Um Método Linguístico que combina Polaridade, Emoção e Aspectos Gramaticais para Detecção de Fake News em Inglês

Gustavo A. Testoni[1], **Marcelo P. Souza**[1], **Paulo Márcio S. Freire**[1], **Ronaldo R. Goldschimidt**[1]

[1]Seção de Engenharia de Computação – Instituto Militar de Engenharia (IME)
Rio de Janeiro – RJ – Brazil

{gustavo.testoni,mpsouza,paulomsfreire,ronaldo.rgold}@ime.eb.br

Abstract. *The growing use of digital media has enhanced Fake News proliferation. Linguistic-based methods that consider sentiment (i.e. polarity or emotions) present in news' texts have shown promising results to detect this kind of news in digital media automatically. Although emotions in texts can provide rich information to identify Fake News, those methods were restricted to polarity extraction when applied to news written in English. Thus, the present work proposes a linguistic-based prototype that considers both polarity and emotions to detect English-written Fake News. The proposed prototype was adapted from a successful sentiment-based method that was conceived and applied to news written in Portuguese. The proposed prototype showed promising results in the experiments, overcoming baselines by up to 3.59 percentage points.*

Resumo. *O uso crescente dos meios digitais aumentou a proliferação de Fake News. Métodos baseados na abordagem linguística que consideram o sentimento (i.e. polaridade ou emoções) presente nos textos das notícias têm mostrado resultados promissores para detectar, automaticamente, esse tipo de notícia nos meios digitais. Embora as emoções nos textos possam fornecer informações significativas para identificar Fake News, esses métodos eram restritos à extração de polaridade quando aplicados a notícias escritas em inglês. Assim, o presente trabalho propõe um protótipo baseado na abordagem linguística que considera a polaridade e as emoções para detectar Fake News escritas em inglês. O protótipo proposto foi adaptado de um método bem-sucedido, baseado no sentimento, que foi concebido e aplicado a notícias escritas em português. O protótipo proposto apresentou resultados promissores nos experimentos, superando os baselines em até 3,59 pontos percentuais.*

1. Introdução

Os meios digitais de divulgação de notícias (*MDDN*), compostos, basicamente, pelas mídias virtuais (ex: jornais on-line) e redes sociais, vêm, a cada dia, aumentando o consumo de notícias *on-line* [Vosoughi et al. 2017]. Entretanto, alguns *MDDN*, apesar de facilitarem o acesso às notícias, permitem que qualquer pessoa, independentemente de sua credibilidade, divulgue notícias com intenso poder de propagação [Shu et al. 2017]. Tal permissividade amplificou a disseminação de *Fake News*, um tipo particular de notícia falsa cuja divulgação acontece de forma intencional [Freire and Goldschmidt 2019].

A desinformação causada pelas *Fake News* é uma preocupação mundial, pois tem o potencial de prejudicar comunidades ou pessoas, criar caos, gerar prejuízos financeiros

ou vantagens políticas [Freire and Goldschmidt 2019]. Como um exemplo atual, pode-se enfatizar o caso da pandemia de COVID-19 que já matou mais de um milhão de pessoas no ano de 2020 ao redor do mundo[1], onde inúmeras *Fake News* têm sido divulgadas em *MDDN* [Mejova and Kalimeri 2020].

Em busca de mitigar os efeitos nocivos das *Fake News*, faz-se necessária a utilização de abordagens computacionais para a sua detecção. A automatização é importante nesta tarefa devido ao grande volume e a rápida velocidade de propagação destas publicações nocivas nos *MDDN* [Vosoughi et al. 2017].

Dentre as abordagens para a detecção automatizada de *Fake News* destaca-se a abordagem linguística, que utiliza as informações extraídas diretamente dos textos das notícias [Bondielli and Marcelloni 2019]. A literatura tem apresentado métodos promissores, baseados neste tipo de abordagem, que utilizam a classificação gramatical (i.e. rotulação de adjetivos, verbos e etc.) associada à análise de sentimentos [Ajao et al. 2019] [Moraes et al. 2019].

Apesar da análise de sentimentos ser caracterizada pela aplicação de técnicas computacionais que identifiquem os aspectos de polaridade (e.g.: sentimento positivo, neutro ou negativo) e/ou da emoção (e.g.: tristeza, angústia ou raiva) presentes em um texto, até onde foi possível observar, somente o trabalho [de Souza et al. 2020] propôs um método, denominado de *FNE*, para detecção de *Fake News* baseado na classificação gramatical, de polaridade e da emoção. Cabe ressaltar que o protótipo do *FNE* proposto em [de Souza et al. 2020] foi implementado para notícias escritas em português.

