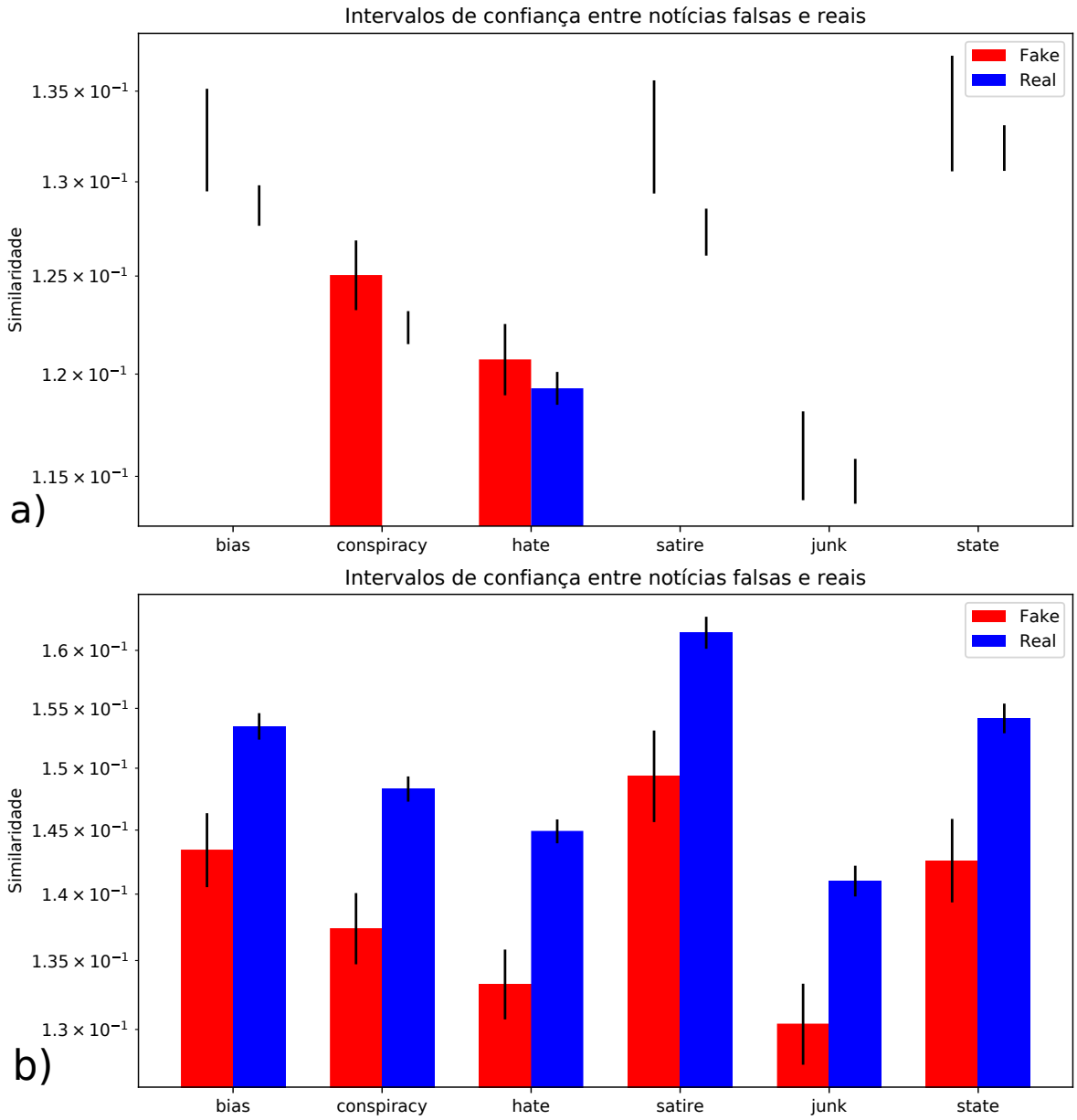

Figura 5.3: Intervalos de confiança das *features* baseadas em similaridade semântica para os dados do BSDetectos (notícias falsas) e "All the news" (notícias reais)

A Figura 5.5 exibe a distribuição das *features* para o *dataset* COVID19. Na Figura, é possível notar que tanto para o cenário que considera as três sentenças iniciais dos documentos 5.5(a), quanto para o cenário considerando todo o corpo das notícias 5.5(b), as notícias falsas apresentam, na maioria dos casos, maior similaridade com cada um dos seis LG, quando comparados com as notícias reais. Na Figura, é possível observar que o léxico gerado a partir das notícias classificadas como *Junk-Science* permitiu a construção de *features* que claramente separam bem as notícias falsas e reais sobre COVID19, indicando que

Similaridade	p-value (3 primeiras sentenças)	p-value (notícias completas)
Bias	0,0048	0,00
Conspiracy	0,0027	0,00
Hate	0,0510	0,00
Satire	0,0004	0,00
Junk-Sci	0,1656	0,00
State	0,1014	0,00

Tabela 5.3: Resultados de testes de hipótese executados sobre as *features* das notícias falsas e reais (BSDetector + "All the News") considerando os seis léxicos gerados automaticamente nesta pesquisa.

Figura 5.4: Nuvem de palavras contendo termos mais frequentes presentes nas notícias falsas do *dataset* COVID19.

Dataset	Termos por documento	Sentenças por documento	Termos únicos
COVID19 - Falsas	689	26	38934
COVID19 - Reais	510	19	36691

Tabela 5.4: Distribuição de termos para o conjunto de notícias relacionadas ao tema COVID19, considerando as notícias falsas e reais presentes nos dados.

este léxico parece adequado à classificação de notícias falsas relacionadas a COVID19.

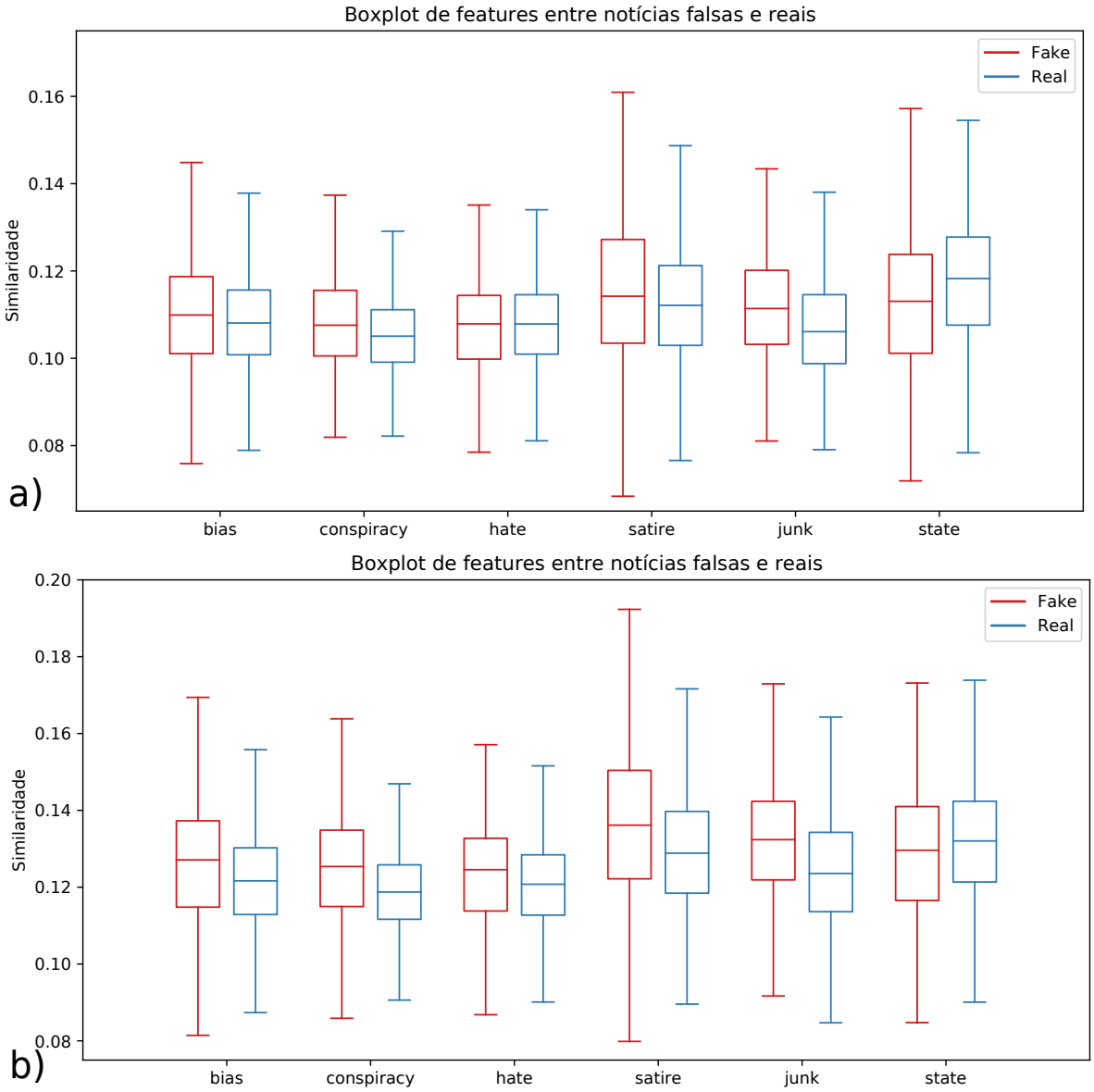

Figura 5.5: Boxplot exibindo a distribuição das *features* baseadas em similaridades semânticas para as notícias falsas e reais do *dataset* COVID19. A figura mais acima, exibe a distribuição para as três primeiras sentenças das notícias, enquanto a figura abaixo exibe a distribuição considerando todo o corpo da notícia.

A Figura 5.6 exibe os intervalos de confiança para as *features* geradas através dos léxicos construídos para os dados de notícias sobre COVID19. Na imagem, é possível obter uma melhor visualização das distribuições das *features*, em especial, para os valores gerados pelo léxico de *Junk-Sciente*, tanto para os dados contendo as sentenças iniciais das notícias (5.6(a)) quanto para o corpo completo dos documentos 5.6(b). A Tabela 5.5 exibe os resultados dos testes de hipóteses para os documentos. Para este *dataset*, é possível notar que

os valores das *features* são significativamente distintos, especialmente para os documentos contendo o corpo completo das notícias.

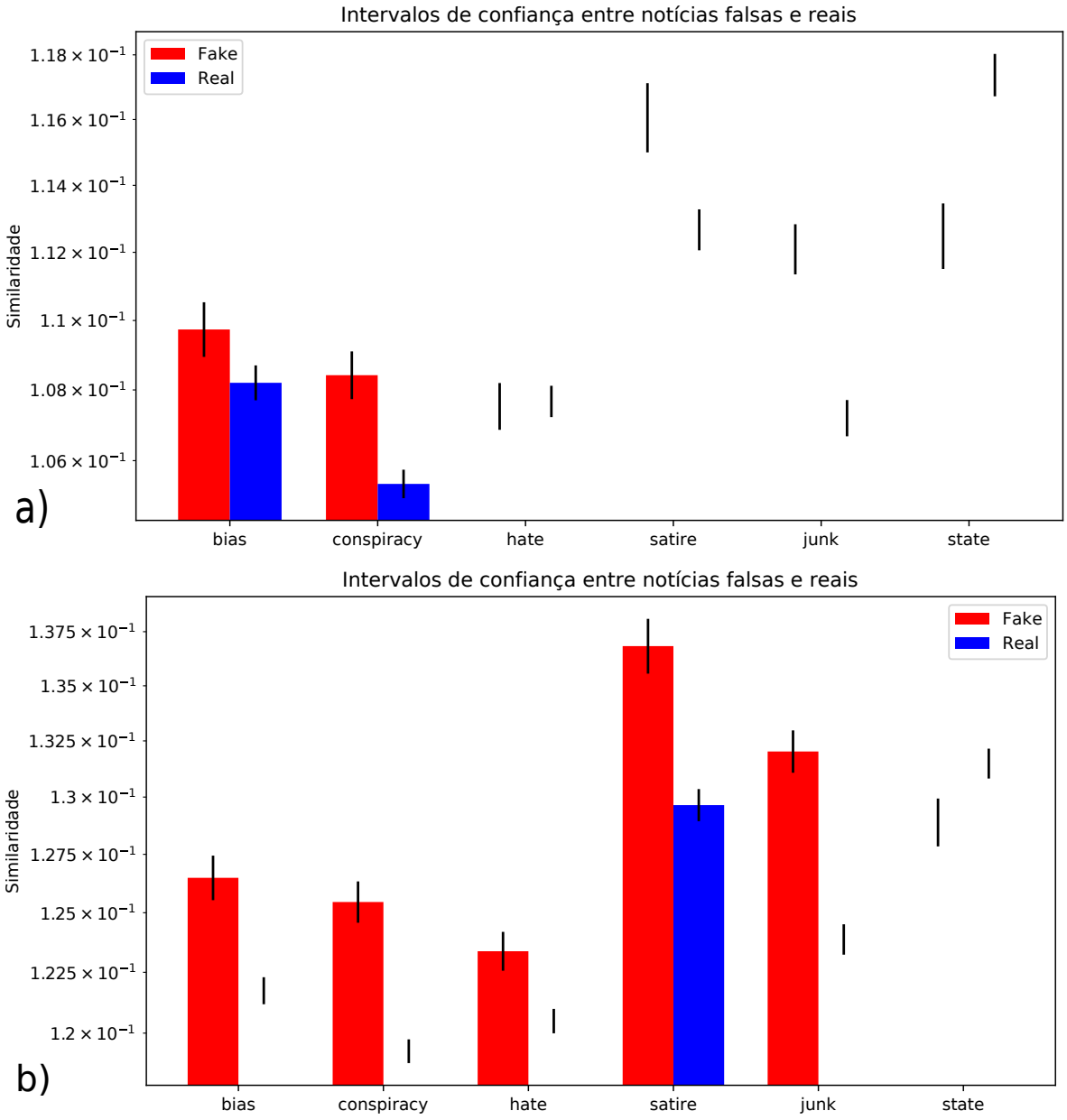

Figura 5.6: Intervalos de confiança das *features* baseadas em similaridade semântica para o *dataset* COVID19

Análise de Dados para o *Dataset Celebrity*

A Tabela 5.6 exibe a distribuição de termos para o *dataset* de notícias *Celebrity*. Neste conjunto de dados, as notícias falsas apresentam uma quantidade média menor de termos e sentenças por documento, quanto comparado com notícias reais. A Figura 5.7 apresenta

Similaridade	p-value (início das notícias)	p-value (notícias completas)
Bias	0.0004	0,00
Conspiracy	0,00	0,00
Hate	0.2389	0,00
Satire	0,00	0,00
Junk-Sci	0,00	0,00
State	0,00	0,00002

Tabela 5.5: Resultados de testes de hipótese executados sobre as *features* das notícias falsas e reais considerando os seis LG para o dataset COVID19.

a nuvem de palavras para as notícias falsas presentes nos dados. Termos como nomes de celebridades podem ser notados na imagem.

Figura 5.7: Nuvem de palavras contendo termos mais frequentes presentes nas notícias falsas do *dataset Celebrity*

A Figura 5.8 exibe a distribuição das *features* para o *dataset Celebrity*. Diferentemente dos conjuntos de notícias anteriores, nota-se que as distribuições das *features* são muito mais similares para as notícias falsas e reais, o que tende a dificultar as classificações para este

Dataset	Termos por documento	Sentenças por documento	Termos únicos
Celebrity - Falsas	399	16	10724
Celebrity - Reais	610	27	14639

Tabela 5.6: Distribuição de termos para o *dataset* de notícias *Celebrity*, considerando as notícias falsas e reais presentes nos dados.

conjunto de dados. A Figura 5.9 exibe os intervalos de confiança para este conjunto de dados *Celebrity*. É possível visualizar sobreposições significativas entre as notícias falsas e reais, denotando distribuições similares entre essas duas classes de documentos, de forma a não ser possível diferenciar, de forma significativa, as notícias falsas e reais para a maioria dos casos.

A Tabela 5.7 exibe os resultados dos testes de hipóteses executados nos dados. É possível verificar que para a maioria dos ensaios, não houve diferença significativa entre as *features* das notícias falsas e reais. Este resultado sugere que os léxicos gerados não foram eficientes para gerar *features* que tornassem as notícias falsas e reais separáveis para este conjunto de dados.

Similaridade	p-value (início das notícias)	p-value (notícias completas)
Bias	0.04136	0.1439
Conspiracy	0.2551	0.0763
Hate	0.2577	0.2144
Satire	0.1693	0.0499
Junk-Sci	0.1232	0.2127
State	0.0332	0.1130

Tabela 5.7: Resultados de testes de hipótese executados sobre as *features* das notícias falsas e reais considerando os seis LG para o *dataset* de notícias *Celebrity*.

