DETECÇÃO AUTOMÁTICA DE FAKE NEWS: UMA ABORDAGEM BASEADA EM DICIONÁRIOS LINGUÍSTICOS

Arthur Gusmão de Matheus & Leandro Nunes de Castro

Apoio: PIBITI CNPq

RESUMO

Nesse artigo é descrita uma aplicação desenvolvida para a identificação de fake news usando uma abordagem baseada em mineração de textos e dicionários sintáticos. A solução conta com dois dicionários de palavras “positivas” e “negativas” para a categorização dos termos de um conjunto de textos representando notícias a serem classificadas em *True News* (TN) ou *Fake News* (FN). A construção da solução foi feita usando uma plataforma analítica *no-code*, filtros e funções do Excel para classificar se uma notícia é TN ou FN. Vale ressaltar que o presente projeto pode ser customizado conforme as necessidades do usuário como, por exemplo, teste de bases em outros idiomas. As etapas de desenvolvimento compreenderam métodos e técnicas algoritmicas focadas na tomada de decisão que tem por base padrões encontrados em textos em língua natural.

Palavras-chave: Identificação de Fake News; Mineração de Textos; Categorização; Léxicos; Processamento de Língua Natural.

ABSTRACT

This paper describes an application developed for identifying fake news using an approach based on text mining and syntax dictionaries. The solution has two dictionaries of “positive” and “negative” words for categorizing the terms of a set of texts representing news to be classified into True News (TN) or Fake News (FN). The construction of the solution was done using a no-code analytics platform, filters and functions in Excel to classify whether a news item is TN or FN. This project can be customized according to the user's needs, such as testing bases in other languages. The development stages comprised algorithmic methods and techniques focused on decision making based on patterns found in natural language texts.

Keywords: Identification of Fake News; Text Mining; Categorization; Lexicons; Natural Language.

1. INTRODUÇÃO

Dado o atual cenário onde qualquer um ao navegar pela rede passa a ser vulnerável às chamadas *fake News*, ou seja, notícias falsas divulgadas na internet e que podem representar algum tipo de malefício à sociedade, buscou-se por reunir dados e soluções computacionais capazes de fazer frente ao seguinte problema de pesquisa: *Construir uma ferramenta computacional capaz de detectar, automaticamente, fake news em um texto através de aplicações baseadas em mineração de textos.*

As notícias falsas espalhadas na internet, principalmente em redes sociais, podem ser usadas para criar boatos ou reforçar um pensamento já existente, por meio de acusações falsas e da disseminação de ódio contra pessoas comuns, celebridades, organizações, políticos, etc. Sendo assim, o presente trabalho apresenta uma ferramenta de detecção de *Fake News* utilizando técnicas de Mineração de Textos e Processamento de Língua Natural (PLN).

O artigo base que deu origem ao método desenvolvido aqui foi o trabalho de Larcker et al. (2013). Em suma, o artigo trata da busca por padrões na linguagem de executivos em teleconferências trimestrais, que levem à identificação de discussões enganosas através de certas categorias de palavras mais utilizadas pelos mentirosos. No contexto de *fake News*, foi possível gerar dois dicionários diferentes contendo milhares de palavras classificadas em “positivas” e “negativas”, onde as palavras positivas estão relacionadas às *fake News* (notícias falsas) e as negativas às *true News* (notícias verdadeiras).

Assim sendo, através da aplicação desenvolvida no presente projeto baseada em algoritmos de mineração de texto, o usuário será capaz de realizar o *download* de dois arquivos em formato *.csv*, onde um mostrará todas as palavras de categoria positivas encontradas em sua base de dados de testes e o outro contendo todas as palavras negativas. Ou seja, através de uma análise das somatórias de todas as palavras encontradas, de ambas as categorias, será possível dizer o quanto tal documento pode ser uma notícia falsa ou uma notícia verdadeira.

