Data mining applied in fake news classification through textual patterns

Marcos Paulo Moraes
marcospaulo.moraes@ufrj.br
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ

Jonice de Oliveira Sampaio
jonice@dcc.ufrj.br
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ

Anderson Cordeiro Charles
andersoncordeironf@gmail.com
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ

ABSTRACT

Fake news has been around for a long time. But with the advancement of social media and internet access, fake news has become a bigger problem. Because of the rapid spread in social media and instant messaging applications, fake news can reach more people in less time by directly influencing democratic processes, leveraging security issues that sometimes lead to tragic ends. In order to promote a fast and automated method of fake news identification, in this study, we performed an analysis of false Brazilian news, identifying writing patterns through natural language processing and machine learning.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → *Information systems applications*; • **Information systems applications** → Data Mining.

KEYWORDS

fake news, data mining, natural language processing, machine learning, sentiment analysis

1 INTRODUÇÃO

Historiadores afirmam que notícias falsas existem desde o século 6 [10]. Contudo, a disseminação das fake news encontra novos patamares na era do mundo conectado. Usuários de mídias, quase sempre na internet, são bombardeados com notícias que nem sempre são verdadeiras, e que em casos extremos causam a morte de inocentes [7].

A velocidade na comunicação impede que uma avaliação possa ser feita no conteúdo que está sendo trafegado e, além disso, os embates sociais resultantes de divergências sócio políticas impulsionam a prática da produção de conteúdo duvidoso que sirva de alicerce para críticas ou fomente discussões na rede. Identificar esses boatos se apresenta como um desafio; estabelecer a confiabilidade de informações online é um desafio assustador mas crítico [4]. Apesar do aumento de ferramentas de fact checking no Brasil, como exemplos, Fato ou Fake, E-farsas, Boatos.org e a Agência Lupa, há um tempo considerável entre o início de compartilhamento e a validação dessa notícia por tais ferramentas, em sua maioria manuais, realizadas por jornalistas ou especialistas no contexto, dependendo assim de esforço humano para validação. Torna-se necessária, então, a criação de mecanismos automáticos para detecção de notícias falsas com o intuito de diminuir o tempo de verificação.

A proposta deste trabalho é analisar textualmente fake news propagadas no Brasil entre 2016 e 2018 a fim de identificar padrões na escrita que possibilitem a criação de um método automatizado capaz de auxiliar a checagem de notícias quanto a sua veracidade.

2 TRABALHOS RELACIONADOS

Estudos dos impactos das fake news na sociedade são diversos, como a influência delas nos processos democráticos ao redor do mundo, com a eleição de Trump [1]. Existem, também, estudos de análise de texto em português em redes sociais [9]. Porém, ainda são raros os estudos sobre a estrutura linguística desses textos em português, devido às características sócio culturais e figuras de linguagem como ironia, sarcasmos, além dos erros ortográficos.

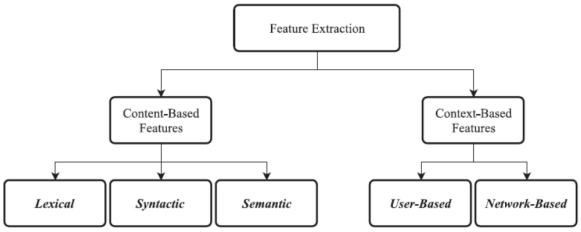

Figure 1: Diferentes tipos de extração de atributos usados na literatura para detecção de fake news (A. Bondielli and F. Marcelloni / Information Sciences 497 (2019))

Em [2], Bondielli e Marcelloni realizaram estudo sobre técnicas de detecção de rumores e notícias falsas. Verificaram duas abordagens de extração de atributos: baseado no conteúdo e no contexto do texto (figura 1). Estudos que consideram textos publicados em mídias sociais devem utilizar a abordagem baseada em contexto, uma vez que se mostrou mais eficaz na detecção devido ao tamanho curto dos textos. Nos estudos que consideram atributos baseados no conteúdo, como é o objetivo deste trabalho, são utilizados métodos de análise léxica (por exemplo, n-gramas), sintática (Part of Speech Tagging) e semântica (análise de sentimento). Shu et. al. [8] conduziram estudo para detectar conteúdo falso compartilhado nas mídias sociais devido ao aumento de consumo de notícias nestas plataformas.