5.4 Resultados de Classificação de Notícias Falsas

Esta seção descreve os resultados de classificação obtidos usando as *features* geradas a partir dos léxicos construídos pelo método proposto nessa pesquisa. Os resultados são avaliados e comparados com léxicos presentes na literatura. Os conjuntos de dados usados para avaliação são descritos na Tabela 5.8. As avaliações são executadas por meio de amostragens aleatórias com cem repetições para cada caso de avaliação, considerando, em cada repetição, partições de 80% para treino e 20% para testes. Esta configuração favorece a obtenção de resultados

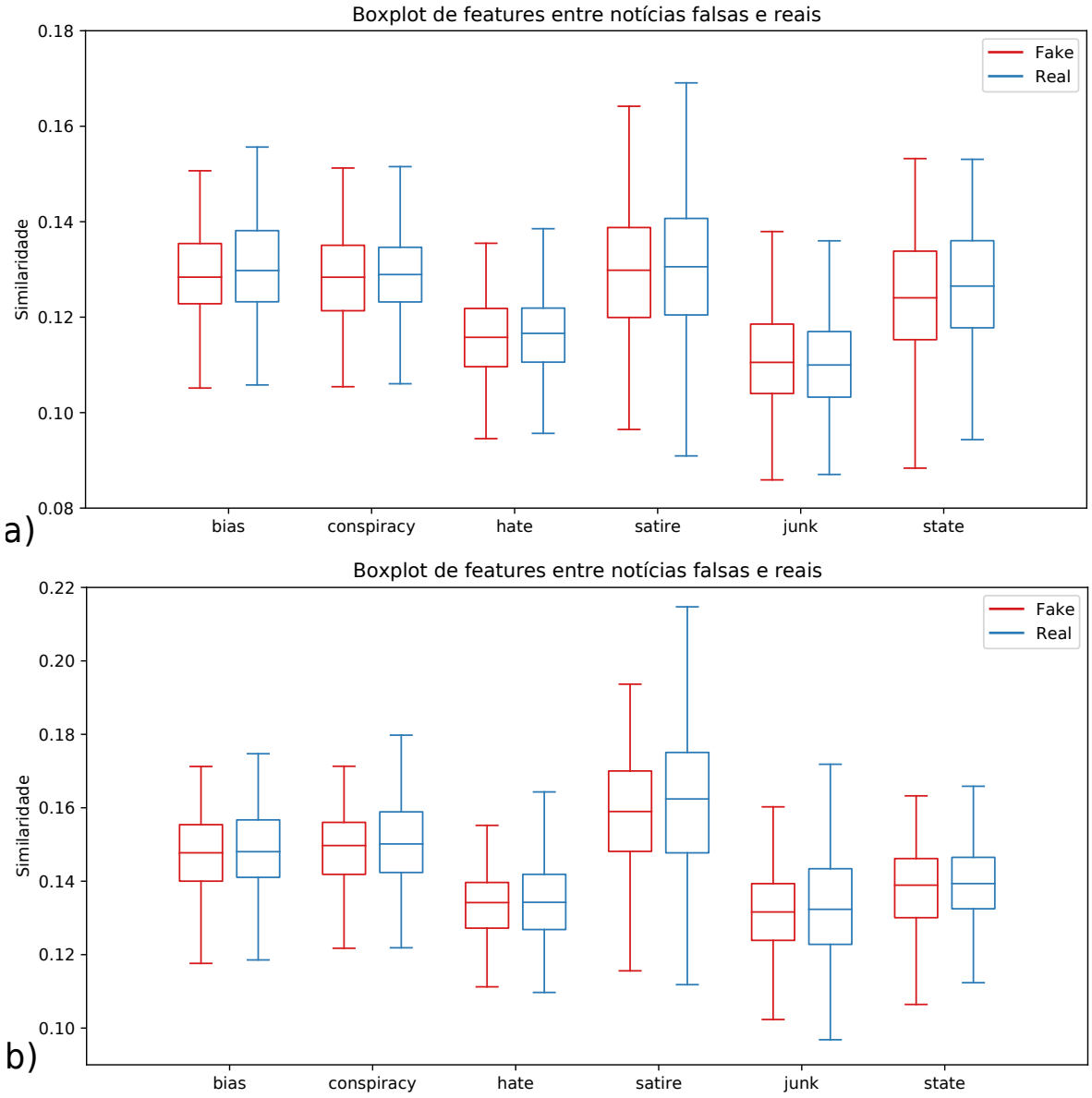

Figura 5.8: Boxplot exibindo a distribuição das *features* baseadas em similaridades semânticas para as notícias falsas e reais do *dataset Celebrity*. A figura mais acima, exibe a distribuição para as três primeiras sentenças das notícias, enquanto a figura mais abaixo exibe a distribuição considerando todo o corpo das notícias.

mais precisos e permite a realização de análises estatísticas mais robustas.

Estudos que avaliam classificações de notícias falsas usualmente utilizam *datasets* balanceados para a avaliação dos modelos preditivos. Isso decorre da dificuldade em se obter dados de notícias falsas com volume suficiente para que os modelos possam aprender como classificar os documentos de forma eficaz. Para manter um balanceamento nos dados, gerando assim um ambiente controlado e propício para a avaliação das *features* construídas,

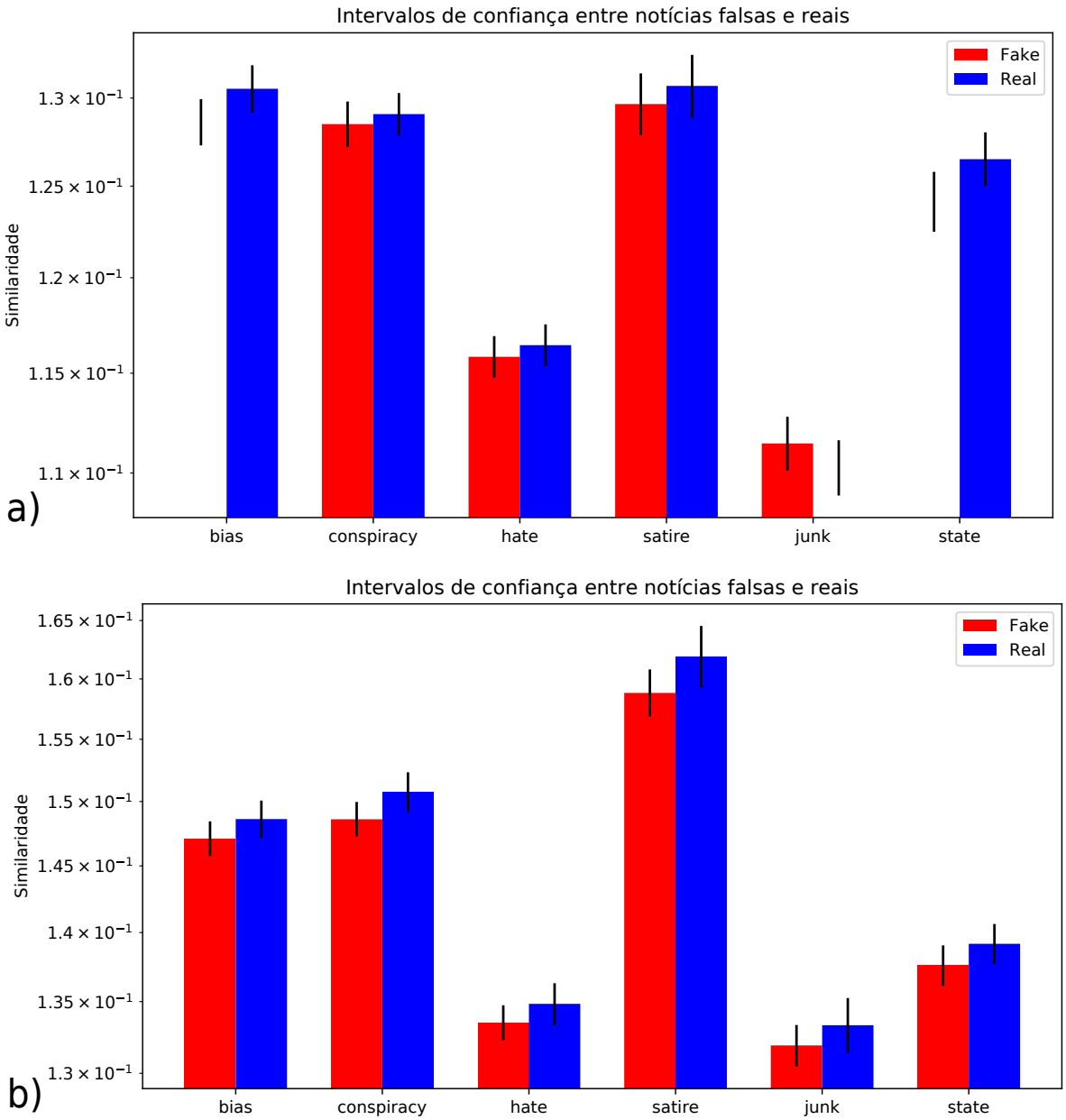

Figura 5.9: Intervalos de confiança das *features* baseadas em similaridade semântica para o *dataset Celebrity*.

é utilizado o algoritmo SMOTE (CHAWLA et al., 2002), o qual realiza um *oversampling* por meio da construção de amostras sintéticas da classe minoritária apenas nos dados de treinamento. Os conjuntos de testes mantém a distribuição original dos dados.

Como descrito anteriormente, os modelos são avaliados considerando as três primeiras sentenças das notícias, bem como com o corpo completo dos documentos. Os resultados serão apresentados, para cada um destes dois cenários, nas próximas subseções.

Dataset	#notícias falsas	#notícias reais
BSDetector (10%)	146	597
COVID19	1055	2035
Celebrity	250	250

Tabela 5.8: Tamanho dos dados usados para avaliação dos modelos de classificação.

5.4.1 Resultados para as Três Primeiras Sentenças dos Documentos

Esta subseção apresenta os resultados de classificação para as notícias considerando as três primeiras sentenças presentes nos documentos de notícias falsas e reais. A Tabela 5.9 apresenta os resultados para os modelos treinados com as *features* derivadas dos LG. Os resultados são apresentados para cada *dataset* avaliado. Os resultados apresentados são as médias obtidas através das cem rodadas de avaliações randomizadas para cada modelo treinado. Os valores entre parênteses representam o desvio padrão dos resultados de classificação.

Na Tabela 5.9, é possível verificar que os melhores resultados foram obtidos para o *dataset* de validação do COVID19, obtendo 0,68 (68%) de ROC-AUC. Este fato sugere que as notícias falsas e reais presentes nesse conjunto de dados em particular, podem simplesmente ser mais separáveis do que os outros dois conjuntos de dados utilizados nas avaliações.

Por outro lado, o *dataset Celebrity* apresentou o pior desempenho de classificação para os modelos treinados com as seis *features* derivadas dos LG. Neste cenário, foi reportado uma ROC-AUC na faixa de 0,48 (48%). Este mesmo comportamento é observado pelos modelos treinados com base nos três léxicos utilizados como *baseline*, sugerindo que as notícias falsas e reais presentes neste *dataset* são menos separáveis, ou seja, de classificação mais difícil. Este fato pode ser observado na seção de análise de dados deste *dataset* (i.e. Seção 5.3.1), onde é possível visualizar diferenças não significativas entre as *features* de classificação das notícias falsas e reais.

	ROC-AUC	PRECISION	RECALL	F1
BSDetector	0.5353 (0.06)	0.2235 (0.07)	0.2862 (0.09)	0.2473 (0.07)
Covid19	0.6887 (0.01)	0.5089 (0.03)	0.5559 (0.03)	0.5307 (0.02)
Celebrity	0.4874 (0.04)	0.4799 (0.06)	0.4905 (0.07)	0.4824 (0.05)

Tabela 5.9: Resultados de classificação com modelos treinados utilizando *features* construías a partir dos LG.

A Tabela 5.10 apresenta os resultados reportados para os modelos treinados a partir das *features* extraídas dos léxicos usados como *baseline*. Os resultados são apresentados em uma

Bias-inducing terms (3 sentenças iniciais)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.5653 (0.05)	0.2633 (0.13)	0.1078 (0.05)	0.1482 (0.07)
Covid19	0.6657 (0.01)	0.5335 (0.04)	0.3783 (0.03)	0.4419 (0.03)
Celebrity	0.5307 (0.05)	0.5165 (0.06)	0.5342 (0.07)	0.5220 (0.05)
MPQA (3 sentenças iniciais)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.5591 (0.05)	0.2450 (0.09)	0.1392 (0.06)	0.1718 (0.06)
Covid19	0.6554 (0.01)	0.5265 (0.03)	0.36626 (0.03)	0.4311 (0.02)
Celebrity	0.5814 (0.04)	0.5409 (0.06)	0.5458 (0.07)	0.5396 (0.05)
Wiebe (2014) (3 sentenças iniciais)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.5696 (0.04)	0.2514 (0.11)	0.1433 (0.06)	0.1761 (0.07)
Covid19	0.6399 (0.02)	0.5139 (0.04)	0.3437 (0.03)	0.4109 (0.02)
Celebrity	0.5289 (0.05)	0.5174 (0.06)	0.5197 (0.06)	0.5158 (0.05)

Tabela 5.10: Resultados de classificação com modelos treinados utilizando *features* construías a partir dos léxicos utilizados como *baseline* considerando as três primeiras sentenças dos documentos.

tabela conjunta, onde é possível visualizar os resultados para cada um dos três *baselines*. A Figura 5.10 exibe a média e os intervalos de confiança em termos de ROC-AUC para os modelos treinados com os LG e os três léxicos usados como *baseline*. A escala do eixo y exibe os resultados de classificação em escala logarítmica, e o eixo x agrupa os resultados para cada *dataset*. Os gráficos presentes na imagem ajudam a visualizar diferenças significativas entre os resultados de classificação dos modelos treinados. Resultados onde os intervalos de confiança não se interceptam representam resultados estatisticamente mais significativos. A Tabela 5.11 apresenta os resultados de testes de hipóteses, em termos de p-valores, executados nos resultados obtidos pelos modelos. Os resultados apresentados endossam os gráficos que apresentam os intervalos de confiança (Figura 5.10) onde é possível notar que os modelos trainados utilizando os LG foram sistematicamente superiores apenas para o dataset de COVID19. Para os outros dois *datasets* (BSDetector e Celebrity), os modelos treinados a partir dos *baselines* foram estatisticamente superiores.

A Tabela 5.12 apresenta os mesmos cenários de classificação, porém, os modelos treinados utilizando as *features* extraídas dos léxicos do *baseline* agora também utilizam as *features* extraídas dos LG. Em outras palavras, os modelos são treinados utilizando *baselines* + LG. Neste cenário, é possível visualizar melhorias significativas para todos os cenários de classificação. A Figura 5.11 apresenta as médias e intervalos de confiança para os resulta-

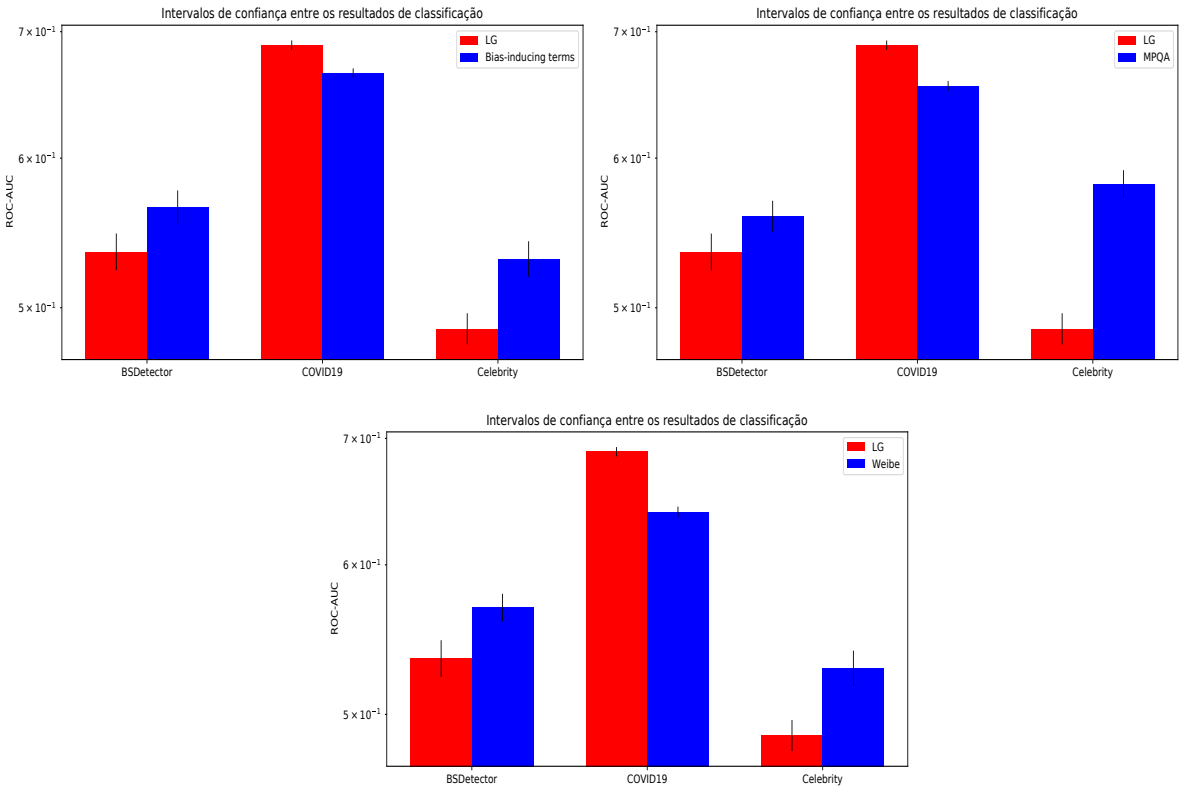

Figura 5.10: Os gráficos apresentam os intervalos de confiança e médias para os resultados de classificação (ROC-AUC) para modelos treinados com os LG comparados com modelos que utilizaram os léxicos do *baseline*. Os resultados foram gerados para modelos treinados com as três primeiras sentenças dos documentos.