O trabalho está organizado da seguinte forma. A Seção 2 apresenta o referencial teórico da pesquisa, descrevendo brevemente o que são *fake News* e quais as principais etapas de análise de textos a serem usadas na pesquisa. Na Seção 3 o método proposto é apresentado e na Seção 4 a metodologia experimental, resultados e discussão são apresentados. O trabalho é concluído na Seção 5 com uma discussão geral e perspectivas de trabalhos futuros.

2. REFERENCIAL TEÓRICO

2.1 Fake News

Há cerca de uma década a comunicação era feita de “poucos para muitos”, onde a sociedade dependia do conteúdo que poucos grupos produziam e propagavam. Contudo, com o recente advento da internet e, principalmente, das redes sociais esse cenário mudou drasticamente. Cada vez mais pessoas estão conectadas à rede, através da qual acessam suas redes sociais.

Segundo um levantamento divulgado com exclusividade na revista Forbes e realizado pela App Annie, agência focada em análise do mercado *mobile*, a média diária gasta em uso de aplicativos, somente no Brasil e feita com base nos resultados do segundo trimestre de 2021, foi de 5.4 horas. Além disso, o estudo contou com um ranking elaborado com base nos aplicativos mais baixados no mundo, onde TikTok, Facebook e WhatsApp encontravam-se entre os dominantes na lista[1].

Portanto, através da internet e das redes sociais atualmente assiste-se a um crescimento do domínio do modelo de comunicação de muitos para muitos em uma escala sem precedentes. Ou seja, cada vez menos pessoas estão a depender de informações produzidas e divulgadas pelos grandes grupos, o que, novamente, mostra a procura absurda de informações criadas, de forma virtual e por qualquer indivíduo, dentro da própria rede. Se, por um lado, se democratizou a geração e o compartilhamento de conteúdo, por outro vimos, também em grande proporção, a explosão da desinformação online (Del Vicario et al., 2016; Lazer et al., 2017; McCright & Dunlap, 2017).

Fake News é um termo do inglês que é usado para referir-se a falsas informações divulgadas, principalmente, em redes sociais e que vem ganhando cada vez mais repercussão ao passo que cada vez mais pessoas estão conectadas à rede. A todo momento o indivíduo é bombardeado por uma “avalanche” de informações, verdadeiras ou falsas, seja rolando comentários nas redes sociais, compartilhando links, ou consumindo materiais audiovisuais dispersos no YouTube (Google™), plataforma com mais de 1 (um) bilhão de horas de vídeos consumidos diariamente por usuários de todo o mundo segundo o *Wall Street Journal*.[2] Todavia, esse termo popularizou-se muito mais e veio à tona durante as eleições de 2016 nos Estados Unidos, na qual Donal Trump tornou-se presidente. Na época em que o ex-presidente

[1] https://olhardigital.com.br/2021/07/17/internet-e-redes-sociais/brasil-e-o-pais-que-passa-mais-tempo-em-aplicativos/

[2] https://www1.folha.uol.com.br/tec/2017/02/1862377-usuarios-passam-1-bilhao-de-horas-por-dia-no-youtube-diz-jornal.shtml

dos EUA havia ganho, algumas empresas especializadas encontraram uma série de websites contendo conteúdos um tanto quanto duvidosos, onde a maioria das notícias divulgadas chamavam para tópicos sensacionalistas, envolvendo algumas vezes personalidades importantes, como a então adversária Hillary Clinton, por exemplo. Porém, além de serem usadas no meio político para prejudicar ou favorecer algum candidato em específico, como também testemunhado no Brasil durante a corrida presidencial de 2018, as *Fake News* podem pairar sobre qualquer pessoa, organização ou instituição em manchetes sensacionalistas, muitas vezes divulgadas em redes sociais por "robôs"[3].