Diante dos resultados promissores obtidos em [de Souza et al. 2020] e da limitação do estado da arte ao realizar detecção baseada em análise de polaridade e emoção somente em notícias escritas na língua portuguesa, este estudo levanta a seguinte hipótese: *considerar que o uso da análise de polaridade, com a inclusão da classificação das emoções, associadas à classificação gramatical dos textos das notícias pode viabilizar a construção de modelos de detecção de Fake News escritas em língua inglesa, mais robustos que os existentes na literatura*.

De forma a obter evidências experimentais que apontem para a validade da hipótese levantada, o presente trabalho propõe e avalia um protótipo adaptado para detectar *Fake News* escritas em inglês, implementado a partir de um método já existente na literatura para detecção de *Fake News* escritas em língua portuguesa que, além da classificação gramatical e da análise de sentimentos por polaridade, utiliza a emoção presente nos textos. Nos experimentos realizados em duas instâncias de *datasets* (i.e. balanceada e desbalanceada) contendo notícias em inglês, o protótipo proposto apresentou resultados superiores aos métodos que não utilizam a classificação de emoções, isto é, utilizam somente a classificação gramatical e a análise de sentimentos por polaridade para detectar *Fake News* escritas em inglês.

Este artigo está organizado como segue. A Seção 2 apresenta alguns artigos do estado da arte que têm métodos relacionados a este trabalho. O método adotado é detalhado na Seção 3, sendo o protótipo adaptado proposto na Seção 4. Em seguida, a Seção 5 descreve os experimentos e discute os resultados obtidos. Por fim, a Seção 6 faz as ponderações finais e destaca as possibilidades de futuras pesquisas.

[1] `www.who.int/emergencies/diseases/novel-coronavirus-2019`

2. Trabalhos Relacionados

Alguns dos mais recentes trabalhos que se relacionam a este estudo propuseram métodos de detecção de *Fake News* com abordagens linguísticas, mais especificamente utilizam técnicas de classificação gramatical e/ou análise de sentimentos. Assim sendo, o restante desta seção apresenta resumidamente esses trabalhos, procurando agrupá-los de acordo com o idioma da notícia (i.e. inglês ou português) a ser detectada como *Fake News*.

O primeiro grupo é formado pelos trabalhos cujas notícias foram escritas na língua inglesa. Neste grupo, o trabalho de revisão elaborado em [Shu et al. 2017] propõe um método que visa a identificar a intenção de um autor manipular a atenção do leitor através dos seus estilos de escrita, para tanto é aplicada a classificação gramatical das palavras presentes no texto da notícia. O artigo publicado em [Ajao et al. 2019], além de características gramaticais, também utiliza a análise de polaridade presente no texto das notícias. Os autores propõem uma métrica formada pela razão entre a quantidade de palavras com polaridade negativa sobre a quantidade de palavras com polaridade positiva. O trabalho proposto em [Bhavika Bhutani and Purwar 2019] procura analisar a polaridade presente no texto das notícias. Essa polaridade é associada com outras métricas, gerando um vetor para alimentar algoritmos de classificação.

No segundo grupo, os trabalhos são voltados para detecção de *Fake News* escritas na língua portuguesa. Em [de Morais et al. 2019] e [Durier and Garcia 2019] é levada em conta apenas a classificação gramatical das palavras presentes nos textos das notícias. Por meio de outra abordagem, o método proposto em [Faustini and Covões 2019] utiliza apenas a análise de polaridade. Já o artigo [Moraes et al. 2019] manipula tanto a classificação gramatical quanto a polaridade do texto. Por fim, o trabalho proposto em [de Souza et al. 2020] apresenta um método de detecção de *Fake News*, denominado de *FNE*, que considera a análise gramatical, assim como a análise de sentimentos através de polaridade e da emoção presentes no texto das notícias.

A fim de sintetizar a análise, a Tabela 1 apresenta os trabalhos de acordo com as seguintes características:

- (C1) Idioma - Informa para qual língua o trabalho é voltado (I: Notícias em Inglês, P: Notícias em Português);
- (C2) Características Gramaticais - Indica se o método apresentado no trabalho leva em conta características gramaticais;
- (C3) Análise de sentimentos com classificação de Polaridade - Exibe se foi considerada a análise de polaridade;
- (C4) Análise de sentimentos com classificação das Emoções - Expõe se foi considerada a análise das emoções.

Tabela 1. Resumo dos trabalhos relacionados.