	BSDetector	**COVID19**	**Celebrity**
LG vs Bias-inducing terms	0.0002	0,00	0,00
LG vs MPQA	0.002	0,00	0,00
LG vs Wiebe	0,00	0,00	0,00

Tabela 5.11: P-valores resultantes de testes de hipóteses (Mann-Whitney) para avaliar os resultados de classificação (ROC-AUC) para os modelos treinados usando os léxicos gerados (LG) e modelos treinados usando os três léxicos do *baseline*. Os resultados foram gerados para modelos treinados com as três primeiras sentenças dos documentos.

dos de classificação para os modelos treinados utilizando *baselines* + LG e comparados com modelos treinados apenas com os *baselines*. A Tabela 5.13 apresenta os p-valores comparando os modelos, permitindo visualizar com mais facilidade os casos onde houve diferenças significativas entre os modelos citados.

LG + Bias-inducing terms (3 sentenças iniciais)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.6012 (0.05)	0.2319 (0.16)	0.0747 (0.05)	0.1073 (0.07)
Covid19	0.7205 (0.01)	0.5878 (0.03)	0.4454 (0.02)	0.5060 (0.02)
Celebrity	0.5524 (0.05)	0.5339 (0.06)	0.5333 (0.06)	0.5301 (0.04)
LG + MPQA (3 sentenças iniciais)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.5867 (0.05)	0.2917 (0.16)	0.1096 (0.05)	0.1538 (0.07)
Covid19	0.7245 (0.01)	0.5918 (0.03)	0.4717 (0.02)	0.5241 (0.02)
Celebrity	0.5774 (0.04)	0.5516 (0.06)	0.5659 (0.06)	0.5553 (0.05)
LG + Wiebe (2014) (3 sentenças iniciais)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.5852 (0.05)	0.2654 (0.15)	0.0966 (0.05)	0.1367 (0.07)
Covid19	0.7110 (0.01)	0.5799 (0.04)	0.4544 (0.03)	0.5086 (0.02)
Celebrity	0.5331 (0.05)	0.5179 (0.06)	0.5294 (0.07)	0.5203 (0.05)

Tabela 5.12: Resultados de classificação com modelos treinados utilizando *features* construías a partir dos léxicos utilizados como *baseline* e também as *features* construídas a partir dos LG, considerando as três primeiras sentenças dos documentos.

	BSDetector	**COVID19**	**Celebrity**
Bias-ind. terms vs LG+Bias-ind. terms	0,00	0,00	0.007
MPQA vs LG+MPQA	0.0002	0,00	0.33
Wiebe vs LG+Wiebe	0.02	0,00	0.35

Tabela 5.13: P-valores resultantes de testes de hipóteses (Mann-Whitney) para avaliar os resultados de classificação (ROC-AUC) para os modelos treinados usando os léxicos gerados (LG) em conjunto com os léxicos do *baseline*. A comparação é feita com modelos treinados apenas com os *baselines*. Os resultados foram gerados para modelos treinados com as três primeiras sentenças dos documentos.

5.4.2 Resultados para o Corpo Inteiro das Notícias

A Tabela 5.14 apresenta os resultados para os modelos treinados com as *features* baseadas nos LG e avaliados utilizando o corpo inteiro das notícias. A princípio, é possível observar melhores resultados quando os modelos são avaliados usando o corpo inteiro das notícias, quando comparados com modelos avaliados com as três sentenças iniciais dos documentos (i.e. Tabela 5.9). Isso sugere que mesmo os léxicos tendo sido construídos a partir das sentenças iniciais dos documentos, eles podem ser utilizados para construir modelos que classifiquem bem em cenários utilizando as notícias completas. Inclusive, obtendo melhores resultados neste cenário.

A Tabela 5.15 apresenta os resultados dos modelos treinados com os léxicos usados como

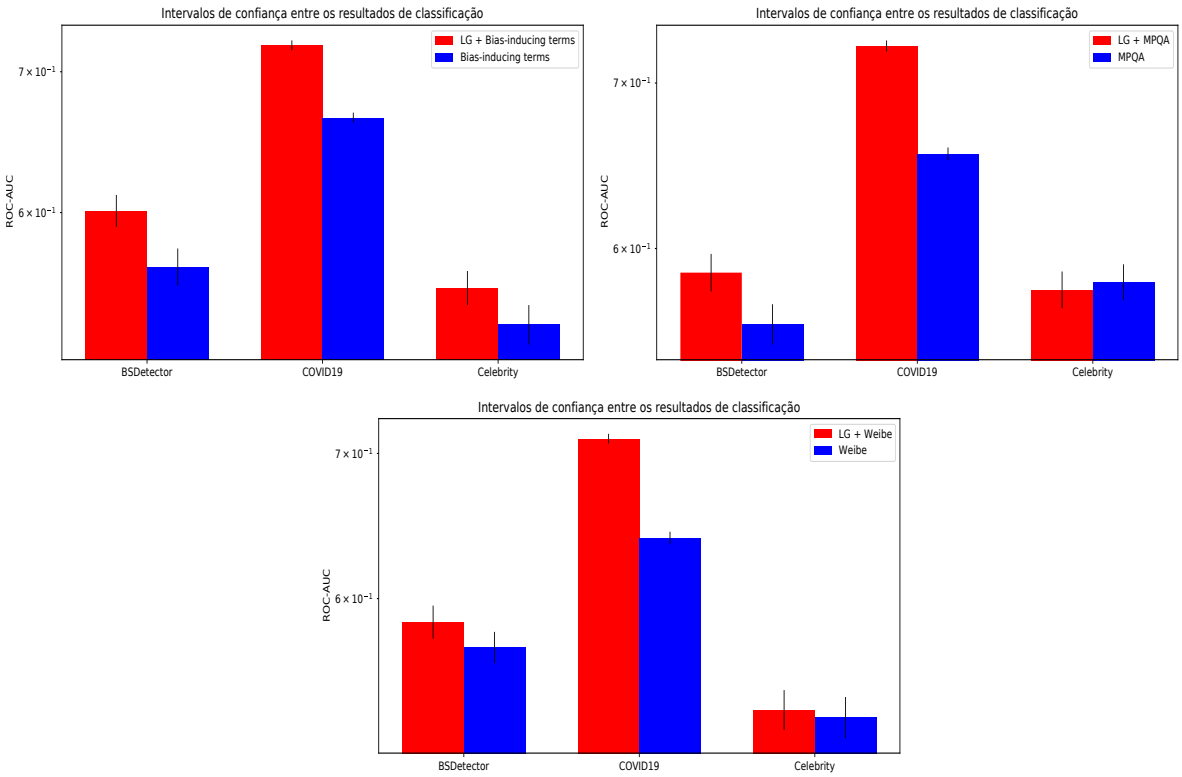

Figura 5.11: Os gráficos apresentam os intervalos de confiança para os resultados de classificação (ROC-AUC) para modelos treinados com os LG em conjunto com as *features* dos léxicos do *baseline*, comparando com modelos treinados apenas com cada um dos três *baselines* separados. Os resultados foram gerados para modelos treinados com as três primeiras sentenças dos documentos.

baseline, porém, avaliados com o corpo inteiro das notícias. É possível notar um melhor desempenho médio para o *dataset* BSDetector quando os modelos são treinados com os léxicos *Bias-inducing terms*. Porém, este resultado não se mostrou significativamente superior quando comparado com os resultados obtidos pelos modelos treinados a partir dos LG e avaliados no mesmo *dataset*. Para os outros dois léxicos, os *baselines* apresentam desempenho inferior, com exceção do *dataset Celebrity* onde os *baselines* superam os LG em todos os ensaios executados. A Figura 5.12 apresenta as médias e intervalos de confiança para os

	ROC-AUC	**PRECISION**	**RECALL**	**F1**
BSDetector	0.6826 (0.04)	0.3677 (0.06)	0.4501 (0.07)	0.4004 (0.05)
Covid19	0.7157 (0.02)	0.5345 (0.03)	0.5717 (0.03)	0.5518 (0.02)
Celebrity	0.5508 (0.05)	0.5316 (0.06)	0.5298 (0.06)	0.5275 (0.05)

Tabela 5.14: Resultados de classificação com modelos treinados utilizando *features* construías a partir dos LG, considerando o corpo inteiro dos documentos para avaliação.

Bias-inducing terms (notícias completas)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.6875 (0.04)	0.5292 (0.10)	0.3042 (0.08)	0.3792 (0.07)
Covid19	0.6796 (0.02)	0.5388 (0.03)	0.4082 (0.03)	0.4639 (0.02)
Celebrity	0.5750 (0.04)	0.5470 (0.06)	0.5563 (0.06)	0.5483 (0.05)
MPQA (notícias completas)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.6143 (0.04)	0.3589 (0.09)	0.2264 (0.07)	0.2703 (0.07)
Covid19	0.6751 (0.01)	0.5389 (0.03)	0.3970 (0.03)	0.4563 (0.02)
Celebrity	0.6539(0.04)	0.6166 (0.06)	0.6204 (0.07)	0.6147 (0.05)
Wiebe (2014) (notícias completas)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.6701 (0.05)	0.4040 (0.09)	0.2723 (0.08)	0.3207 (0.08)
Covid19	0.6351 (0.02)	0.5026 (0.03)	0.3502 (0.02)	0.4118 (0.02)
Celebrity	0.5763 (0.04)	0.5546 (0.05)	0.5502 (0.06)	0.5491 (0.04)

Tabela 5.15: Resultados de classificação com modelos treinados utilizando *features* construías a partir dos léxicos utilizados como *baseline* considerando o corpo inteiro dos documentos.

resultados de classificação dos modelos treinados com os LG comparando-os com modelos treinados com os *baselines*. A Tabela 5.16 apresenta os p-valores comparando os resultados de classificação dos modelos.

A Tabela 5.17 apresenta os resultados para os modelos treinados utilizando tanto as *features* construídas a partir dos LG, quanto com os léxicos utilizados como *baselines*. Os resultados reportados neste cenário são os melhores encontrados neste trabalho. Isso demonstra que os léxicos construídos nesta pesquisa podem também gerar modelos que apresentam predições significativamente superiores, quando usado em conjunto com os léxicos já existentes. A Figura 5.13 apresenta as médias e intervalos de confiança para este cenário, bem como a Tabela 5.18 apresenta os p-valores para o cenário. Neste cenário, é possível verificar que para todos os casos, os modelos treinados com os LG + *baselines* foram melhores que modelos treinados apenas usando os *baselines*.

5.5 Discussão dos Resultados

Esta seção tem como objetivo descrever os resultados obtidos na Seção 5.4 de forma a responder, com base nos resultados, as questões de pesquisa propostas na Seção 1.2.

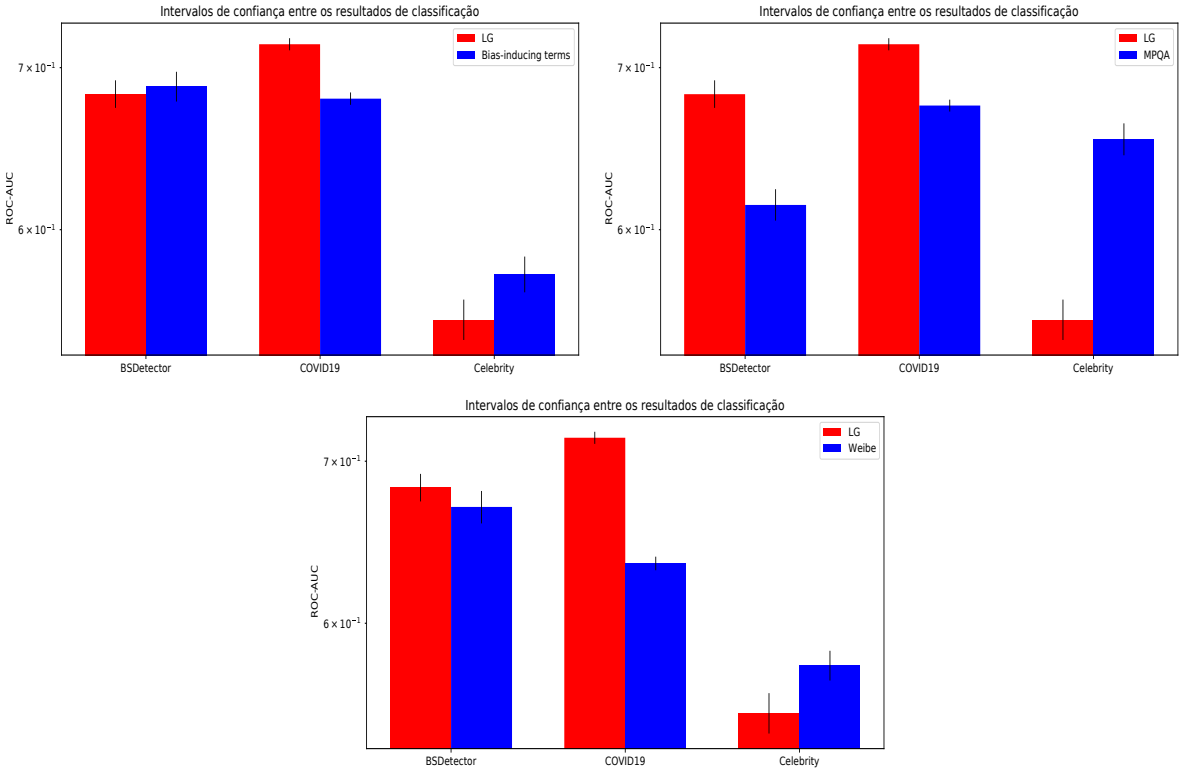

Figura 5.12: Os gráficos apresentam os intervalos de confiança e médias para os resultados de classificação (ROC-AUC) para modelos treinados com os LG comparados com modelos que utilizaram os léxicos do *baseline*. Os resultados foram gerados para modelos treinados com o corpo completo dos documentos.

	BSDetector	**COVID19**	**Celebrity**
LG vs Bias-inducing terms	0.20	0,00	0.0005
LG vs MPQA	0,00	0,00	0,00
LG vs Wiebe	0.03	0,00	0.0001

Tabela 5.16: P-valores resultantes de testes de hipóteses (Mann-Whitney) para avaliar os resultados de classificação (ROC-AUC) para os modelos treinados usando os léxicos gerados (LG) e modelos treinados usando os três léxicos usados como *baseline*. Os resultados foram gerados para modelos treinados com o corpo inteiro dos documentos.

5.5.1 Discussão Geral dos Resultados Encontrados

Para a questão de pesquisa Q1, que pergunta "É possível encontrar diferenças significativas entre notícias falsas e reais utilizando as *features* construídas a partir dos LG?", a resposta a essa questão pode ser obtida a partir da análise preliminar dos dados descrita na Seção 5.3.1. Considerando os dados apresentados na referida seção, é possível observar que, para o *dataset* BSDetector (notícias falsas) e "All the News" (notícias reais), os dados presentes na Tabela 5.3 mostram que houve diferença significativa para três das seis *features* gera-

LG + Bias-inducing terms (notícias completas)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.7728 (0.04)	0.5782 (0.11)	0.3351 (0.09)	0.4164 (0.08)
Covid19	0.7452 (0.01)	0.6077 (0.03)	0.4814 (0.03)	0.5363 (0.02)
Celebrity	0.6210 (0.05)	0.5826 (0.07)	0.6008 (0.07)	0.5879 (0.06)
LG + MPQA (notícias completas)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.7743 (0.04)	0.5541 (0.10)	0.3706 (0.08)	0.4367 (0.07)
Covid19	0.7484 (0.01)	0.6192 (0.03)	0.4904 (0.03)	0.5464 (0.02)
Celebrity	0.6534 (0.04)	0.6042 (0.06)	0.5972 (0.06)	0.5973 (0.05)
LG + Wiebe (2014) (notícias completas)				
	ROC_AUC	PRECISION	RECALL	F1
BS_detector	0.7275 (0.04)	0.5466 (0.11)	0.3338 (0.08)	0.4077 (0.08)
Covid19	0.7354 (0.01)	0.6006 (0.03)	0.4756 (0.03)	0.5300 (0.02)
Celebrity	0.5912 (0.04)	0.5603 (0.06)	0.5673 (0.07)	0.5600 (0.04)

Tabela 5.17: Resultados de classificação com modelos treinados utilizando *features* construías a partir dos léxicos utilizados como *baseline* e também as *features* construídas a partir dos LG, considerando o corpo inteiro dos documentos. Neste cenário, é possível observar resultados significativamente melhores.

	BSDetector	**COVID19**	**Celebrity**
Bias-ind. terms vs LG+Bias-ind. terms	0,00	0,00	0,00
MPQA vs LG+MPQA	0,00	0,00	0.37
Wiebe vs LG+Wiebe	0,00	0,00	0.005

Tabela 5.18: P-valores resultantes de testes de hipóteses (Mann-Whitney) para avaliar os resultados de classificação (ROC-AUC) para os modelos treinados usando os léxicos gerados (LG) em conjunto com os léxicos do *baseline*. A comparação é feita com modelos treinados apenas com os *baselines*. Os resultados foram gerados para modelos treinados com o corpo inteiro dos documentos.

das a partir dos léxicos construídos na pesquisa. Neste caso, foram encontradas diferenças significativas para as *features* de *Bias* (p-value = 0,0048), *Conspiracy* (p-value = 0,0027) e *Satire* (p-value = 0,0004) para modelos treinados e avaliados com as três primeiras sentenças das notícias. Ao considerar as avaliações dos modelos utilizando o corpo inteiro das notícias, foram observadas diferenças significativas ainda mais relevantes e para as seis *features* presentes na mesma Tabela 5.3.