Allcott & Gentzkow (2017) definem as Fake News como sendo o tipo de notícia que é propositalmente criada com a intenção de ser falsa e com o objetivo de enganar e/ou manipular os leitores. Se tratando, portanto, de um modelo de desinformação, as *Fake News* são notícias intencionalmente e verificavelmente falsas (Shu et al., 2017). Tal definição exclui erros jornalísticos não intencionais; rumores, ou seja, informações que não são verificadas no momento da postagem (Zubiaga et al., 2018); teorias de conspiração, entendidas como explicações sobre eventos históricos em termos do agente causal de um grupo relativamente pequeno de pessoas agindo em segredo (Keeley, 1999); sátiras e fofocas.

2.2 Análise de Textos

A *análise de textos* é um campo multidiciplinar responsável por tomar decisões beaseadas em padrões e regularidades extraídas de textos em língua natural. Ela pode, portanto, requerer conhecimentos de diversas áreas, como Recuperação de Informação (Baeza-Yates & Ribeiro-Neto, 2011), Estatística (Triola, 2017), Mineração de Dados (de Castro & Ferrari, 2016); Mineração de Textos (Weiss et al., 2005; Ignataw & Mihalcea, 2017), Linguística e Processamento de Língua Natural (Jurafsky & Martin, 2008; Goldberg & Hirst, 2017). Ou seja, a mineração de textos busca analisar computacionalmente padrões e regularidades encontradas em textos para processos de tomada de decisão e, para tanto, atribui uma identidade para cada termo de um documento previamente estruturado.

Portanto, faz-se necessária uma breve introdução ao tema com base no clássico modelo de *espaço vetorial*, o qual requer o pré-processamento dos documentos usando algumas técnicas como:

1. Análise léxica ou *tokenização* do texto: primeiro passo para a manipulação de textos e responsável pela conversão dos documentos (conjunto de caracteres) em conjuntos de

[3] https://mundoeducacao.uol.com.br/curiosidades/fake-news.htm

palavras conhecidas como termos (tokens), o que normalmente é feito através da escolha de algumas delimitações, como as palavras, espaços em branco, pontuações, etc.;

2. Remoção de palavras frequentes (stopwords): *stopword* é o nome dado às palavras que são filtradas antes ou depois do processamento do texto. Elas podem ser entendidas como uma espécie de ruído na informação, dificultando a habilidade de identificar rapidamente a relevância do resultado da busca ou o significado e importância das palavras em um documento. A filtragem das stoipwords é geralmente feita por uma lista de *stopwords* (p. ex., artigos, preposições e conjunções) e permite que o documento se torne mais claro, útil e de mais fácil análise.

Uma vez estruturado o documento, torna-se possível a utilização de técnicas convencionais de aprendizagem de máquina para a análise. Em princípio, qualquer algoritmo de análise de dados pode ser usado para extrair informações dos textos estruturados. Por exemplo, classificadores como Naive Bayes, K-NN, máquinas de vetores suporte (SVMs), árvores de decisão, redes neurais e muitos outros (de Castro & Ferrari, 2016), podem ser aplicados para classificar os documentos em *Fake* ou *True News*.

3. UM MÉTODO BASEADO EM LÉXICO PARA DETECÇÃO DE FAKE NEWS

A detecção de fake news consiste no processo de captura, pré-processamento, análise e validação de dados, a fim de predizer se uma determinada notícia é verdadeira (*true news*) ou falsa (*fake news*). Contudo, a essas etapas compreendem-se métodos e técnicas algoritmicas focadas na tomada de decisão que tem por base padrões encontrados em textos em língua natural. Ou seja, pode-se trabalhar em cima de uma solução para identificação de fake news baseadas em conteúdo, na fonte da informação e na difusão da informação (Figueira & Oliveira, 2017). Os métodos de detecção de *fake News* podem ser divididos em linguísticos, baseados em rede, ou baseados em mineração de dados (Conroy et al. 2017; Shu et al. 2017).