Referência	C1	C2	C3	C4
[Shu et al. 2017]	I	X	-	-
[Ajao et al. 2019]	I	X	X	-
[Bhavika Bhutani and Purwar 2019]	I	-	X	-
[de Morais et al. 2019]	P	X	-	-
[Durier and Garcia 2019]	P	X	-	-
[Faustini and Covões 2019]	P	-	X	-
[Moraes et al. 2019]	P	X	X	-
[de Souza et al. 2020]	P	X	X	X

Diante do exposto, identificou-se que, dentre os trabalhos relacionados que detectam *Fake News* escritas em inglês, nenhum utiliza explicitamente a emoção contida no texto, deixando de considerar, portanto, nuances da linguagem (e.g.: raiva, tristeza, etc.) que evidenciem a presença de informações falsas no referido texto.

3. Método Adotado

Denominado *FNE*, o método utilizado neste trabalho é análogo ao proposto pelos autores do artigo [de Souza et al. 2020]. De uma forma geral, o *FNE* constrói modelos de aprendizado de máquina voltados à detecção de *Fake News* escritas em língua portuguesa. Para tanto, baseia-se na combinação de informações linguísticas, entre elas classificações gramaticais, polaridade e emoções extraídas de textos escritos de notícias previamente rotuladas como *fake* e *não fake*. A Figura 1 apresenta uma visão macro-funcional do método adotado.

O *FNE* recebe como entrada um conjunto de notícias N, onde cada notícia $n \in N$ possui dois atributos: $n.t$ que contém o texto divulgado em n e $n.r$ o rótulo indicando se n é *fake* ou *não fake*. As próximas Subseções detalham as etapas do método utilizado.

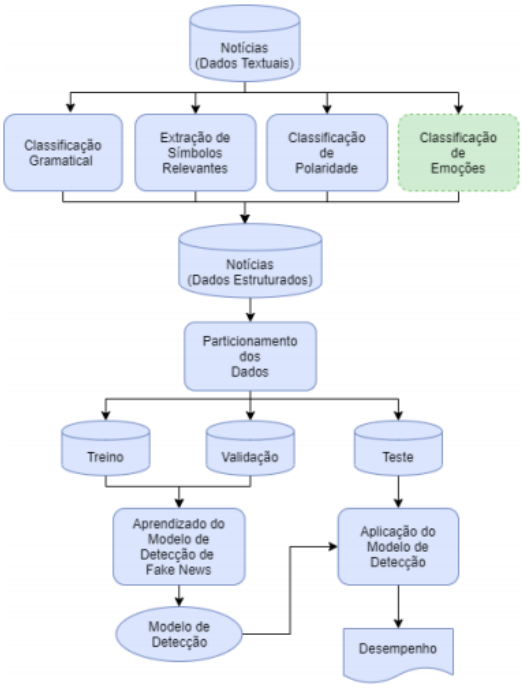

Figura 1. visão macro-funcional do FNE. Fonte: [de Souza et al. 2020]

3.1. Classificação Gramatical

Para cada notícia $n \in N$, esta etapa tem como objetivo identificar todas as palavras ou símbolos de pontuação, também chamados de *tokens*, existentes em n.t, gerando um conjunto ordenado $TK_{n.t} = \{p_1, p_2, ..., p_k\}$, onde cada p_i é um *token*. Assim, para cada token $p_i \in TK_{n.t}$, esta etapa aplica um processo de etiquetagem $\alpha(p_i)$ cujo objetivo é identificar a classe gramatical $cg \in CG$, a qual p_i pertence , onde $CG = \{cg_1, cg_2, ..., cg_{|CG|}\}$ é o conjunto de classes gramaticais consideradas. É importante observar, neste ponto, que o

FNE é configurável, cabendo ao analista de dados escolher a implementação de α a ser adotada e, consequentemente, o conjunto CG de classes gramaticais a ser utilizado.

Para otimizar a análise pelos algoritmos de aprendizado de máquina, a frequência de *tokens* em cada classe é contabilizada, sendo os valores resultantes organizados em uma matriz linha $T_{n.t}$ de forma que cada coluna indica a quantidade de *tokens* etiquetados em uma das $|CG|$ classes de CG, conforme mostra a Equação (1). Nela $|n.t_{cg_j}|$ corresponde à quantidade de *tokens* de $n.t$ cuja classe gramatical é cg_j. Cabe destacar que as referidas quantidades são normalizadas pelo total de *tokens* em $TK_{n.t}$ (i.e. k).

$$T_{n.t} = \frac{1}{k}\Big[\mid n.t_{cg_1} \mid, \mid n.t_{cg_2} \mid, ..., \mid n.t_{cg_{|CG|}} \mid\Big] \tag{1}$$

3.2. Extração de Símbolos Relevantes

Esta etapa analisa cada símbolo s em $n.t$ de forma a contabilizar símbolos considerados relevantes para o processo de detecção de *Fake News*, como, por exemplo, pontos de exclamação, aspas, caracteres em caixa alta, entre outros. Segundo [Moraes et al. 2019], quanto mais frequentes forem esses símbolos em um texto, maior a possibilidade de que esse texto seja *fake*. A fim de flexibilizar a escolha de quais símbolos especiais devem ser considerados, o *FNE* permite que o analista de dados especifique o conjunto de símbolos relevantes $CSR = \{s_1, s_2, ..., s_{|CSR|}\}$, onde, conforme o nome sugere, cada s_j é um símbolo relevante.