Comportamento similar pôde ser observado ao analisar os testes de hipóteses presentes na tabela 5.5 que considera as *features* geradas para o *dataset* COVID19. Neste cenário, foram observados resultados ainda mais significativos, onde apenas em uma ocasião, foi obtido um p-valor > 0,05. Este caso pode ser visualizado na referida Tabela 5.5 para o LG a partir

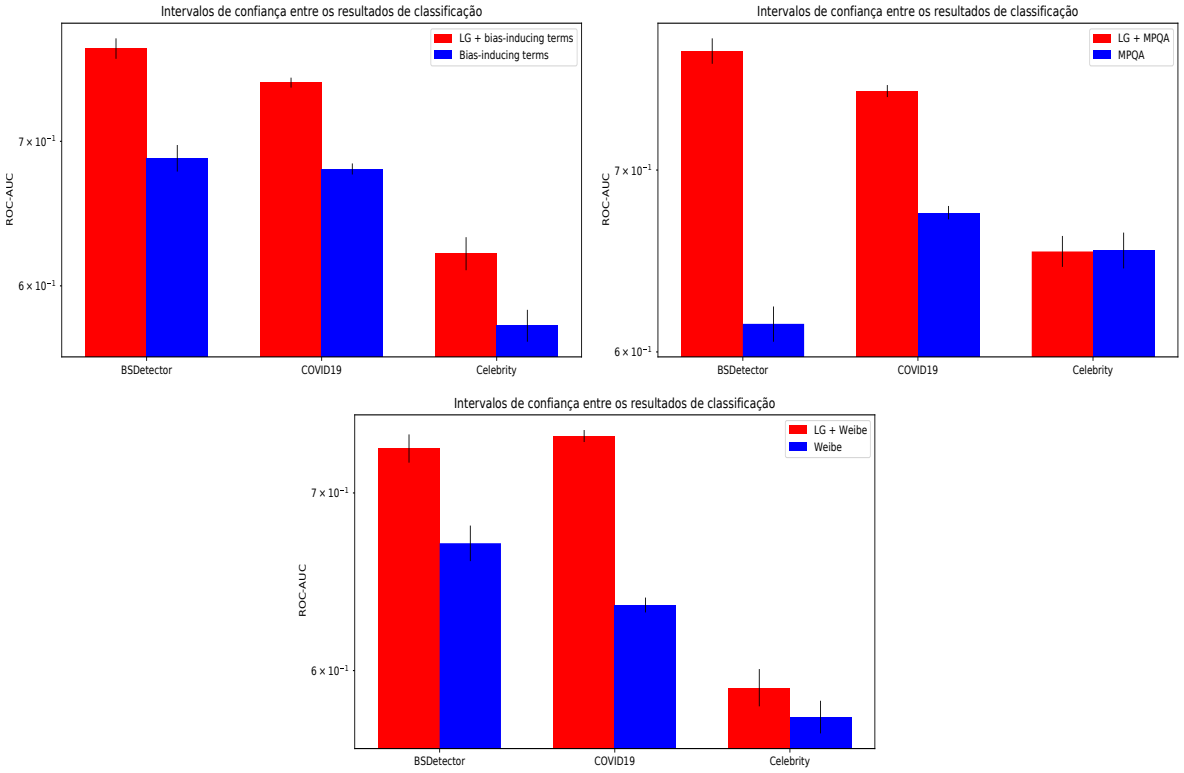

Figura 5.13: Os gráficos apresentam os intervalos de confiança para os resultados de classificação (ROC-AUC) para modelos treinados com os LG em conjunto com as *features* dos três léxicos do *baseline*, comparando com modelos treinados apenas com os *baselines*. Os resultados foram gerados para modelos treinados com o corpo completo dos documentos.

das notícias falsas de ódio (*hate*. Apenas para o *dataset Celebrity* os resultados seguiram um padrão de poucas diferenças significativas entre notícias falsas e reais. Neste conjunto de dados, foram encontradas diferenças significativas apenas para as *features* de *State* (modelos avaliados utilizando as três primeiras sentenças das notícias) com p-valor = 0,0332 e para a *feature Satire*, com p-valor = 0,0499. Dados os resultados dos testes de hipóteses, a Tabela 5.19 quantifica o número de testes onde foi possível encontrar diferenças significativas (H0 rejeitada) para os três *datasets* utilizados. Dados os resultados apresentados na Tabela 5.19 , podemos responder a questão de pesquisa Q1 afirmando que os conjuntos notícias BSDetector e COVID19 apresentaram diferenças significativas para a maior parte dos testes realizados nestes dados (i.e. mais de 50% dos 12 ensaios possíveis para cada base de dados de notícias). Já para o conjunto de dados *Celebrity*, apenas três testes reportaram diferenças significativas entre as notícias falsas e reais, sugerindo que os léxicos gerados não foram tão eficazes para uso com este *dataset*.

Já para a questão de pesquisa Q2, que pergunta “Modelos preditivos treinados a partir

Datasets avaliados	# H0 rejeitadas
BSDetector	75% (9)
COVID19	91% (11)
Celebrity	25% (3)

Tabela 5.19: Quantificação dos testes de hipóteses que reportaram diferenças significativas (p-valor < 0,05) entre as notícias falsas e reais considerando as *features* construídas a partir dos léxicos gerados automaticamente. Esta tabela quantifica o número de testes de hipóteses em que a $H0$ foi rejeitada, considerando as Tabelas 5.3, 5.5 e 5.7.

dos LG possuem desempenho superior a modelos treinados a partir de léxicos construídos manualmente?" pode ser respondida através dos resultados de classificação apresentados pelas Tabelas 5.9 e 5.10 para o caso que considera classificações usando as três primeiras sentenças das notícias, e também nas Tabelas 5.14 e 5.15 que consideram os cenários onde os modelos gerados são avaliados utilizando o corpo completo das notícias.

Em termos gerais, é possível observar, para o cenário de modelos avaliados usando as três primeiras sentenças das notícias que, para a maioria dos cenários, os modelos treinados utilizando os LG (Tabela 5.9) não apresentam melhores resultados quando comparados com os modelos que utilizaram os léxicos do *baseline* proposto (Tabela 5.10). Ao observar os resultados apresentados pela Figura 5.10 em conjunto com a Tabela 5.11, é possível observar que os modelos treinados utilizando os LG foram significativamente superiores apenas para o *dataset* COVID19, ou seja, em 3 dos 9 ensaios de classificação executados. Já para o cenário onde os modelos foram avaliados utilizando o corpo completo das notícias, ao analisarmos os intervalos de confiança presentes na Figura 5.12 bem como os respectivos testes de hipóteses presentes na Tabela 5.16, podemos notar que para 5 casos dos 9 analisados (55%) foi observado desempenho significativamente superior para os modelos treinados a partir dos LG quando comparado com modelos treinados com os três léxicos do *baseline*. Este aumento médio foi de 7,39% para a ROC-AUC. Logo, respondemos a Q2 de forma positiva apenas para o cenário de avaliação que utilizou os documentos de notícias completos, onde o método de construção automática de léxicos permitiu a construção de modelos de classificação de notícias falsas que foram significativamente superiores, em mais de 50% dos casos, a modelos que utilizaram léxicos já existentes na literatura.

Para a questão de pesquisa Q3, que pergunta "Modelos preditivos treinados com base nos LG + *baseline* podem obter um melhor desempenho de classificação de notícias falsas,

quando comparados com os outros modelos gerados na pesquisa?" também foram analisados os intervalos de confiança bem como os testes estatísticos acerca dos resultados de classificação dos modelos. Porém, para a Q3, a comparação é feita considerando os resultados apresentados pelos modelos treinados usando os léxicos do *baseline* + LG. Ou seja, as *features* extraídas a partir dos dois conjuntos de léxicos (i.e. *baseline* e LG) são usadas em conjunto para o treinamento dos modelos preditivos. Estes modelos são comparados com outros modelos treinados unicamente com os léxicos presentes no *baseline*. Para esta questão, os resultados também consideram os dois cenários, onde os modelos são avaliados utilizando apenas as três primeiras sentenças das notícias, bem como com o corpo completo dos documentos. Para o primeiro cenário, ao analisarmos também os intervalos de confiança presentes na Figura 5.11 e os respectivos p-valores presentes na Tabela 5.13, podemos notar que, para 7 dos 9 cenários de avaliação (77%), os modelos treinados com a combinação *baseline* + LG foram significativamente superiores aos modelos treinados usando apenas os léxicos adotados como *baselines*. Este aumento foi de 6,85% na média.

Ao considerarmos o segundo cenário, onde os modelos gerados são avaliados considerando o corpo inteiro das notícias, podemos observar um resultado semelhante. A Figura 5.13 que exibe as médias e os intervalos de confiança para este cenário, bem como a Tabela 5.18 que exibe os respectivos p-valores e que confirmam os resultados. Neste cenário, apenas em um caso não foi possível encontrar uma diferença estatisticamente significativa. Este cenário foi relativo ao modelo treinado com LG + MPQA comparado com um modelo treinado utilizando apenas os léxicos do MPQA para o *dataset Celebrity*. Neste cenário, o p-valor encontrado foi de 0,37 (Tabela 5.18). Ainda assim, para todos os demais cenários, os modelos treinados utilizando a combinação de *features* LG + *Baseline* foram significativamente superiores a modelos treinados utilizando apenas os *baselines* (88% dos casos), com um aumento médio de 11,73% em termos de ROC-AUC. Dado estes resultados, respondemos a Q3 de forma afirmativa, onde modelos que utilizam os LG em conjunto com léxicos já existentes na literatura apresentam desempenho de classificação de notícias falsas significativamente superior a modelos que utilizam apenas os léxicos já existentes. Este resultado, em particular, demonstra que os LG parecem, de fato, contribuir de forma significativa para a classificação de notícias falsas.

De modo geral, os resultados sugerem que os léxicos construídos na pesquisa permitem

a construção de modelos preditivos que são, na maior parte dos caos, significativamente superiores quando avaliados utilizando o corpo completo das notícias. Porém, é necessário destacar que, dada a natureza das notícias falsas de tentar mimetizar as notícias reais, nem sempre será possível identificar diferenças linguísticas de permitam uma boa separação entre estes documentos, levando assim, a uma possível degradação dos resultados em determinado cenários. Também foi possível notar que a combinação de *features* LG + *baselines* permitiu a construção dos melhores modelos preditivos obtidos nesta pesquisa.

5.6 Explicação dos modelos

Para avaliar os modelos do ponto de vista de suas explicações, são utilizados nesta pesquisa os valores reportados pelo SHAP (LUNDBERG; LEE, 2017), que demonstram a relevância de cada *feature* para a predição 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.

5.6.1 Análise Explicativa para Modelos Treinados Utilizando os LG

A Figura 5.14 exibe a plotagem do gráfico de relevância em barras (imagem superior) e a plotagem sumarizada, citada no trabalho dos autores como *summary plot* (imagem inferior) para um modelo treinado tendo como base os LG e treinado com o *dataset* BSDetector. O gráfico de barras exibe, de forma resumida, a relevância de cada uma das *features* do modelo, de forma ordenada (eixo y). O eixo X exibe a média dos valores SHAP (i.e. *shap values*), denotando o impacto que as *features* exercem nas classificações do modelo. Já a plotagem sumarizada exibe, de forma mais detalhada, a influência que cada *feature* exerce sobre as amostras (pontos). Na imagem, o eixo y apresenta as seis *features* de utilizadas em ordem de importância. Estas *features* representam a similaridade semântica de cada documento para um léxico específico. No eixo x, tem-se a faixa de valores que consistem nos valores SHAP, que expressam o peso que cada *feature* exerce para determinar a classificação de uma amostra. O segmento positivo do eixo X 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 no 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* para uma amostra específica.

Na imagem superior da Figura 5.14, é possível notar a magnitude da influência que as três *features* mais relevantes (i.e. *Satire*, *Conspiracy* e *Bias*) têm nas classificações. Na plotagem sumarizada, é possível notar que, para as três *features* mais relevantes (i.e. Satire, Conspiracy e Bias) há uma prevalência de pontos em vermelho na escala positiva do eixo X. Este comportamento expresso no gráfico sugerem que similaridades semânticas maiores entre as notícias e esses três léxicos podem fazer com que os modelos preditivos tendam a classificar as notícias como falsas. Este comportamento pode indicar que, para o *dataset* BSDetector, as notícias falsas parecem possuir, em ordem de relevância, características satíricas, de conspiração e enviesamento do texto mais marcantes que as notícias reais.

Já a Figura 5.15 exibe as mesmas plotagens, porém, para um modelo treinado com o *dataset* COVID19. Na imagem, é possível notar que as três *features* mais relevantes para classificação foram, de acordo com os valores SHAP, as similaridades semânticas extraídas através dos léxicos *State*, *Conspiracy* e *Bias*. Neste cenário, as decisões de classificação dos modelos parecem estar fortemente vinculada a baixos valores de similaridade para o léxico *State*. Esta característica faz sentido, dado que o léxico *State* deriva de documentos de notícias com foco em conflitos armados. Este tópico diverge de assuntos relacionados à disseminação de COVID19. Para os léxicos *Conspiracy* e *Bias*, o comportamento é similar ao apresentado pelo modelo treinado com o conjunto de dados BSDetector. Inclusive, os léxicos de *Conspiracy* e *Bias* aparecem novamente como sendo bastante relevantes para a classificação de notícias falsas.

A Figura 5.16 exibe as plotagens dos valores SHAP para um modelo treinado com o *dataset Celebrity*. Na plotagem sumarizada é possível perceber a ausência de um padrão claro nos pontos, sugerindo que este *dataset* parece ser menos separável ao utilizarmos a representação baseada nas *features* propostas.

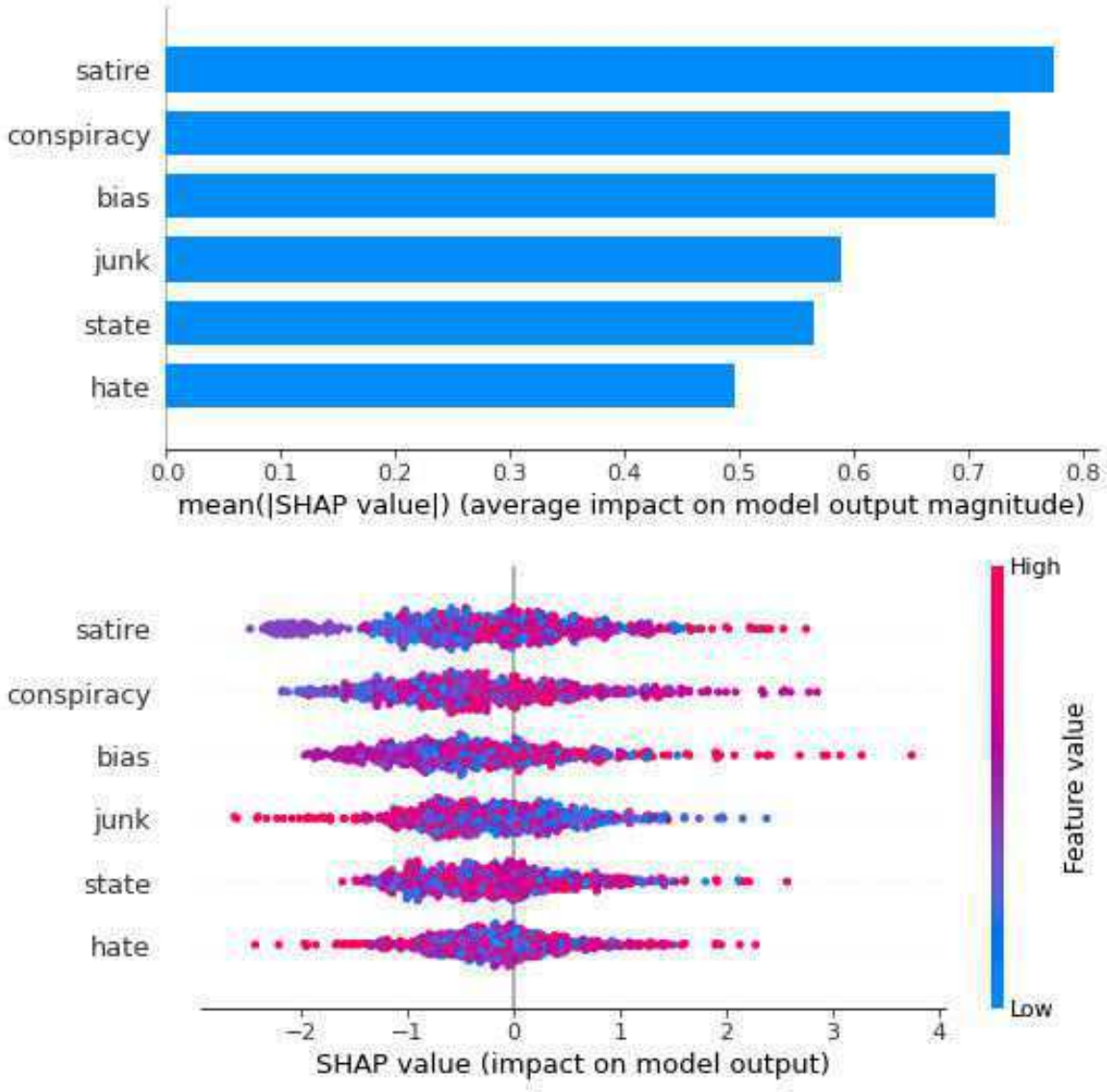

Figura 5.14: Gráfico de barras laterais exibindo a magnitude da relevância que cada *feature* exerce nas classificações do modelo (imagem superior). Plotagem sumarizada (imagem inferior) exibindo o peso que as *features* exercem sobre a decisão de classificação do modelo. No eixo y, estão listadas as seis *features* que formam a representação vetorial de um documento. 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). O gráfico apresenta valores para um modelo treinado com o conjunto de dados BSDetector usando os seis léxicos construídos com a abordagem proposta.