Tanto em notícias verdadeiras, quanto em falsas, existem o autor, o difusor, o receptor e o conteúdo. O primeiro é o sujeito que criou a mensagem e que possuí características como idade e gênero, o difusor é quem a compartilhou, o receptor quem a recebeu e o conteúdo é o conjunto de características que compõem a mensagem, como o texto, cabeçalho, imagens etc. Entretanto, cada notícia possui um conjunto de engajamentos sociais que ilustram sua difusão com o tempo.

Na abordagem a ser utilizada nessa pesquisa, o processo de identificação de *fake news* se dará, basicamente, em cima de dois processos centrais: a estruturação dos documentos como etapa principal do pré-processamento e a análise dos dados estruturados. Aqui,

a “mineração” se dará em cima das notícias em análise e o conhecimento a ser obtido será a classificação delas em verdadeiras ou falsas. O método a ser utilizado foi baseado no trabalho de Larcker et al. (2013) e implementado na plataforma analítica *SOMMA.ai*, conforme ilustrado na Figura 1. A essência deste método reside na criação de um dicionário de palavras características de *fake news* (Dic_Positivo) e um dicionário de palavras características de *true news* (Dic_Negativo), seguida da verificação da existência e contagem dessas palavras nos documentos. Ao final do processo é atribuída a classe de acordo com a predominância das palavras de cada classe (*fake* ou *true*) no documento.

O projeto basicamente se desenvolveu em cima de componentes específicos da plataforma para processameto de texto (*Text Processing*), união e cruzamento de dados (*Join*), agrupamento de dados (*Group Join*) e consultas SQL (*query’s* SQL). Toda a criação e integração entre essas operações foi feita usando a *SOMMA.ai*, uma plataforma online capaz de trabalhar com as mais variadas técnicas de mineração de dados e de texto de forma simplificada e *no-code*. Ou seja, ela possibilita que um usuário comum, sem grandes conhecimentos no campo da inteligência artificial, seja capaz de criar suas aplicações de *machine learning* sem precisar utilizar de linguagens de programação.

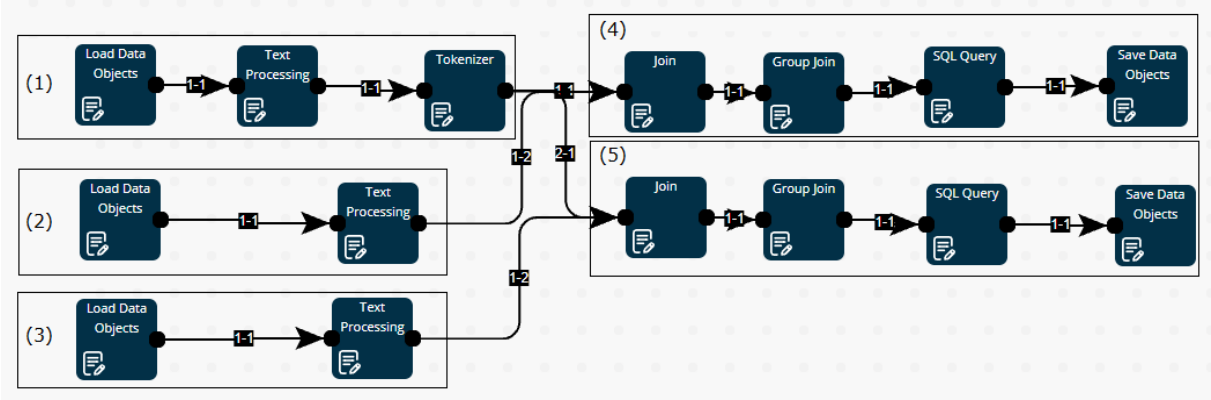

Figura 1: Fluxo da Aplicação construída na SOMMA para classificar notícias em verdadeiras (true News) ou falsas (fake News).