Assim, esta etapa percorre a cadeia $n.t$ de forma a contabilizar, para cada $s_j \in CSR$, $|n.t_{s_j}|$, i.e., o número de vezes que o símbolo s_j ocorreu em $n.t$. Após contabilizar a frequência de todos os termos de CSR, uma matriz linha $C_{n.t}$ é construída, como mostra a Equação 2, onde z é a quantidade total de caracteres em $n.t$.

$$C_{n.t} = \frac{1}{z}\Big[\mid n.t_{s_1} \mid, \mid n.t_{s_2} \mid, ..., \mid n.t_{s_{|CSR|}} \mid\Big] \tag{2}$$

3.3. Classificação de Polaridade

Para cada *token* $p_i \in TK_{n.t}$, esta etapa aplica a função parcial $P : L \rightarrow \{-1, 0, 1\}$ que procura se p_i pertence ao léxico L a fim de recuperar o valor de polaridade associado a p_i. Os valores de polaridade recuperados são somados à polaridade total de $n.t$, conforme indicado na Equação (3). Para estabelecer uma independência em relação aos diferentes tamanhos de texto, os valores de polaridade $P_{n.t}$ são normalizados (i.e. processados de forma a assumir valores no intervalo $[0, 1]$), considerando os resultados de polaridade obtidos para todas as notícias do *dataset*. É importante destacar que, neste ponto, cabe ao analista de dados configurar o *FNE* com o léxico de polaridade desejado, incluindo a função P a ser utilizada.

$$P_{n.t} = \sum_{i=1}^{k} P(p_i) \tag{3}$$

3.4. Classificação de Emoções

Usando léxicos afetivos (i.e. dicionários que relacionam emoções às palavras de um texto), é possível extrair atributos que permitem identificar a presença de emoções em textos [Tausczik and Pennebaker 2010]. Para tanto, cada palavra $p_i \in n.t$, é classificada inicialmente segundo uma função binária $affect(p_i)$, cujo resultado é 1, caso p_i pertença ao léxico afetivo escolhido pelo analista, ou 0, caso contrário. Então, cada p_i em que $affect(p_i) = 1$ é submetida às funções binárias mutuamente exclusivas $posemo(p_i)$ ou $negemo(p_i)$. Caso a emoção relacionada à p_i seja positiva no léxico afetivo, então $posemo(p_i) = 1$ e $negemo(p_i) = 0$. Caso seja negativa, então $posemo(p_i) = 0$ e $negemo(p_i) = 1$. Cada p_i com $negemo(p_i) = 1$ pode ainda ser subclassificada em uma das três emoções negativas: raiva, ansiedade ou tristeza. Para tanto, são utilizadas, respectivamente, as funções binárias $anger(p_i), anx(p_i)$ e $sad(p_i)$, que retornam valor 1, caso p_i se enquadre na referida emoção, e 0, caso contrário. Como resultado, tem-se uma matriz linha $E_{n.t}$ representada na Equação (4), onde cada coluna indica o somatório de ocorrências de *tokens* em cada uma das seis categorias afetivas apresentadas acima.

$$E_{n.t} = \left[\sum_{i=1}^{k} affect(p_i), \sum_{i=1}^{k} posemo(p_i), \sum_{i=1}^{k} negemo(p_i), \sum_{i=1}^{k} anger(p_i), \sum_{i=1}^{k} anx(p_i), \sum_{i=1}^{k} sad(p_i) \right] \tag{4}$$

3.5. Formação do Conjunto de Dados Estruturados

Ao final do processo de categorização descrito nas etapas anteriores, para cada notícia $n \in N$ um conjunto de atributos estruturados Ne_n é formado e definido pela tupla indicada na Equação 5.

$$Ne_n = (T_{n.t}, C_{n.t}, P_{n.t}, E_{n.t}) \tag{5}$$

O conjunto Ne formado pelas tuplas geradas pelas etapas descritas anteriormente, a partir de todas as notícias de N, é então armazenado para posterior processamento pelas etapas seguintes. A Equação 6 descreve formalmente tal conjunto.

$$Ne = \{Ne_n | n \in N\} \tag{6}$$

3.6. Particionamento de Dados, Aprendizado e Aplicação do Modelo

A etapa de particionamento de dados é responsável por separar Ne em três conjuntos: treino, validação e teste. Tal separação ocorre de forma aleatória, porém estratificada, assegurando a mesma distribuição de classes em cada conjunto.