5.6.2 Análise Explicativa para Modelos Utilizando os léxicos *Bias-inducing Terms*

Já a Figura 5.17 apresenta os gráficos reportando os valores SHAP para um modelo utilizando as *features* extraídas a partir dos léxicos *Bias inducing terms* presentes no *baseline*.

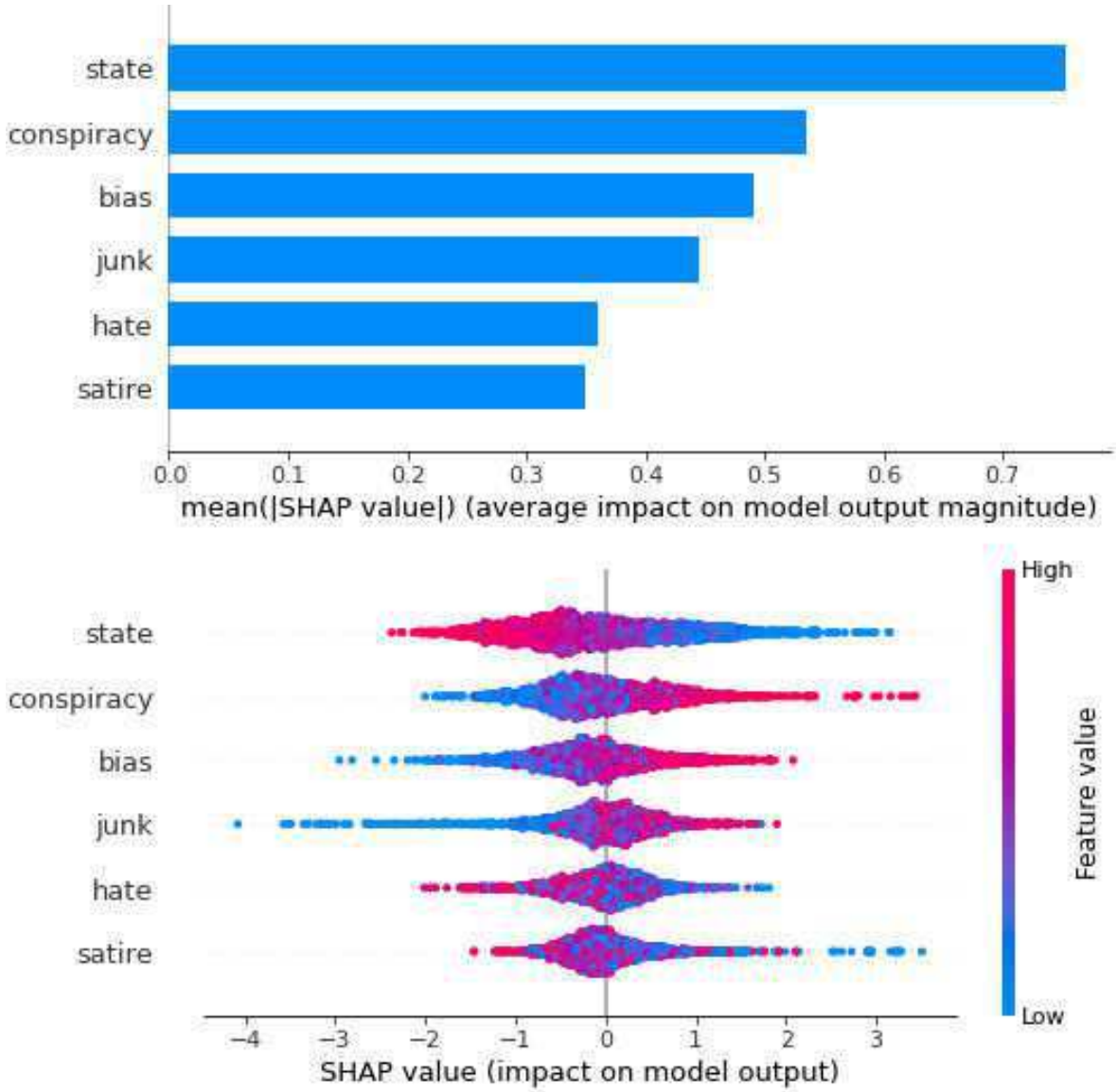

Figura 5.15: Gráfico de barras laterais exibindo a magnitude da relevância que cada *feature* exerce nas classificações do modelo (imagem superior). Plotagem sumarizada (imagem inferior) exibindo o peso que as *features* exercem sobre a decisão de classificação do modelo. No eixo y, estão listadas as seis *features* que formam a representação vetorial de um documento. 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). O gráfico apresenta valores para um modelo treinado com o conjunto de dados COVID19 usando os seis léxicos construídos com a abordagem proposta.

Nesta Figura, o modelo foi treinado utilizando o *dataset* BSDetector. Ao comparar com as imagens apresentadas pelo modelo treinado com os LG utilizando o mesmo *dataset* (Figura 5.14) , é possível verificar que as *features* geradas a partir dos LG permitem uma compre-

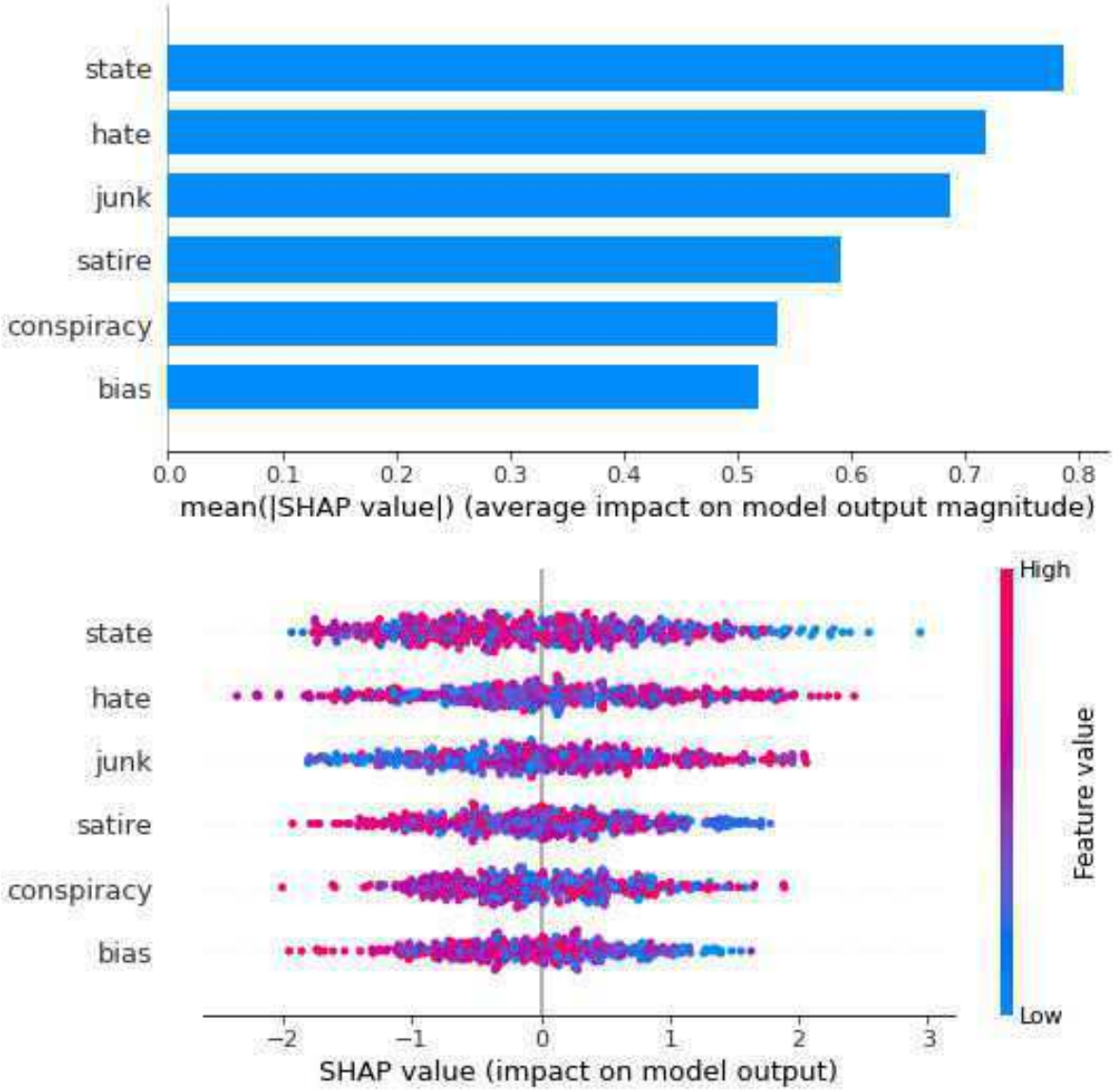

Figura 5.16: Gráfico de barras laterais exibindo a magnitude da relevância que cada *feature* exerce nas classificações do modelo (imagem superior). Plotagem sumarizada (imagem inferior) exibindo o peso que as *features* exercem sobre a decisão de classificação do modelo. No eixo y, estão listadas as seis *features* que formam a representação vetorial de um documento. 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). O gráfico apresenta valores para um modelo treinado com o conjunto de dados *Celebrity* usando os seis léxicos construídos com a abordagem proposta.

ensão mais intuitiva das nuances relacionadas às notícias falsas. Por exemplo, é muito mais objetivo e intuitivo perceber que uma notícia falsa possui mais características satíricas e de viés, do que afirmar que o documento tende a apresentar mais verbos implicativos, quando

comparado com notícias reais. Desta forma, a construção automática de léxicos permite dar mais liberdade para a construção de léxicos que podem ajudar na obtenção de explicações mais claras, o que pode, em alguma medida, auxiliar no entendimento de nuances que eventualmente estejam ocultas nos documentos de notícias.

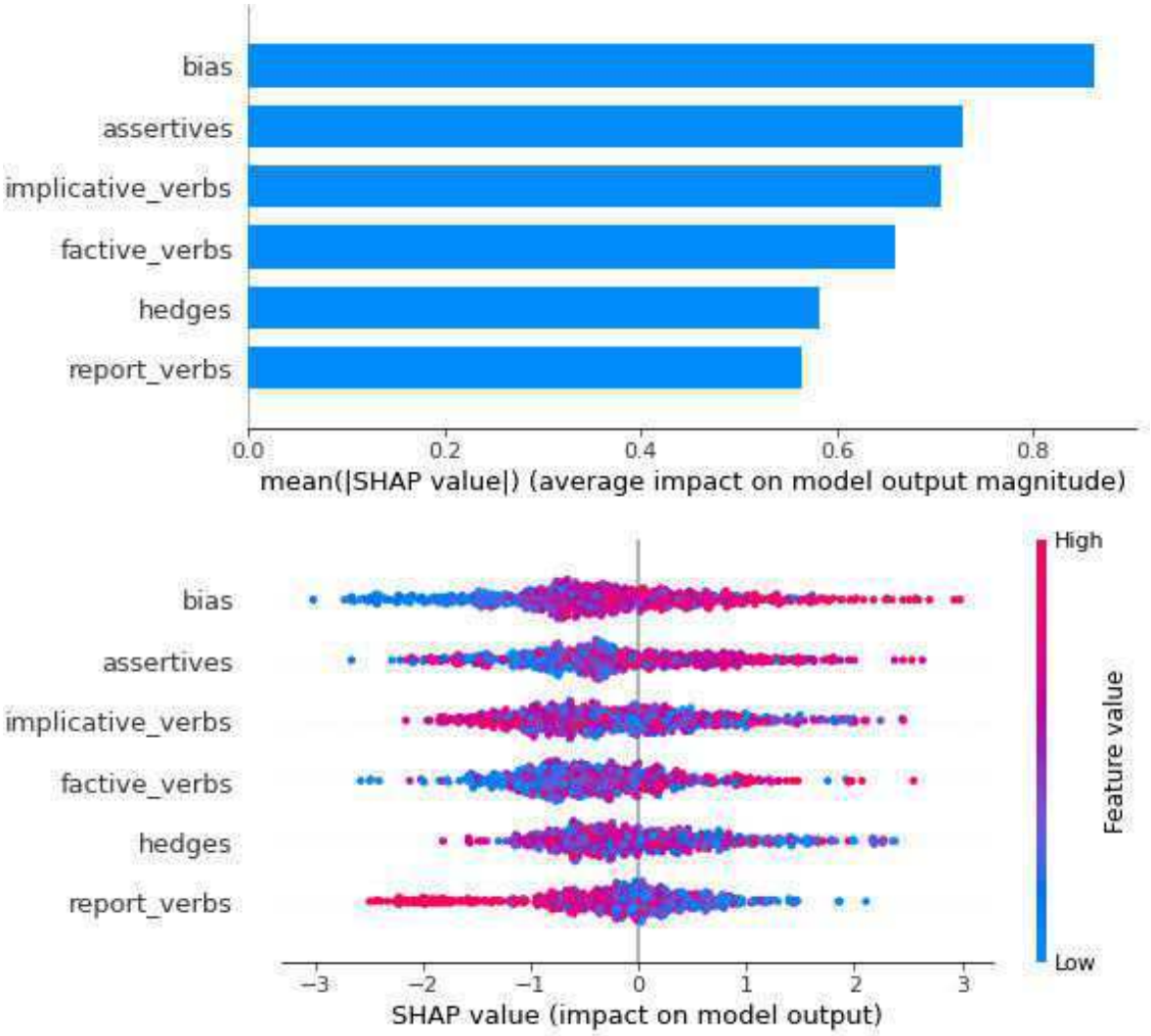

Figura 5.17: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset* BSDetector usando as *features* extraídas a partir dos léxicos "Bias-inducing terms" que compõem o *baseline* de léxicos usado na pesquisa.

As Figuras 5.18 e 5.19 apresentam as plotagens do modelo, porém agora treinado utilizando os *datasets* COVID19 e *Celebrity*, respectivamente. É possível notar que, para os três conjuntos de dados de notícias, não parece haver um padrão claro na ordem de relevância das *features*. O contrário pode ser observado quando o modelo é treinado com base nos LG, onde as *features* relacionadas aos léxicos *Conspiracy* e *Bias* se revelaram entre as três mais

relevantes tando para os dados de notícias BSDetector quanto para as notícias relacionadas a COVID19.

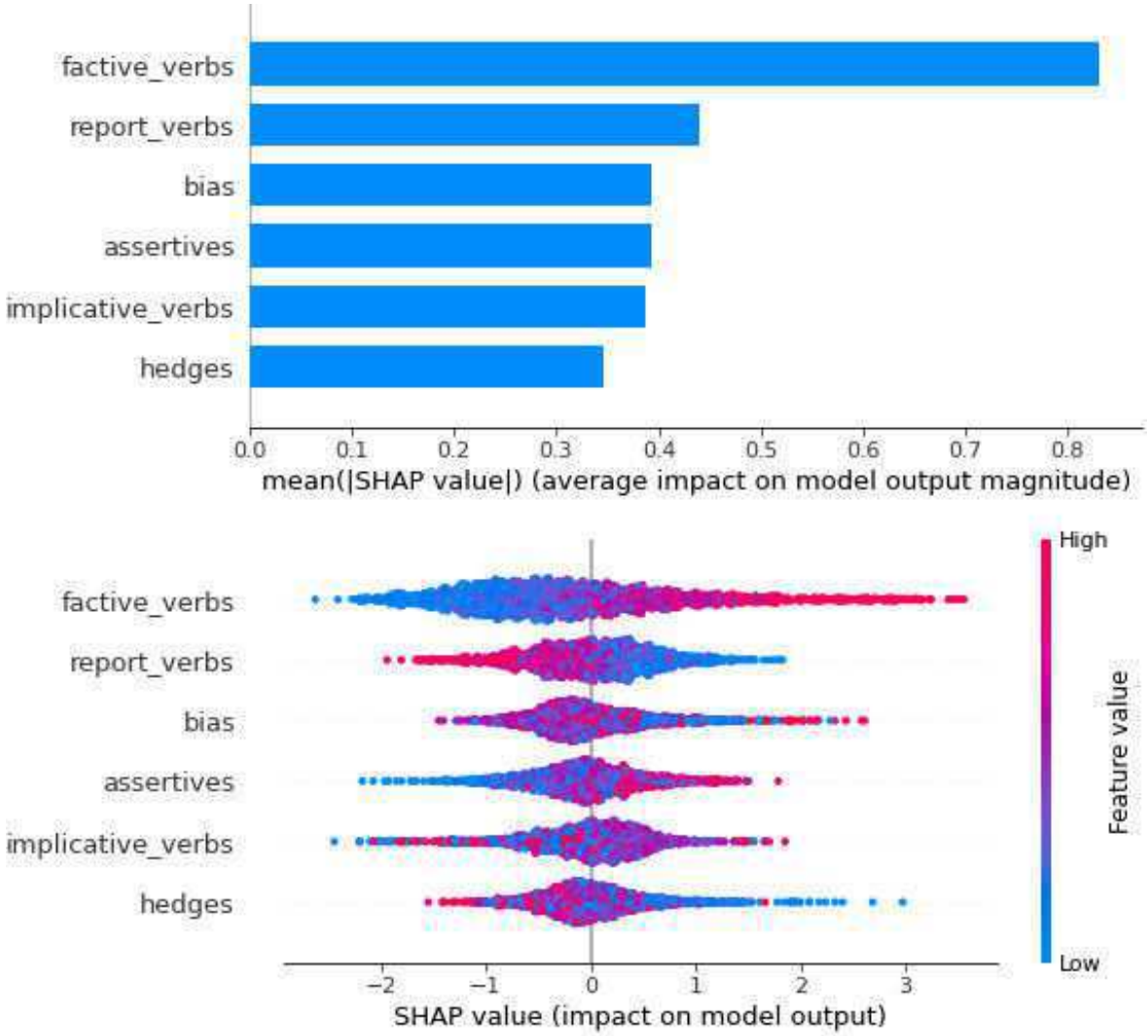

Figura 5.18: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset* COVID19 usando as *features* extraídas a partir dos léxicos "Bias-inducing terms" que compõem o *baseline* de léxicos usado na pesquisa.