A aplicação funciona da seguinte forma. O bloco (1) da aplicação permite a entrada (*Load Data Objects*) das notícias a serem analisadas, realiza uma padronização (remove acentos, coloca tudo em minúsculo e remove pontuações) (*Text Processing*), e toqueniza (*Tokenizer*) o documento. Nos blocos (2) e (3) temos a padronização dos textos contidos nos dicionários de palavras positivas e negativas, respectivamente. O bloco (4) verifica a interseção entre as palavras dos documentos e o dicionário de palavras positivas (*Join+ Group Join*), enquanto o bloco (5) faz a interseção (*Join+ Group Join*) com o dicionário negativo.

Além disso, esses blocos realizam a contagem das palavras de cada classe e gravam um arquivo na saída (*Save Data Objects*) com o resultado do processo.

4. METODOLOGIA EXPERIMENTAL

Para a classificação automática de *fake* e *true news*, o projeto utilizou uma base de dados da literatura e dois dicionários que serviram para a categorização das palavras encontradas em *positivas* (*fake news*) ou *negativas* (*true news*). A base utilizada, no presente projeto, está disponível para download gratuito no Kaggle[4] e trata de um conjunto de textos e metadados retirados de 244 sites, envolvendo 12.999 postagens. Os dados foram extraídos usando a API webhose.io e os sites rotulados de acordo com a extensão *BS Detector*, um plug-in que usa uma lista de fontes de notícias falsas como referência. Ele pode ser adicionado aos navegadores Chrome e Mozilla e emite um aviso de *"este site é considerado uma fonte* questionável" para quando encontrar alguma notícia potencialmente falsa.[5] Já os dicionários para a categorização das palavras da base de dados em “positivas” (*Fake News*) e “negativas” (*True News*), foram construídos com base no artigo de Larcker & Zakolyunkina (2012), o qual basicamente informa quais classes de palavras pertencem a cada categoria. Para a construção dos dicionários de palavras positivas e negativas o modelo proposto por Larcker & Zakolyunkina (2012) faz a categorização resumida na *Tabela* 1.

Tabela 1: Categorias de palavras positivas e negativas do modelo de Larcker & Zakolyunkina (2012).

Negativos	Positivos
1st person singular pronouns	1st person plural pronouns
Assent	Negations
Nonextreme positive emotions	Anxiety
Certainty	Anger
	Swear words
	Extreme negative emotions
	Tentative

O desempenho e a efetividade da aplicação será medida pela acurácia preditiva das classificações obtidas pelo método em comparação com as classificações (rótulos) disponíveis na própria base de dados. No método proposto a classificação da mensagem em positiva ou negativa é feita considerando a soma das palavras de cada dicionário encontradas no documento, sendo que a maior soma determina a classe da mensagem. Também será

[4] Getting Real about Fake News | Kaggle
[5] https://www.bbc.com/news/technology-38181158

construída a matriz de confusão dos resultados para avaliarmos a *taxa de verdadeiros positivos* (TVP), que corresponde ao percentual de objetos positivos classificados corretamente, e a *taxa de falsos positivos* (TFP).

5. RESULTADOS E DISCUSSÃO

Foi extraída uma amostra contendo 1000 objetos da base de dados do *BS Detector* e gerados dois dicionários para a categorização das palavras em positivas e negativas. Assim, foi possível dar início aos testes com a aplicação de identificação de *fake News*, conforme mostrado nos passos abaixo:

a) Após executar a aplicação são gerados dois arquivos, cada um com a contagem da quantidade de palavras positivas e negativas de cada texto de acordo com os dicionários utilizados (*Figura* 2);