Em seguida, na etapa de aprendizado do modelo de detecção de Fake News, o *FNE* treina cada um dos algoritmos de classificação indicados pelo analista de dados com as notícias do conjunto de treino. O conjunto de validação é utilizado de forma a selecionar o melhor modelo gerado por cada algoritmo, a partir de diferentes configurações de

parâmetros especificadas pelo analista. Note que, cabe ao analista escolher a métrica de avaliação dos modelos de classificação a ser utilizada (e.g.: acurácia).

Por fim, a etapa de aplicação do modelo de detecção é responsável por avaliar, no conjunto de teste, o desempenho do melhor modelo de classificação gerado por cada algoritmo na etapa anterior. A mesma métrica utilizada no treinamento dos modelos deve ser utilizada nesta etapa.

4. Protótipo Proposto

O protótipo do *FNE* proposto neste trabalho para detectar *Fake News* escritas em inglês é uma adaptação do apresentado pelos autores em [de Souza et al. 2020] para detectar *Fake News* escritas em português. Para tanto, o protótipo proposto neste estudo foi implementado em linguagem Python, utilizando técnicas de análise de sentimentos com funções disponíveis na plataforma KNIME [2] e integrado ao banco de dados MySQL. Este protótipo permite ao analista de dados configurar os métodos de cálculo dos atributos de alguns dos componentes da Equação 5. A seguir, são descritas as bibliotecas que foram utilizadas em cada uma das opções de implementação.

Para o componente $T_{n.t}$ (i.e. classificação gramatical), duas opções foram implementadas. Primeiramente a função *token.pos* da biblioteca SpaCy[3] desenvolvida em *Python*. Esta biblioteca foi aplicada por ter suporte consistente ao idioma inglês na tarefa de marcação de partes do texto (*Parts-of-Speech Tagging*). Como alternativa, foi utilizada uma parte do dicionário para o inglês do LIWC[4], com os atributos que representam classes gramaticais.

O conjunto de símbolos relevantes CSR (componente $C_{n.t}$) foi configurado com as letras maiúsculas do alfabeto romano (MAIÚSCULAS), o ponto de exclamação (!) e as aspas (”). Denominado Extrator *FNE* (*FNE*-CSR), um módulo foi implementado de forma a contabilizar a quantidade de ocorrências de cada símbolo do conjunto CSR.

A classificação de polaridade (componente $P_{n.t}$) foi desenvolvida com base no dicionário *SocialSent*[5], que consiste em um léxico que possue a polaridade para diversas palavras da língua inglesa para cada década, desde 1850. No total o léxico conta com 9550 palavras ou expressões diferentes. Foi desenvolvido um *script* que unifica os léxicos para cada década e mapeia a palavra escolhida com a polaridade mais adequada, i.e, a polaridade que está mapeada no léxico da década mais atual.

Quanto à classificação das emoções, utilizou-se o léxico do LIWC referente às emoções das palavras. Para isso, foi implementado um *script* que identifica os atributos relacionados à emoção de cada palavra do texto e contabiliza o componente $E_{n.t}$. Os valores são normalizados entre 0 e 1 para todo o conjunto de notícias analisado.

Por fim, os algoritmos de classificação implementados nesta versão do *FNE* foram os mesmos utilizados em [de Souza et al. 2020]: *Naive Bayes, AdaBoost, SVM, Gradient Boost e KNN*.

[2]https://www.knime.com/

[3]https://spacy.io/

[4]http://liwc.wpengine.com/

[5]https://nlp.stanford.edu/projects/socialsent/

5. Experimentos e Resultados

A fim de validar o protótipo proposto, foi escolhido o repositório *FakeNewsNet* que consiste em dois *datasets*: *GossipCop* e *PolitiFact*, ambos contendo notícias escritas em inglês. A escolha desses *datasets* foi guiada por duas razões principais. Primeiro, esses *datasets* foram criados para o específico propósito de detecção de *Fake News* e contêm, para cada notícia escrita em inglês, seu rótulo real, ou seja, a indicação de que a notícia é *fake* ou não. Em segundo lugar, eles foram usados e disponibilizados por publicações recentes e relevantes [Shu et al. 2019] [Sharma et al. 2019] [Shu et al. 2020]. Com o objetivo de ampliar os experimentos, foram criadas duas instâncias a partir dos dois *datasets* escolhidos. A primeira instância contém um número balanceado de notícias *fake* e não *fake*, para tanto, foi aplicado um sorteio aleatório sem reposição nas notícias pertencentes aos dois *datasets* originais. Já a segunda instância representa os dois *datasets* originais, onde o número de notícias *fake* e não *fake* está desbalanceado. A principal razão para a utilização dessa segunda instância desbalanceada é a busca em realizar experimentos com dados mais próximos da realidade, onde, espera-se que a quantidade de notícias não *fake* divulgadas em um meio digital (e.g.: Twitter) seja superior a quantidade de notícias *fake*. As Tabelas 2 e 3 fornecem, respectivamente, uma visão estatística geral das instâncias balanceada (total de 10.000 notícias) e desbalanceada (total de 21.576 notícias) dos *datasets* escolhidos, onde cada instância é formada pela concatenação dos respectivos *datasets* (i.e. PolitiFact + GossipCop).