5.6.3 Análise Explicativa para Modelos Utilizando os léxicos MPQA

As Figuras 5.20 apresentam os gráficos dos modelos com os valores SHAP para as *features* extraídas a partir dos léxicos do projeto MPQA e avaliados considerando as notícias do BSDetector. Na Figura, é possível notar que, para este conjunto de notícias falsas e reais, a ocorrência de termos que apresentam uma semântica negativa (i.e. dentro da polaridade de sentimentos negativos) parece tendenciar o modelos a classificar documentos como sendo

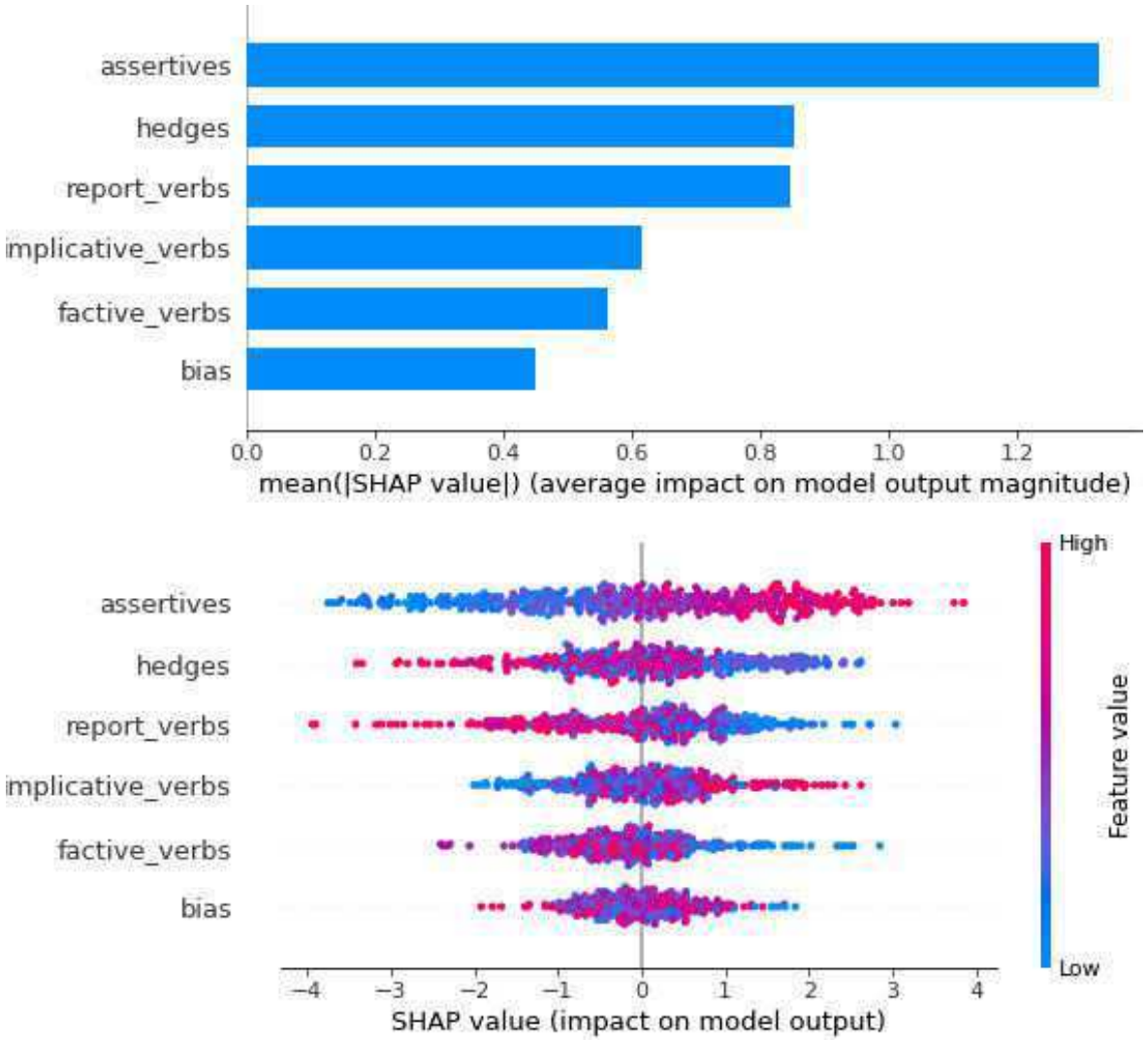

Figura 5.19: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset Celebrity* usando as *features* extraídas a partir dos léxicos "Bias-inducing terms" que compõem o *baseline* de léxicos usado na pesquisa.

notícias falsas. Para o léxico de termos de polaridade positiva, o efeito é o inverso. Em outras palavras, segundo os valores SHAP obtidos para este modelo e treinado com este conjunto de dados, uma notícia que apresenta uma maior similaridade semântica com o léxico de termos positivos tenderia a ser classificada como uma notícia real.

Já para a Figura 5.21, que apresenta os gráficos para as notícias o conjunto de dados de notícias relacionadas à disseminação de COVID19, o efeito se mostra diferente. Na Figura, é possível notar que uma maior similaridade semântica das notícia, tanto para o léxico de termos positivos e negativos, parece aumentar as chances de uma notícia ser classificada como falsa pelo modelo. Esta situação pode indicar que as notícias falsas relacionadas a COVID19

presentes no conjunto de dados apresentam uma correspondência semântica tanto para as polaridades positivas quanto as negativas. Já as notícias reais sobre COVID, possivelmente por se tratarem de documentos relacionados com tópicos de saúde, podem apresentar uma semântica menos emotiva e mais sóbria. A Figura 5.22 apresenta a mesma análise, porém agora relacionada ao conjunto de notícias do *dataset Celebrity*. Para este modelo, a visualização das influências exercidas pelas *features* de classificação parece ser mais difícil de se observar. Ainda assim, ao observar as extremidades dos pontos apresentados na plotagem sumarizada, percebe-se que a prevalência de uma semântica negativa nos documentos tende a influenciar o modelo a classificar uma notícia como falsa. O efeito inverso é observado para o léxico de termos positivos.

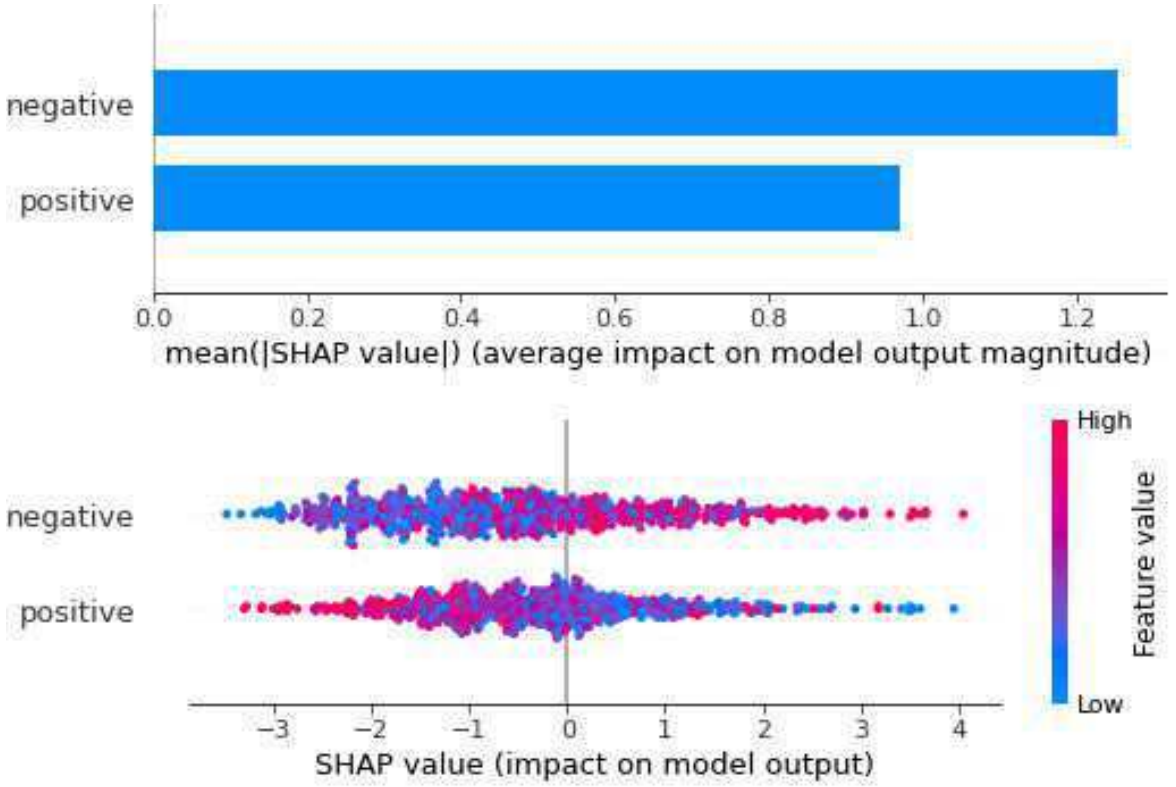

Figura 5.20: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset* BSDetector usando as *features* extraídas a partir dos léxicos do projeto MPQA, que compõem o *baseline* de léxicos usado na pesquisa.

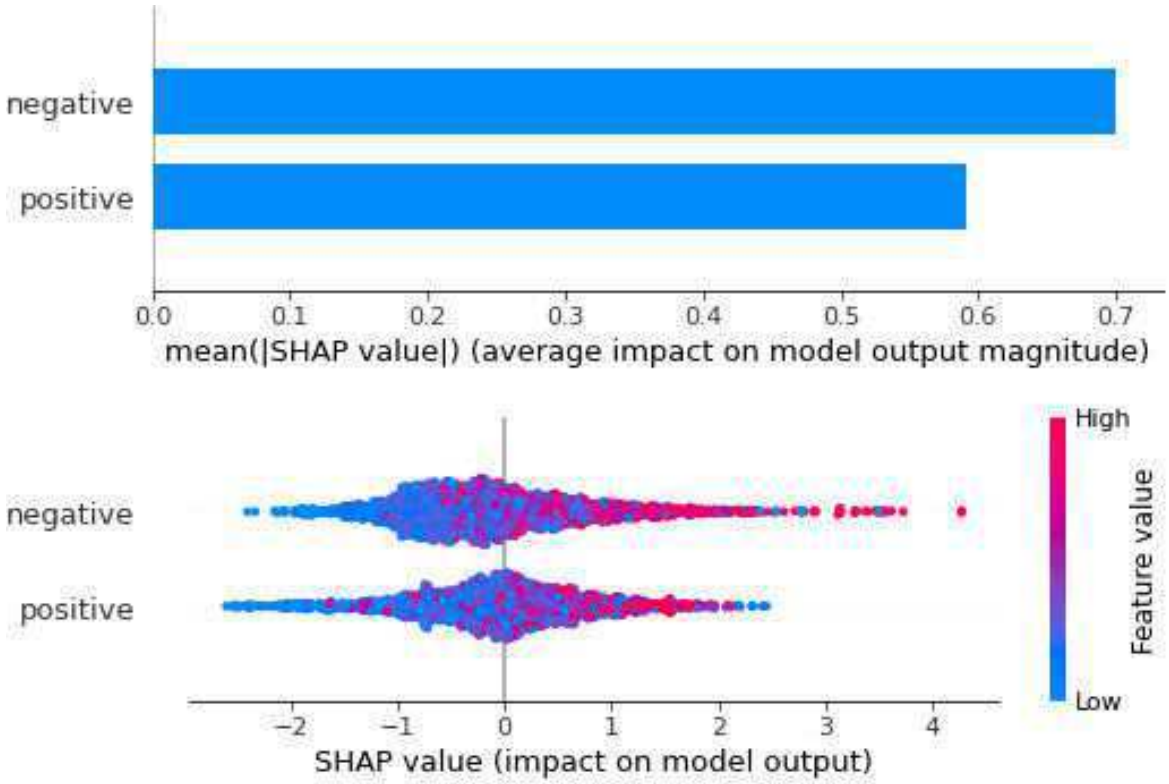

Figura 5.21: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset* COVID19 usando as *features* extraídas a partir dos léxicos do projeto MPQA, que compõem o *baseline* de léxicos usado na pesquisa.

5.6.4 Análise Explicativa para Modelos Utilizando os léxicos Wiebe

A Figura 5.23 apresenta a análise de valores SHAP para o modelo treinado utilizando as *features* extraídas para os léxicos apresentados por Choi e Wiebe (2014) e avaliado utilizando o dados de notícias do BSDetector. Neste cenário, a influência das *features* geradas a partir deste conjunto de léxicos também é de difícil observação. Já para a Figura 5.24, onde o modelo é avaliado utilizando o *dataset* COVID19, o mesmo padrão observado para os léxicos do MPQA são observados. Neste caso, a presença de uma semântica positiva e negativa parece tendenciar o modelo a classificar uma notícia como falsa.

Já para a Figura 5.25, onde o modelo é treinado utilizando os dados do conjunto de dados *Celebrity*, não é possível visualizar, de forma clara, algum padrão de influência das *features* no modelo. Este fato corrobora com os resultados pouco relevantes na classificação das notícias presentes neste conjunto de dados.

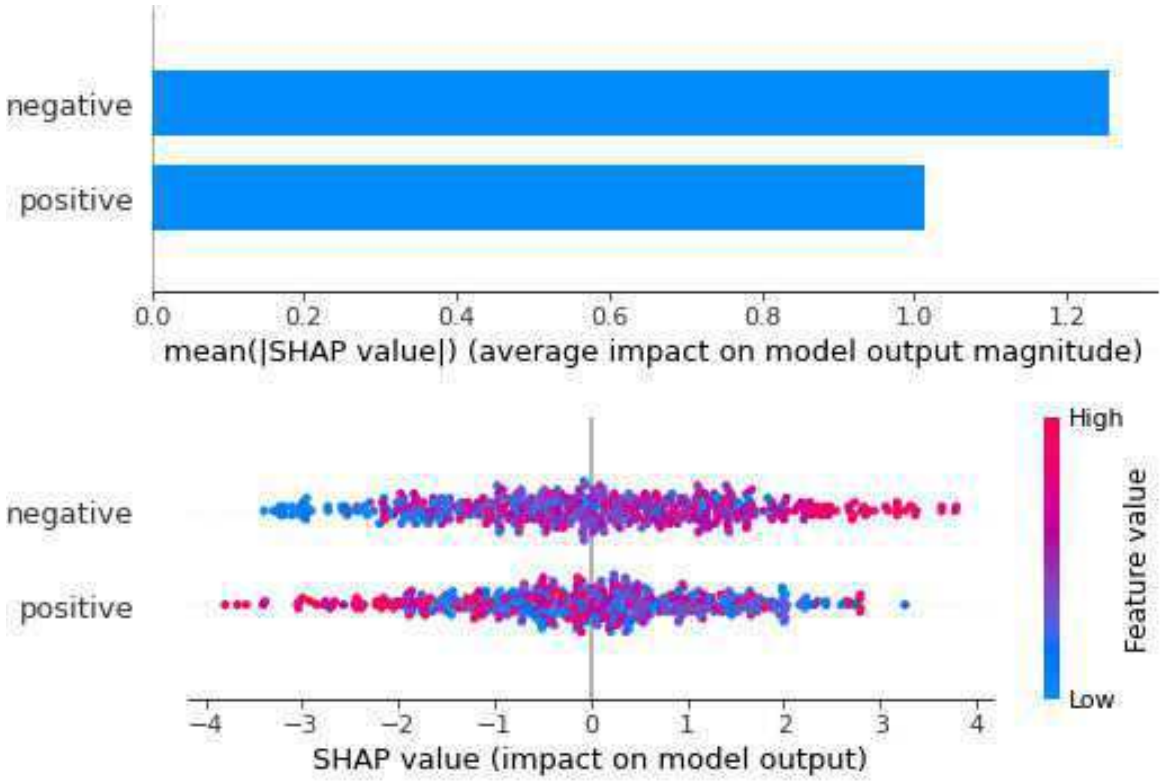

Figura 5.22: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset Celebrity* usando as *features* extraídas a partir dos léxicos do projeto MPQA, que compõem o *baseline* de léxicos usado na pesquisa.