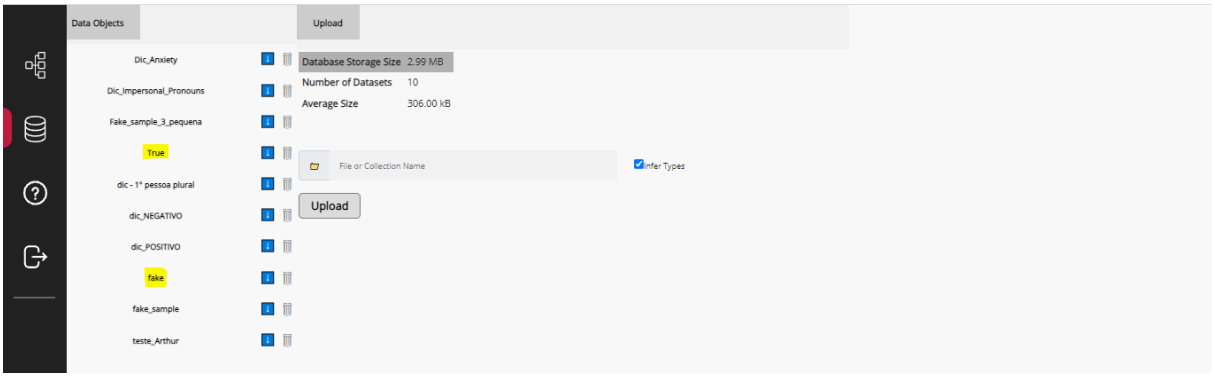

Figura 2: Exportação dos dicionários de palavras positivas e negativas na plataforma.

b) Após o download dos arquivos, foi criada uma pasta na área de trabalho para o armazenamento de ambas as bases geradas pela SOMMA juntamente com uma planilha do Microsoft Excel, configurada através das funções PROCV e SE para a classificação em FN ou TN. A base contendo as palavras negativas foi renomeada para “DTrueNews” e a base contendo as palavras positivas foi renomeada para “DFakeNews” (*Figura 3*);

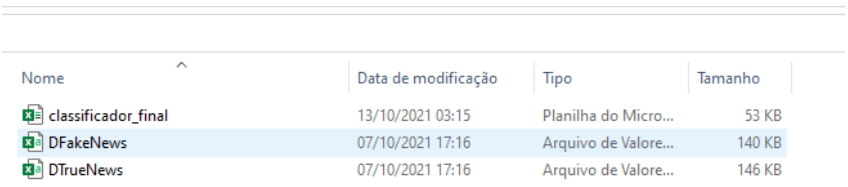

Figura 3: *Planilha de classificação e dicionários organizados em um diretório único.*

c) Em seguida foi necessário formatar os arquivos das bases de dados geradas pela plataforma (originalmente em formato .csv) (*Figura* 4) e gerar uma matriz com a quantidade de termos positivos e negativos de cada texto, permitindo a classificação automática dos documentos e comparação com os rótulos originais.

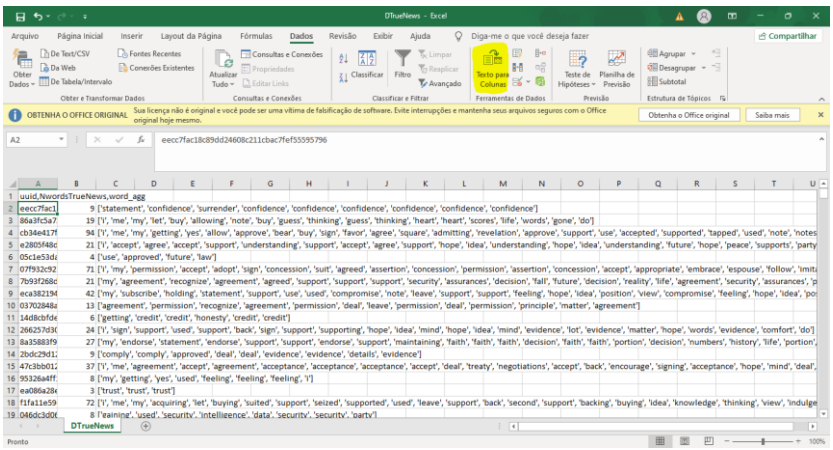

Figura 4:Dicionário contendo o total de palavras de cada classe encontradas na base de dados.

d) Avaliação de desempenho. A Figura 5 ilustra a distribuição de frequência da base de dados original, que possui os seguintes rótulos: *bs* (não rotulado); *bias* (viés de confirmação), conspiracy (teoria da conspiração); e fake (*fake news*).