Tabela 2. Instância balanceada dos *datasets* (10.000 notícias).

	Fake	Não Fake
PolitiFact	300	300
GossipCop	4700	4700
PolitiFact + GossipCop	5000	5000

Tabela 3. Instância desbalanceada dos *datasets* (21.576 notícias).

	Fake	Não Fake
PolitiFact	379	435
GossipCop	4768	15994
PolitiFact + GossipCop	5147	16429

Para a realização dos experimentos, foram utilizadas quatro versões do *FNE*. Como apresentado abaixo, variando a escolha do léxico gramatical (Spacy ou LIWC), as duas primeiras versões representam os dois *baselines* que não utilizaram a classificação de emoções (i.e. o caractér '-' representa a não utilização da emoção). Já para a realização dos experimentos levando em conta a emoção, as outras duas versões do *FNE* utilizaram o léxico LIWC (i.e. LIWC no lugar do caractér '-', representando a utilização da emoção).

(B1) baseline1 $\rightarrow$ *FNE*(Spacy, *FNE*-CSR, SocialSent, -)

(B2) baseline2 $\rightarrow$ *FNE*(LIWC, *FNE*-CSR, SocialSent, -)

(B1+E) *FNE*(Spacy, *FNE*-CSR, SocialSent, LIWC)

(B2+E) *FNE*(LIWC, *FNE*-CSR, SocialSent, LIWC)

Para comparar as quatro versões implementadas, em cada instância de *dataset* foi aplicada a validação cruzada [Kohavi et al. 1995] com 10 conjuntos, onde a acurácia,

precisão e recall foram as métricas de avaliação de desempenho adotadas. A partir dos melhores resultados obtidos, a Tabela 4 resume a configuração utilizada em cada um dos algoritmos classificadores utilizados nos experimentos.

Tabela 4. Configurações dos classificadores.

Algoritmo	Parâmetros
Naive Bayes (NB)	padrão
AdaBoost (AB)	Random Forest I = 10 iterações depth = 0
SVM	C = 1000 gama = 0.001 kernel linear
Gradient Boost (GB)	níveis = 5 modelos = 100 learning rate = 0.1
KNN	k = 3

Para cada um dos *baselines*, os resultados dos experimentos com as instâncias balanceada e desbalanceada são apresentados nas Tabelas 5, 6, 7 e 8. De uma forma geral, existem evidências que apontam para a validade da hipótese levantada neste trabalho de que a combinação de informações gramaticais, com a polaridade e as emoções presentes nos textos das notícias pode levar a melhores modelos de detecção de *Fake News*, escritas em inglês, do que aqueles modelos que consideram apenas informações gramaticais e a polaridade desses textos. Abaixo segue uma análise mais detalhada sobre os resultados obtidos.

Em ambas as instâncias (i.e. balanceada e desbalanceada) percebe-se que a acurácia obtida pela versão baseada nas emoções superou o *baseline* correspondente (i.e. em 70% (7 em 10) das comparações entre as versões do *FNE* com o uso das emoções e seus respectivos *baselines*).

Ademais, comparando os *baselines* em ambas as instâncias, percebe-se que, para a língua inglesa, os melhores resultados foram obtidos com o léxico gramatical Spacy em comparação com o léxico gramatical LIWC. Na instância balanceada, o LIWC foi superior somente na acurácia e precisão para o classificador GB e no recall para o classificador NB. Já na instância desbalanceada, o LIWC foi superior só na acurácia e recall para o classificador NB e na precisão para o classificador SVM.

Pode-se ainda observar que a escolha do algoritmo de classificação pode influenciar nos resultados, apesar da inclusão dos atributos de emoção. Esse comportamento com viés negativo pode ser visto, utilizando o *baseline1*, nos resultados de acurácia com o classificador AB na instância balanceada e do classificador GB na instância desbalanceada. Entretanto, esse comportamento com viés positivo, pode ser notado com a utilização do classificador NB que apresentou as maiores diferenças na acurácia em relação ao *baseline2*, sendo (3.59pp) na instância balanceada e (3.38pp) na instância desbalanceada.