5.7 Análise Explicativa para Modelo utilizando *Bag-of-Words*

Como pôde ser observado no Capítulo 3 de Trabalhos Relacionados desta Tese, diversas abordagens de classificação utilizam *features* baseadas em BoW, sendo uma das abordagens mais utilizadas para a classificação de notícias falsas. Frequentemente, o seu uso está associado a elevadas performances de classificação (i.e. acurácia superior a 90%), superando em muitos casos, modelos de Aprendizagem Profunda mais robustos (KHAN et al., 2019). Porém, este tipo de representação não permite a construção de explicações intuitivas quando utilizado em problemas de classificação de notícias falsas. Para exemplificar esta dificuldade, a Figura 5.26 apresenta um exemplo de explicação de um modelo preditivo treinado para a classificação de notícias falsas utilizando BoW e TFIDF. Na imagem, é possível notar que as principais *features* estão fortemente ligadas a assuntos específicos presentes nos dados. Situações como essa tendem a gerar problemas de *overfitting* nos modelos de classificação,

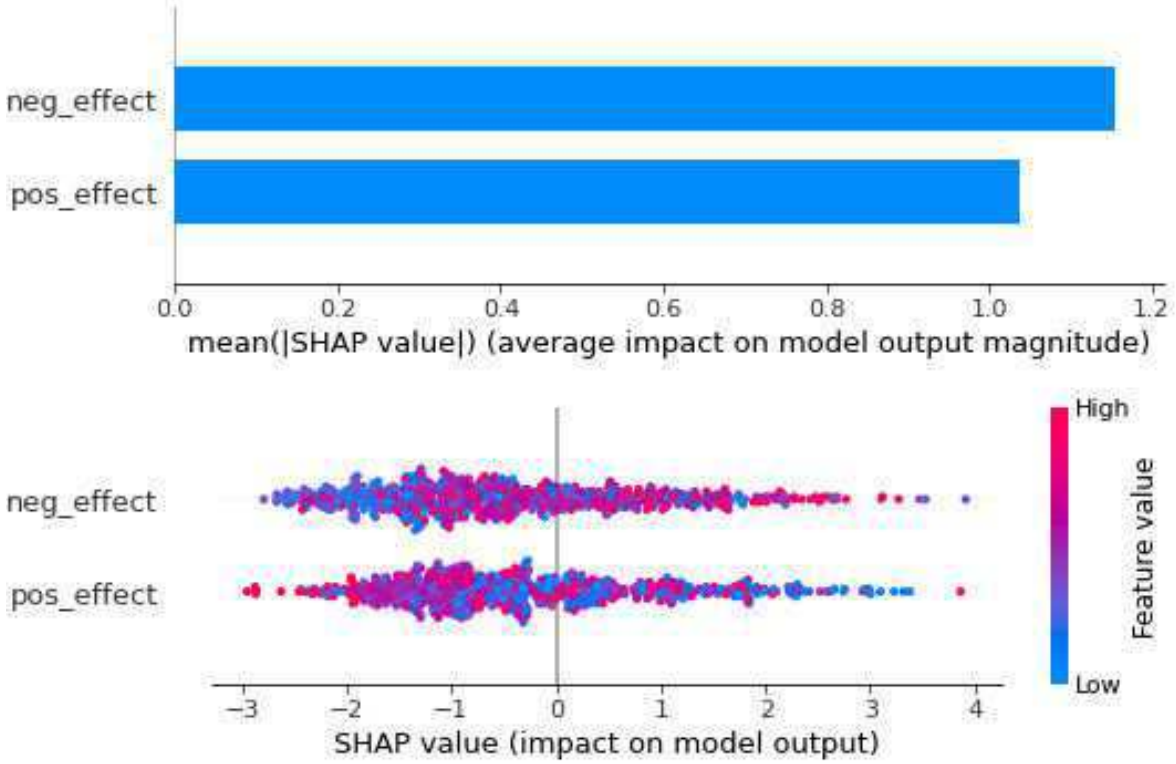

Figura 5.23: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset* BSDetector usando as *features* extraídas a partir dos léxicos apresentados por Choi e Wiebe (2014), que compõem o *baseline* de léxicos usado na pesquisa.

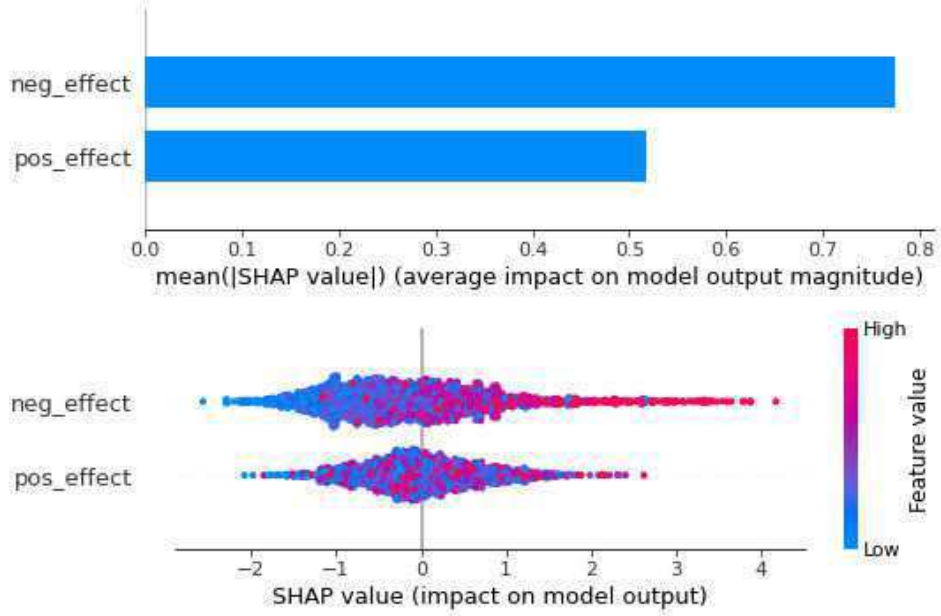

Figura 5.24: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset* COVID19 usando as *features* extraídas a partir dos léxicos apresentados por Choi e Wiebe (2014), que compõem o *baseline* de léxicos usado na pesquisa.

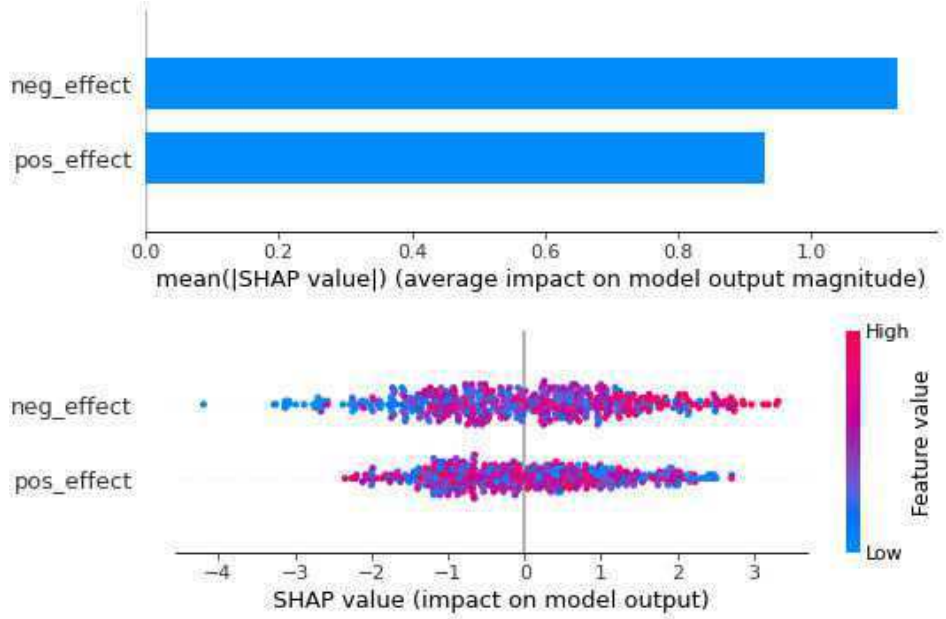

Figura 5.25: Os gráficos apresentam os valores SHAP para um modelo treinado com o *dataset Celebrity* usando as *features* extraídas a partir dos léxicos apresentados por Choi e Wiebe (2014), que compõem o *baseline* de léxicos usado na pesquisa.

pois muitas *features* tendem a raramente aparecer nos documentos, ou mesmo só ocorrer com mais frequência em documentos de uma determinada classe. Essas condições acabam por reduzir a capacidade de generalização de tais modelos (CHEN et al., 2013a; FAUSTINI; COVOES, 2020). A Figura 5.27 exibe, para o mesmo modelo, a plotagem sumarizada para as mesmas vinte *features* mais relevantes. É possível notar, para o termo "Trump" que, pesos TFIDF maiores tentem aumentam as chances de que o modelo classifique uma notícia como falsa. Esta explicação apenas sugere que o termo "Trump" parece estar mais vinculado às notícias falsas do que nas reais, não permitindo compreender nada além da simples presença ou não de termos nos documentos.

Ao contrário das explicações fornecidas pelo modelo utilizando BoW, os modelos treinados a partir dos léxicos construídos nesta pesquisa podem gerar explicações que estejam mais relacionadas ao escopo de notícias falsas. Por exemplo, na Figura 5.15, através da plotagem sumarizada, é possível notar claramente que as notícias falsas relacionadas à COVID19 parecem possuir claras tendencias conspiratórias e de enviesamento textual.

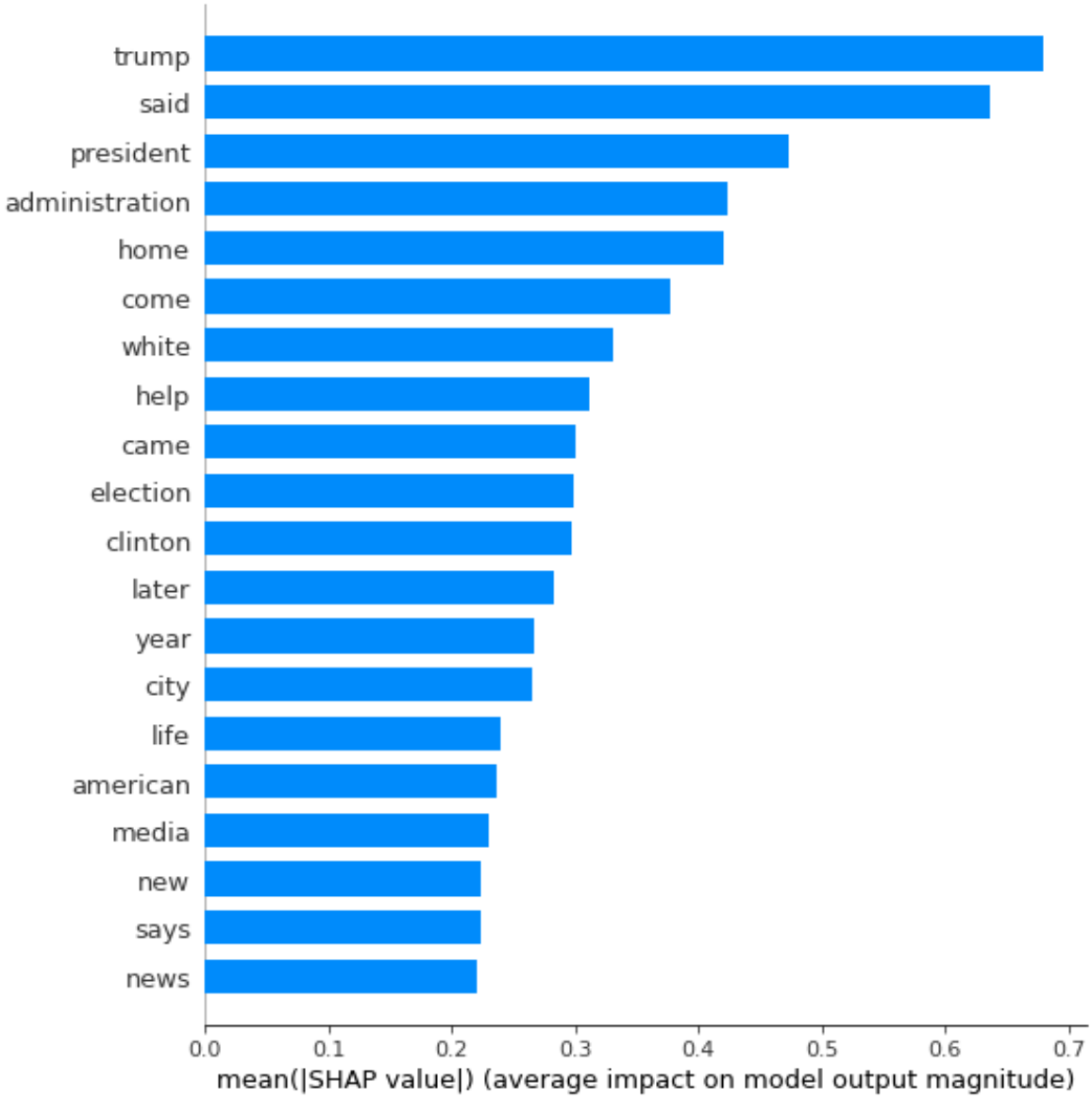

Figura 5.26: Gráfico de barras exibindo as vinte *features* mais relevantes para um modelo de classificação utilizando BoW e TFIDF.

5.8 Considerações Finais

Neste capítulo, foram apresentados os principais resultados encontrados nesta pesquisa. Os modelos construídos a partir dos LC foram comparados com modelos treinados a partir de léxicos construídos manualmente e já utilizados na literatura (i.e.*baselines*. Como principais resultados, pôde ser observado que os léxicos construídos utilizando a abordagem proposta nesta Tese permitiram a construção de modelos preditivos que, na maioria dos casos, superaram os modelos treinados a partir de léxicos construídos manualmente. Adicionalmente,

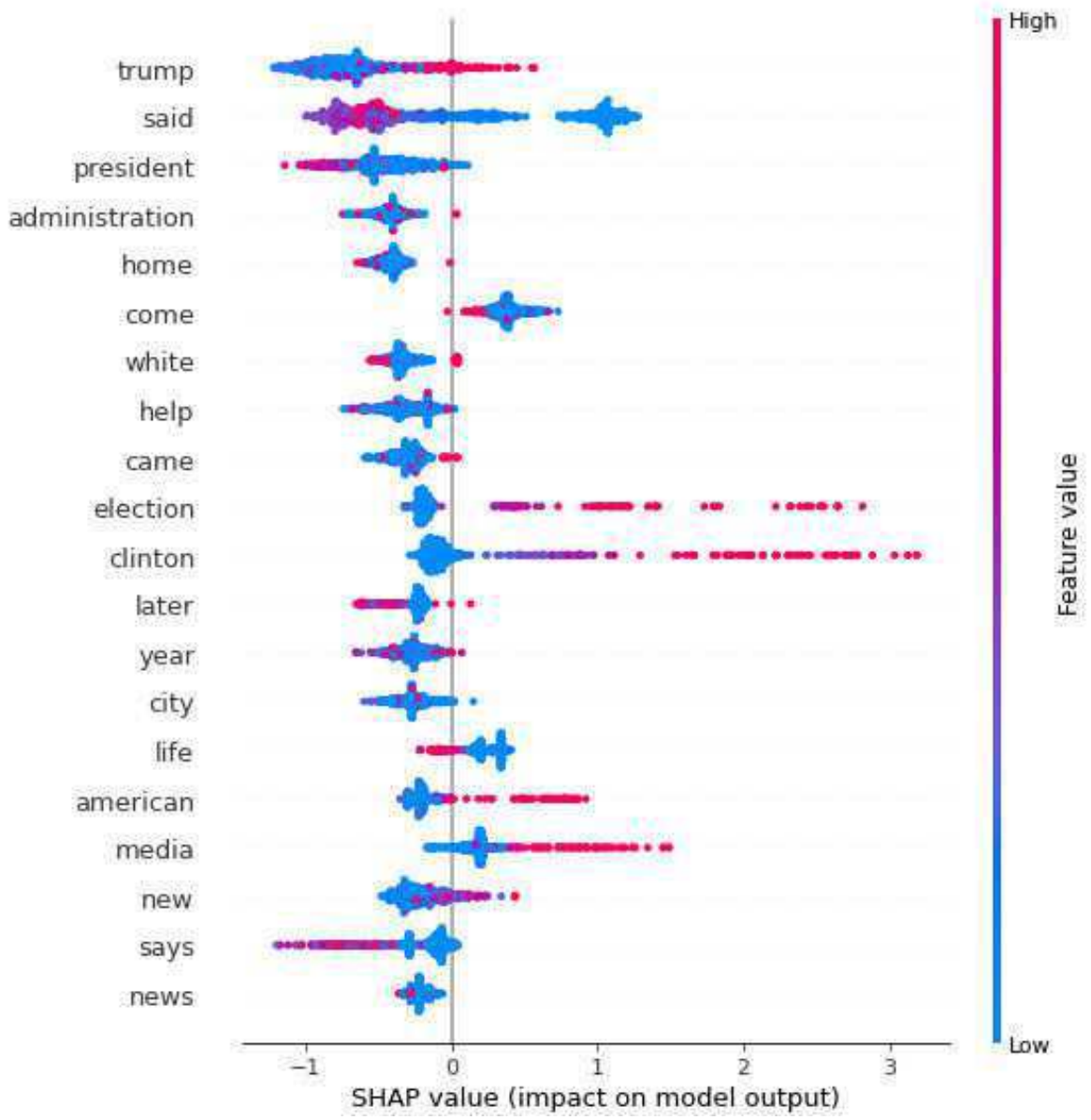

Figura 5.27: A plotagem sumário exibindo as vinte *features* mais relevantes para um modelo de classificação utilizando BoW e TFIDF. Na imagem, é possível observar como cada uma das *features* influenciam na tomada de decisão do modelo preditivo.

quando usadas em conjunto, ou seja, LC + *baselines*, os modelos puderam atingir os melhores resultados de classificação. Por fim, foi apresentada uma análise da explicabilidade de tais modelos, sendo demonstrado como essas explicações podem contribuir para um aprofundamento geral no entendimento de características ocultas nas notícias falsas.