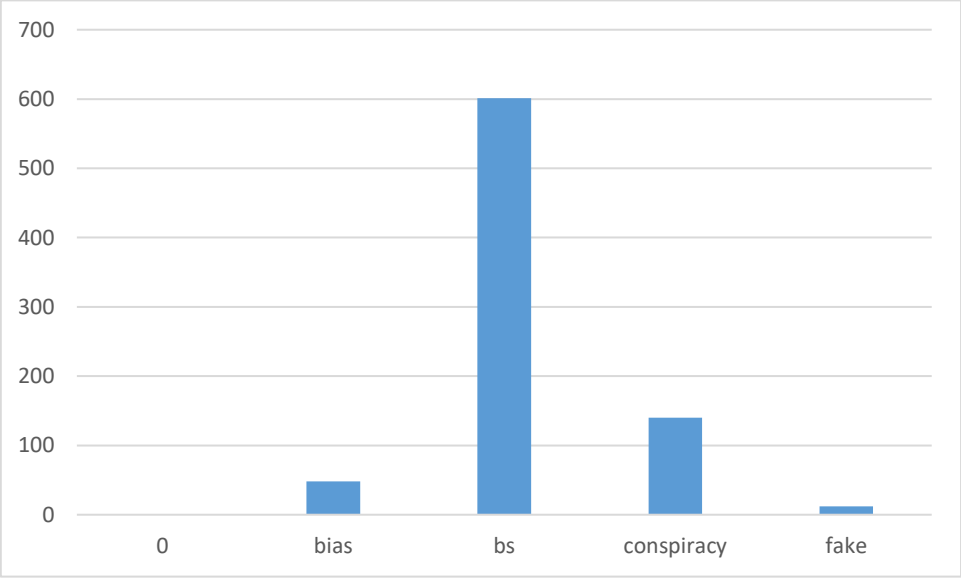

Figura 5: Distribuição de frequência dos rótulos na base original: bs (ausência de rótulo); bias (viés de confirmação); conspiracy (teoria da conspiração); fake (fake news).

Após realizar as etapas anteriores os seguintes resultados foram obtidos:

- Quantidade de mensagens classificadas em TN ou FN: 802. Portanto, 199 mensagens receberam o rótulo *bs* (não rotulado), seguindo o mesmo padrão da base original.
- Quantidade de mensagens classificadas como TN: 413.
- Quantidade de mensagens classificadas como FN: 389.
- Taxa de Verdadeiros Positivos (TVP): 25%.
- Taxa de Falsos Negativos (TFN): 75%.

Como podemos observar nos resultados apresentados, o algoritmo teve desempenho muito baixo, apresentando uma alta TFN e uma baixa TVP. Isso sugere alguns caminhos possíveis de pesquisa, como a melhoria dos dicionários para um melhor poder discriminatório, uma análise mais detalhada dos rótulos da base original e a utilização de métodos baseados em aprendizagem para a classificação das mensagens.

6. CONSIDERAÇÕES FINAIS

Ao passo em que a disseminação da informação aumenta entra a sociedade por meios digitais, também aumenta a ocorrência das chamadas *fake news* entre aqueles que sempre estão a consumir algum conteúdo gerado e dipersado diariamente dentro da própria rede. Portanto, partindo desse problema o atual projeto busca ajudar o indivíduo a saber se determinada notícia pode ou não ser considerada uma FN ou uma TN.

A aplicação aqui construída fornece uma solução simples e personalizável de identificação de notícias falsas. Isto é, através de léxicos e dicionários disponíveis para *download* na internet, o usuário consegue moldar a aplicação para as suas necessidades como, por exemplo, testá-la em um outro idioma diferente do inglês. Aqui, os dicionários construídos permitiram testar a aplicação e foram produzidos de acordo com as categorias de palavras informadas pelo de Larcker & Zakolyunkina (2012), no qual constavam os diferentes tipos de palavras que mais costumam aparecer em discursos que, comprovadamente, possuíam alguma informação falsa. Devido a questões de praticidade, a construção aqui deu-se através do idioma inglês e se decidiu por utilizar apenas um dicionário para cada uma das categorias de “positivo” e “negativo” por assim ser uma maneira mais simples de se tratar os dados.