Com relação somente à instância balanceada, destaca-se que o uso da classificação da emoção superou todos os valores de acurácia alcançados pelo *baseline2*, além da maior parte dos valores de precisão e recall (3 em 5). No que diz respeito somente à instância desbalanceada, onde buscou-se realizar experimentos com dados mais próximos da realidade (i.e. qtd notícias não *fake* $>$ qtd notícias *fake*), os resultados com a inclusão dos atributos de emoção superou a maioria dos valores de acurácia e precisão obtidos pelo

baseline2 correspondente (4 em 5). Entretanto, a maior parte dos resultados obtidos com recall no *baseline2* foram superiores (3 em 5).

Além disso, constatou-se uma superioridade nas acurácias obtidas na maioria dos resultados deste trabalho (i.e. detecção de *Fake News* escritas em inglês), ao serem comparados com as apresentadas pelos autores em [de Souza et al. 2020] (i.e. detecção de *Fake News* escritas em português). Uma razão para essa constatação, seria uma possível melhor qualidade dos léxicos disponíveis para aplicação em textos em língua inglesa, se comparados aos léxicos para língua portuguesa.

Por fim, apesar do desafio das *Fake News* estar longe de ser resolvido, os resultados obtidos mostram que a utilização da classificação da emoção pode contribuir positivamente para a detecção de *Fake News* escritas em língua inglesa.

Tabela 5. Resumo dos resultados na instância balanceada para Baseline 1.

	B1			B1+E		
Classificadores	Acurácia	Precisão	Recall	Acurácia	Precisão	Recall
NB	91.64%	**93.54%**	89.46%	**91.91%**	**93.54%**	**90.04%**
GB	99.97%	99.96%	**99.98%**	**99.98%**	**99.98%**	**99.98%**
AB	**99.98%**	**99.98%**	**99.98%**	99.97%	99.97%	99.96%
SVM	**95.00%**	**96.24%**	93.66%	94.65%	93.59%	**95.86%**
KNN	**99.19%**	**99.04%**	**99.34%**	98.61%	98.58%	98.64%

Tabela 6. Resumo dos resultados na instância balanceada para Baseline 2.

	B2			B2+E		
Classificadores	Acurácia	Precisão	Recall	Acurácia	Precisão	Recall
NB	89.42%	87.46%	92.04%	**93.01%**	**92.22%**	**93.96%**
GB	99.98%	99.98%	**99.98%**	**99.99%**	**99.99%**	**99.98%**
AB	99.89%	**99.98%**	**99.96%**	**99.90%**	99.86%	99.92%
SVM	94.67%	**95.81%**	93.42%	**94.87%**	93.67%	**96.24%**
KNN	96.23%	96.18%	96.28%	**97.06%**	**97.06%**	**97.06%**

Tabela 7. Resumo dos resultados na instância desbalanceada para Baseline 1.

	B1			B1+E		
Classificadores	Acurácia	Precisão	Recall	Acurácia	Precisão	Recall
NB	87.21%	97.20%	85.67%	**87.76%**	**97.24%**	**86.21%**
GB	**99.99%**	**99.99%**	**99.99%**	99.98%	**99.99%**	**99.99%**
AB	99.97%	**99.98%**	99.98%	**99.98%**	**99.98%**	**99.99%**
SVM	96.96%	95.08%	**92.05%**	**97.01%**	**95.42%**	91.89%
KNN	**99.35%**	**99.56%**	**99.58%**	99.04%	99.41%	99.33%

Tabela 8. Resumo dos resultados na instância desbalanceada para Baseline 2.

	B2			B2+E		
Classificadores	Acurácia	Precisão	Recall	Acurácia	Precisão	Recall
NB	87.31%	89.32%	**94.65%**	**90.69%**	**96.06%**	91.53%
GB	99.98%	**99.99%**	**99.99%**	**99.99%**	**99.99%**	**99.99%**
AB	99.88%	99.86%	**99.98%**	**99.94%**	**99.97%**	99.95%
SVM	**96.94%**	95.49%	**91.49%**	96.84%	**95.66%**	90.88%
KNN	97.40%	98.09%	98.50%	**97.74%**	**98.43%**	**98.60%**

6. Considerações Finais

Cada vez mais pessoas estão consumindo notícias dos meios digitais, ao invés dos canais tradicionais. Tal tendência amplificou a disseminação de *Fake News*, isto é, as notícias intencionalmente falsas. Este tipo de notícia pode ter significativos impactos sociais negativos, por exemplo, a manipulação da opinião em larga escala. Este cenário levou ao desenvolvimento de abordagens automáticas para detecção de *Fake News* em várias frentes de atuação, de forma a abranger diferentes aspectos para a solução do problema.