Capítulo 6

Conclusões e Trabalhos Futuros

Este capítulo apresenta as principais conclusões que puderam ser extraídas a partir desta pesquisa, bem como aponta caminhos promissores para trabalhos futuros.

6.1 Conclusões

O uso massivo de conteúdo enganoso, em especial nas redes sociais, vem representando um grave risco que afeta as sociedades sob diversos aspectos. Esses riscos vão desde intervenções deliberados sobre processos eleitorais, à disseminação de informações falsas que podem colocar a saúde das pessoas em risco. Como resposta, o notável desenvolvimento de métodos e estratégias de identificação de notícias falsas vêm permitindo grandes avanços no combate a este tipo de conteúdo.

Neste trabalho, foram considerados três pilares de estudo: (i) construção automatizada de léxicos para o estudo e classificação de notícias falsas; (ii) construção de *features* de classificação baseadas nos léxicos gerados e (iii) estudo explicativo dos modelos preditivos construídos. Para a construção automatizada de léxicos voltados para problemas de classificação, adaptamos uma estratégia genérica de construção de léxicos já existente, tornando-a capaz de gerar léxicos que permitiram classificar notícias falsas de forma competitiva, e em muitos cenários, até mesmo superior a outros léxicos construídos manualmente por especialistas e já utilizados na literatura. A partir dos léxicos construídos, propomos também uma abordagem para a construção de *features* baseadas no conceito de similaridade semântica, onde ao invés de considerar uma palavra em si como *feature* de classificação, como normalmente

ocorre na literatura, é utilizado todo o espaço semântico existente entre os léxicos e os documentos de notícias. Este espaço semântico é provido por uma camada de *word embeddings*. Para a avaliação dos léxicos de notícias falsas gerados na pesquisa, avaliamos o desempenho de classificação de modelos treinados com base nos léxicos construídos automaticamente e léxicos já presentes na literatura. Com os resultados obtidos, foi possível observar que a construção de léxicos específicos para a classificação de notícias falsas contribuiu, de forma significativa, para uma melhoria evidente no desempenho de classificação dos modelos, em especial, quando treinados juntos com os léxicos já existentes. Este achado evidencia que aspectos intrínsecos dos documentos de notícias falsas só puderam ser revelados a partir da construção de léxicos específicos para este tipo de notícia.

O terceiro aspecto abordado nesta pesquisa, e que ainda é pouco explorado na literatura, é a explicação dos modelos de classificação construídos. Para tal estudo, foi considerado o impacto que as *features* de classificação exercem sobre os modelos preditivos. Para a análise explicativa, os léxicos gerados a partir de notícias falsas permitem revelar *insigts* que não poderiam ser notados quando utilizados léxicos construídos para outros cenários que não o de notícias falsas.

A abordagem de construção de léxicos apresentada tem como principal característica positiva, possibilitar a construção de léxicos a partir de quaisquer documentos textuais que possam ser classificados de forma binária. Estes léxicos podem ser usados como *features* de classificação, sendo assim, especialmente úteis em cenários onde a construção manual de léxicos se mostra custosa e complexa. Ainda como um possível cenário de aplicação, os léxicos construídos podem auxiliar especialistas no processo de construção manual de léxicos. Por exemplo, um léxico construído automaticamente a partir de notícias satíricas pode guiar especialistas na construção de um léxico de sátiras mais completo. Os recursos produzidos nesta pesquisa estão publicamente acessíveis[1].

6.1.1 Limitações

Algumas limitações percebidas nesta pesquisa precisam ser descritas. Dentre elas, o fato de que os léxicos gerados nesta pesquisa foram construídos a partir de um conjunto relativamente pequeno de amostras. Por exemplo, o subconjunto de notícias falsas satíricas do

[1] https://github.com/caiolibanio/lex_build

BSDetector, o qual foi utilizado para a construção do léxico de Sátiras, possui apenas 99 notícias. Dado este número reduzido de amostras para a construção desse léxico, o léxico de sátiras ficou com apenas 19 termos. Outra limitação importante é a ausência de uma validação do arcabouço de classificação proposto em um outro cenário, além do de notícias falsas. Nesta pesquisa, foram utilizados como *baselines* apenas léxicos publicamente acessíveis. Porém, léxicos como os presentes no LIWC representam um importante *baseline* a ser considerado para análise. Já no que tange os léxicos construídos nesta pesquisa, é importante destacar que eles podem, eventualmente, ficarem desatualizados ao longo do tempo, devido às mudanças de contexto das notícias. Logo, avaliações neste sentido se fazem necessárias. Outra importante limitação desta pesquisa é a ausência de avaliações utilizando notícias em Português. A limitação na quantidade de notícias falsas em português colaborou para a escolha da língua inglesa como único idioma para avaliação.

6.2 Trabalhos Futuros

Como trabalhos futuros e oportunidades de estudo diretamente relacionados a esta pesquisa, destacam-se:

1. Um grande ponto de partida para trabalhos futuros seria avaliar o método proposto considerando outros tipos de notícias falsas. Por exemplo, a construção de léxicos específicos para caracterizar notícias falsas de política, ou mesmo notícias falsas que eventualmente possam ameaçar a saúde das pessoas (e.g. criação de léxicos a partir de notícias falsas sobre COVID19);

2. Validar o método proposto considerando outros domínios além do de notícias falsas;

3. O uso de *word embeddings* criados utilizando abordagens mais recentes, como BERT (DEVLIN et al., 2018) pode contribuir para a construção de *features* de classificação mais representativas;

4. O uso de *word embeddings* voltados especificamente para notícias falsas pode também representar uma oportunidade de aprimoramento dos resultados obtidos no método, por permitir uma maior representação deste tipo de notícia, e consequentemente promover uma maior diferenciação entre notícias falsas e reais;

5. O uso do método proposto para construção de léxicos em Português também representa uma direção importante para a continuidade desta pesquisa;

6. Experimentos que permitam, por exemplo, definir parâmetros mínimos para a execução do método ainda precisam ser definidos. Como exemplo, a definição de um tamanho mínimo de *dataset* para a construção de um léxico pode ser considerado para trabalhos futuros.

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>.

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.

BHUTANI, B.; RASTOGI, N.; SEHGAL, P.; PURWAR, A. Fake news detection using sentiment analysis. In: IEEE. **2019 twelfth international conference on contemporary computing (IC3)**. [S.l.], 2019. p. 1–5.

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>.

CARVALHO, F.; OKUNO, H. Y.; BARONI, L.; GUEDES, G. A brazilian portuguese

moral foundations dictionary for fake news classification. In: **2020 39th International Conference of the Chilean Computer Science Society (SCCC)**. [S.l.: s.n.], 2020. p. 1–5.

CASTELO, S.; ALMEIDA, T.; ELGHAFARI, A.; SANTOS, A.; PHAM, K.; NAKAMURA, E.; FREIRE, J. A topic-agnostic approach for identifying fake news pages. In: **Companion proceedings of the 2019 World Wide Web conference**. [S.l.: s.n.], 2019. p. 975–980.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHEN, B.; CHEN, B.; GAO, D.; CHEN, Q.; HUO, C.; MENG, X.; REN, W.; ZHOU, Y. Transformer-based language model fine-tuning methods for covid-19 fake news detection. In: SPRINGER. **International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation**. [S.l.], 2021. p. 83–92.

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, M.; WEINBERGER, K. Q.; SHA, F. et al. An alternative text representation to tf-idf and bag-of-words. **arXiv preprint arXiv:1301.6770**, 2013.

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

CHOI, Y.; WIEBE, J. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1181–1191.

CHOUDHARY, M.; CHOUHAN, S. S.; PILLI, E. S.; VIPPARTHI, S. K. Berconvonet: A deep learning framework for fake news classification. **Applied Soft Computing**, v. 110, p. 107614, 2021. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494621005354>.

CONROY, N. K.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. **Proceedings of the association for information science and technology**, Wiley Online Library, v. 52, n. 1, p. 1–4, 2015.

DARWICH, M.; MOHD, S. A.; OMAR, N.; OSMAN, N. A. Corpus-based techniques for sentiment lexicon generation: A review. **J. Digit. Inf. Manag.**, v. 17, n. 5, p. 296, 2019.

DENG, D.; JING, L.; YU, J.; SUN, S. Sparse self-attention lstm for sentiment lexicon construction. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 27, n. 11, p. 1777–1790, 2019.

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.

DOGO, M. S.; DEEPAK, P.; JUREK-LOUGHREY, A. Exploring thematic coherence in fake news. In: SPRINGER. **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. [S.l.], 2020. p. 571–580.

FAST, E.; CHEN, B.; BERNSTEIN, M. S. Empath: Understanding topic signals in large-scale text. In: **Proceedings of the 2016 CHI conference on human factors in computing systems**. [S.l.: s.n.], 2016. p. 4647–4657.

FAUSTINI, P. H. A.; COVOES, T. F. Fake news detection in multiple platforms and languages. **Expert Systems with Applications**, Elsevier, v. 158, p. 113503, 2020.

FAUSTINI, P. H. A.; COVõES, T. F. Fake news detection in multiple platforms and languages. **Expert Systems with Applications**, v. 158, p. 113503, 2020. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417420303274>.

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>.

GONG, H.; SAKAKINI, T.; BHAT, S.; XIONG, J. Document similarity for texts of varying lengths via hidden topics. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2018. p. 2341–2351.

GRAHAM, J.; HAIDT, J.; KOLEVA, S.; MOTYL, M.; IYER, R.; WOJCIK, S. P.; DITTO, P. H. Moral foundations theory: The pragmatic validity of moral pluralism. In: **Advances in experimental social psychology**. [S.l.]: Elsevier, 2013. v. 47, p. 55–130.

GUTHRIE, L.; PUSTEJOVSKY, J.; WILKS, Y.; SLATOR, B. M. The role of lexicons in natural language processing. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 1, p. 63–72, 1996.

HAUMAHU, J.; PERMANA, S.; YADDARABULLAH, Y. Fake news classification for indonesian news using extreme gradient boosting (xgboost). In: IOP PUBLISHING. **IOP Conference Series: Materials Science and Engineering**. [S.l.], 2021. v. 1098, n. 5, p. 052081.

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>.

HUANG, S.; NIU, Z.; SHI, C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. **Knowledge-Based Systems**, v. 56, p. 191–200, 2014. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705113003596>.

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

JERONIMO, C. L.; CAMPELO, C. E.; MARINHO, L. B.; SALES, A.; VELOSO, A.; VIOLA, R. Computing with subjectivity lexicons. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. [S.l.: s.n.], 2020. p. 3272–3280.

JERONIMO, C. L. M.; MARINHO, L. B.; CAMPELO, C. E.; VELOSO, A.; MELO, A. S. da C. Fake news classification based on subjective language. In: **Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services**. [S.l.: s.n.], 2019. p. 15–24.

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.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in Neural Information Processing Systems - NIPS**, v. 30, p. 3146–3154, 2017.

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.

KOIRALA, A. Covid-19 fake news dataset. **Mendeley Data**, v. 1, 2021.

KUMAR, S.; ASTHANA, R.; UPADHYAY, S.; UPRETI, N.; AKBAR, M. Fake news detection using deep learning models: A novel approach. **Transactions on Emerging Telecommunications Technologies**, Wiley Online Library, v. 31, n. 2, p. e3767, 2020.

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>.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: XING, E. P.; JEBARA, T. (Ed.). **Proceedings of the 31st International Conference on Machine Learning**. Bejing, China: PMLR, 2014. (Proceedings of Machine Learning Research, 2), p. 1188–1196. Disponível em: <https://proceedings.mlr.press/v32/le14.html>.

LE, T.; WANG, S.; LEE, D. Malcom: Generating malicious comments to attack neural fake news detection models. In: IEEE. **2020 IEEE International Conference on Data Mining (ICDM)**. [S.l.], 2020. p. 282–291.

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>.

LI, X.; XIA, Y.; LONG, X.; LI, Z.; LI, S. Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english. **arXiv preprint arXiv:2101.02359**, 2021.

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.

MERTOĞLU, U.; GENÇ, B. Lexicon generation for detecting fake news. **arXiv preprint arXiv:2010.11089**, 2020.

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

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.

MONTI, F.; FRASCA, F.; EYNARD, D.; MANNION, D.; BRONSTEIN, M. M. Fake news detection on social media using geometric deep learning. **arXiv preprint arXiv:1902.06673**, 2019.

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.

OSHIKAWA, R.; QIAN, J.; WANG, W. Y. A survey on natural language processing for fake news detection. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 6086–6093. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.747>.

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.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. Supervised learning for fake news detection. **IEEE Intelligent Systems**, IEEE, v. 34, n. 2, p. 76–81, 2019.

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.

SALES, A.; BALBY, L.; VELOSO, A. Media bias characterization in brazilian presidential elections. In: **Proceedings of the 30th ACM Conference on Hypertext and Social Media**. [S.l.: s.n.], 2019. p. 231–240.

SCHWARZ, S.; THEÓPHILO, A.; ROCHA, A. Emet: Embeddings from multilingual-encoder transformer for fake news detection. In: IEEE. **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.], 2020. p. 2777–2781.

SHARMA, K.; QIAN, F.; JIANG, H.; RUCHANSKY, N.; ZHANG, M.; LIU, Y. Combating fake news: A survey on identification and mitigation techniques. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 10, n. 3, p. 1–42, 2019.

SHU, K.; MAHUDESWARAN, D.; WANG, S.; LEE, D.; LIU, H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. **Big data**, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 8, n. 3, p. 171–188, 2020.

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.

TAI, Y.-J.; KAO, H.-Y. Automatic domain-specific sentiment lexicon generation with label propagation. In: **Proceedings of International Conference on Information Integration and Web-based Applications & Services**. [S.l.: s.n.], 2013. p. 53–62.

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.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2017. p. 5998–6008.

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, L.; WANG, Y.; MELO, G. de; WEIKUM, G. Understanding archetypes of fake news via fine-grained classification. **Social Network Analysis and Mining**, Springer, v. 9, n. 1, p. 1–17, 2019.

WANG, L.; XIA, R. Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 502–510. Disponível em: <https://aclanthology.org/D17-1052>.

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.

WANI, A.; JOSHI, I.; KHANDVE, S.; WAGH, V.; JOSHI, R. Evaluating deep learning approaches for covid19 fake news detection. In: SPRINGER. **International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation**. [S.l.], 2021. p. 153–163.

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>.

WYNNE, H. E.; WINT, Z. Z. Content based fake news detection using n-gram models. In: **Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services**. [S.l.: s.n.], 2019. p. 669–673.

ZELLERS, R.; HOLTZMAN, A.; RASHKIN, H.; BISK, Y.; FARHADI, A.; ROESNER, F.; CHOI, Y. Defending against neural fake news. **arXiv preprint arXiv:1905.12616**, 2019.

ZHANG, X.; GHORBANI, A. A. An overview of online fake news: Characterization, detection, and discussion. **Information Processing & Management**, v. 57, n. 2, p. 102025, 2020. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457318306794>.

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.; JAIN, A.; PHOHA, V. V.; ZAFARANI, R. Fake news early detection: A

theory-driven model. **Digital Threats: Research and Practice**, ACM New York, NY, USA, v. 1, n. 2, p. 1–25, 2020.

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

Apêndice A

Léxicos Gerados pela Pesquisa

Esta seção apresenta os léxicos gerados na pesquisa, referenciados ao longo do texto desta tese como LG. Ao todo, foram construídos seis léxicos para classificar e analisar notícias falsas. Estes léxicos foram construídos a partir do conjunto de noticias falsas BSDetector.

- **Léxico de Viés (*Bias*)**: investigation, season, calling, wants, big, republicans, archive, enough, october, tuesday, gop, including, part, voting, black, campaigning, wnd, yearold, united, presidentelect, me, another, cover, death, talk, years, little, rally, economic, gold, went, believe, role, month, min, near, give, present, using, sessions, fraud, takes, rep, building;

- **Léxico de Conspiração (*Conspiracy*)**: voter, daily, white, now, however, rt, radio, head, sexual, infowars, sunday, legal, actually, executive, daughter, took, foundation, wikileaks, image, wire, nominee, attacks, video, good, look, held, left, job, fall, middle, looked, result, nothing, children, child, act, event, old, with, described, six, growth, makes, study, flickr, choice, reasons, muslim, swing, aide, modern, sitting, played, thought, noted, paid, wife;

- **Léxico de Ódio (*Hate*)**: greenfield, dr, men, popular, that, book, written, well, statement, college, later, police, says, say, change, kings, hit, court, interview, known, threat, around, recently, share, vote, called, three, enjoys, south, radical, international, train, jewish, outcome, research, started;

- **Léxico de pseudo-ciência (*Junk-Science*)**: home, today, help, scientific, presidential,

great, chemical, effective, weight, form, real, president, things, remedies, we, twain, leading, play, treat, campaign, others, parents, system, protests;

- **Léxico de Sátiras (*Satire*)**: seen, published, time, story, told, got, posted, pretty, first, election, government, full, he, days, ago, always, think, woman, work;

- **Léxico de Estados Repressores (*State*)**: tv, reuters, bulletin, politics, ap, file, army, national, foreign, capital, indian, northern, yearold, party, fighters, issues, monday, took, several, match, control, emergency, protest, region, head, building, better, movement;