Todavia, se o usuário quiser enriquecer, e/ou criar, os seus dicionários de categorização da sua base de dados, basta procurar na rede pelo “LIWC” que melhor atenderá às suas necessidades. Por exemplo, caso o mesmo opte por configurar a aplicação para que ela trabalhe com o português, será necessária a utilização do LIWC do mesmo idioma, que encontra-se disponível em http://143.107.183.175:21380/portlex/index.php/en/projetos/liwc.

Porém, caso queira aprimorar ainda mais os dicionários desenvolvidos para a presente aplicação, basta ele adiquirir o seu LIWC 2015, pago e em inglês, no seguinte link: http://liwc.wpengine.com.

7. REFERÊNCIAS

[1] Alott, H.; Gentzkow, M (2017), Social media and Fake News in the 2016 Election, Journal of Economic Perspectives, 31(2), pp. 211-236.

[2] Baeza-Yates, R.; Ribeiro-Neto, B. (2011), Modern Information Retrieval, 2a. Edição, Pearson.

[3] Conroy, N. J.; Rubin, V. L.; Chen, Y. (2015), Automatic deception detection: Methods for finding Fake News, Proc. of the Association for Information Science and Technology, 52(1), pp. 1-4.

[4] Figueira, A.; Oliveira, L. (2017), The current state of the fake news: Challenges and opportunities, Procedia Computer Science, 121(2017), pp. 817-825.

[5] Goldberg, Y.; Hirst, G. (2017), Neural Network Methods in Natural Language Processing, Morgan & Claypool Publishers.

[6] Ignataw, G.; Mihalcea, R. F. (2017), An Introduction to Text Mining: Research Design, Data Collection, and Analysis, Sage Publications.

[7] Jurafsky, D.; Martin, J. H. (2008), Speech and Language Processing, Prentice Hall.

[8] Keeley, B. L. (1999), Of Conspiracy Theories, The Journal of Philosophy, 96(3), pp. 109-126.

[9] Larcker, D. F.; Zakolyunkina, A. A. (2012), Detecting Deceptive Discussions in Conference Calls, Journal of Accounting Research, 50(2), pp. 495-540.

[10] Lazer, D.; Baum, M.; Grinberg, N.; Friedland, L.; Joseph, K.; Hobbs, W.; Mattsson, C. (2017), Combating Fake News: An agenda for research and action, Conference Report, p. 19.

[11] Lima, A. C. E. S.; de Castro, L. N.; Corchado, J. M. (2015), A polarity analysis framework for Twitter messages, Applied Mathematics and Computation, 270, pp. 756-767.

[12] McCright, A. M.; Dunlap, R. E. (2017), Combating misinformation requires recognizing its types and factors that facilitate its spread and resonance, Journal of Applied Research in Memory and Cognition, Elsevier, 6(2017), pp. 389-396.

[13] Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. (2017), Fake News Detection on Social Media: A Data Mining Perspective, ACM SIGKDD Explorations Newsletter, 19(1), pp. 22-36.

[14] Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. (2017), Fake News Detection on Social Media: A Data Mining Perspective, ACM SIGKDD Explorations Newsletter, 19(1), pp. 22-36.

[15] Triola, M. F. (2017), Elementary Statistics, 13a. Edição, Pearson.

[16] Weiss, S. M.; Indurkhya, N.; Zhang, T.; Damerau, F. (2005), Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer.

[17] Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; Procter, R. (2018), Detection and Resolution of Rumours in Social Media: A Survey, ACM Computing Surveys 51(2), Article 32, p. 36.