Dentre essas abordagens, destacam-se as linguísticas que utilizam informações que podem ser extraídas diretamente do texto da notícia. Baseados nessas abordagens, foram desenvolvidos métodos que utilizam a classificação gramatical e/ou a análise de sentimentos presentes na escrita das notícias em língua inglesa e portuguesa. Entretanto, até onde foi possível observar, a análise de sentimentos presente nos trabalhos voltados para *Fake News* em inglês se limitam à utilização da polaridade. Tendo como base essa limitação, este estudo levantou a hipótese de que a ampliação do uso de técnicas de análise de sentimentos, em particular com a inclusão da classificação das emoções, associadas a classificação gramatical dos textos das notícias pode viabilizar a construção de modelos de detecção de *Fake News* em língua inglesa, mais robustos que os existentes na literatura.

Durante esse trabalho, foram identificados léxicos que permitem classificar emoções humanas retiradas de textos escritos em inglês e transformar essas informações em elementos que possam ser processados por algoritmos. Desta forma, este estudo propôs e aplicou um protótipo adaptado da língua portuguesa para a inglesa, onde os resultados obtidos com os experimentos realizados apresentaram evidências que apontam para a validade da hipótese levantada.

Com base nas limitações deste trabalho, destacam-se como trabalhos futuros a realização de estudos visando avaliar o custo da complexidade ao inserir a classificação da emoção na análise de sentimentos, assim como, a ampliação dos experimentos em busca de uma maior variedade de resultados. Para tanto, podem ser utilizados outros *datasets* (i.e. outros vieses), a aplicação de algoritmos de classificação capazes de trabalhar com grande quantidade de dimensões e complexas fronteiras de separação (e.g.: redes neurais e *deep learning*) e o uso de seletores de características antes da classificação.

7. Agradecimento

O presente trabalho foi realizado com apoio do Ministério da Ciência, Tecnologia, Inovações e Comunicações, do Ministério da Saúde e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - Código de Financiamento (CNPq) 401662/2020-9.

Referências

Ajao, O., Bhowmik, D., and Zargari, S. (2019). Sentiment Aware Fake News Detection on Online Social Networks. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2019-May, pages 2507–2511. Institute of Electrical and Electronics Engineers Inc.

Bhavika Bhutani, Neha Rastogi, P. S. and Purwar, A. (2019). Fake news detection using sentiment analysis. In *Institute of Electrical and Electronics Engineers (IEEE)*.

Bondielli, A. and Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

de Morais, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Barbon, S. (2019). Deciding among Fake, Satirical, Objective and Legitimate news. pages 1–8. Association for Computing Machinery (ACM).

de Souza, M. P., da Silva, F. R. M., Freire, P. M. S., and Goldschmidt, R. R. (2020). A linguistic-based method that combines polarity, emotion and grammatical characteristics to detect fake news in portuguese. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 217–224, New York, NY, USA. Association for Computing Machinery.

Durier, F. and Garcia, A. C. (2019). Fake news and sarcasm, what is the limit of a critic and what is intentionally fake? In *Anais do Simpósio Brasileiro de Sistemas Colaborativos*, pages 58–61. Sociedade Brasileira de Computação - SBC.

Faustini, P. and Covões, T. (2019). Fake news detection using one-class classification. In *Proceedings - 2019 Brazilian Conference on Intelligent Systems*, pages 592–597. BRACIS 2019.

Freire, P. M. S. and Goldschmidt, R. R. (2019). Uma introdução ao combate automático às fake news em redes sociais virtuais. In *Tópicos de Gerenciamento de Dados e Informação*, 34th SBBD, pages 38–67, Fortaleza, CE, Brazil. SBC.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, Montreal, Canada. Montreal, Canada, researchgate.

Mejova, Y. and Kalimeri, K. (2020). Advertisers jump on coronavirus bandwagon: Politics, news, and business. *ArXiv*, abs/2003.00923.

Moraes, M. P., Sampaio, J. d. O., and Charles, A. C. (2019). Data mining applied in fake news classification through textual patterns. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web - WebMedia '19*, pages 321–324. ACM Press, New York, New York, USA.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM TIST*.

Shu, K., Mahudeswaran, D., and Liu, H. (2019). Fakenewstracker: A tool for fake news collection, detection, and visualization. *Comput. Math. Organ. Theory*, 25(1):60–71.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Shu, K., Sliva, A., WNAG, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. In *SIGKDD Explor. Newsl. 19, 1*, pages 22—-36.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Vosoughi, S., Mohsenvand, M. N., and Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on twitter. *ACM Trans. Knowl. Discov. Data*, 11(4):50:1–50:36.