Em [3], Charles et al propuseram a criação de uma fonte de dados para consulta de notícias falsas e verdadeiras: a Fakepedia, com o

In: XVII Workshop de Trabalhos de Iniciação Científica (WTIC 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

intuito de auxiliar a tarefa de verificação de informação, concentrando em um mesmo lugar a checagem de notícias realizadas por outros portais especializados. Além disso, a plataforma fará uso de crowdsourcing para alimentar seu banco de dados de notícias visando mantê-la atualizada e disponível para consultas. Além de uma interface web de busca e recuperação de notícias, a ferramenta disponibiliza uma API, utilizada neste trabalho para extração de notícias verificadas.

Monteiro et al [6], analisou manualmente 7200 notícias e identificou que os percentuais de classes gramaticais foram próximos para notícias falsas e verdadeiras. Utilizando o algoritmo SVM, obtiveram uma acurácia de 89%. Monteiro calculou a quantidade de erros gramaticais e verificou que notícias falsas possuem 10 vezes mais erros que notícias verdadeiras. E utilizou outras variáveis como pausalidade, incerteza, emotividade e não imediatismo do texto. Além dos resultados obtidos, o trabalho de Monteiro resultou na criação de um corpus de notícias chamada Fake.Br [6] que possibilita a identificação de alguns padrões na escrita de fake news.

Em seu estudo, Stiilpen e Merschmann [9] propuseram uma metodologia para análise de textos em português compartilhados no Twitter e revisões do Google Play Store, que muitas vezes são escritos em língua informal e curtos, onde a falta de contexto adiciona obstáculos na mineração de texto. A metodologia é parecida com a aplicada neste estudo, fazendo uso de bibliotecas para processamento de linguagem natural e extraindo métricas como utilização de classes gramaticais para auxiliar a categorização de texto e análise de sentimento atingindo acurácia de 81% na classificação utilizando o algoritmo SVM.

Nesta pesquisa, adicionaremos outras variáveis e métricas para identificação de notícias falsas não contempladas nos estudos anteriores, como a distribuição dos valores de cada classe gramatical, e seu desvio padrão, para verificar que existem diferenças na escrita de notícias falsas e verdadeiras. Aplicando em conjunto com análise de sentimento e uso da pontuação de exclamações, bastante utilizado em notícias falsas. Assim, ratificando conclusões de estudos anteriores e os incrementando com informações relevantes.

3 MATERIAIS E MÉTODOS

A condução das atividades de análise e classificação de notícias baseia-se nos seguintes passos: pesquisa de estudos anteriores, seleção dos dados, unificação de base, preparação e pré processamento, cálculo de campos, seleção dos algoritmos, treinamento e classificação para posterior análise dos resultados frente a estudos relacionados. [5]

Com o objetivo de reunir dados suficientes para aplicação dos algoritmos de classificação, foram utilizadas duas bases de notícias que circularam no Brasil e já verificadas: a Fakepedia [3], desenvolvida como ferramenta de crowdsourcing para verificação de notícias que reúne as verificações realizadas por diversas agências de fact checking no Brasil, e disponibiliza uma API para realização de consultas no banco de dados; e o corpus do projeto Fake.br [8], que possui notícias verdadeiras e falsas, manualmente verificadas, criado por grupos de pesquisa da USP e UFSCar [6].

3.1 Arquitetura

Para realizar a análise, utilizou-se a linguagem Python, bibliotecas de processamento de linguagem natural (NLTK e spaCy) e aplicação de algoritmos para aprendizado de máquina (scikit-learn). Para o uso de uma base unificada, utilizou-se ferramentas da stack ELK: elasticsearch e kibana. Como a arquitetura (figura 2) do projeto contempla a coleta de dados das duas fontes e inserção em base única, a adição de notícias ou outras fontes de dados é facilitada pela padronização da base e escalabilidade da ferramenta utilizada.

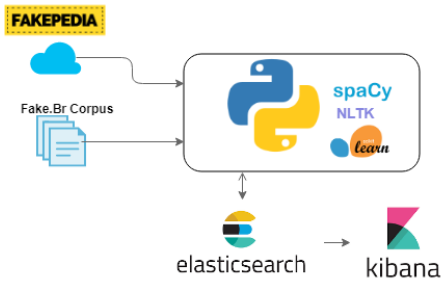

Figure 2: Arquitetura proposta do estudo.

3.2 Dados

O corpus Fake.Br possui uma base totalmente balanceada de notícias, com 7204 matérias, (3602 falsas e 3602 verdadeiras) criado a partir de notícias de sites entre o período de janeiro de 2016 e janeiro de 2018, o formato do arquivo é texto com organização posicional, com métricas da notícia em arquivo separado.

Como o intuito do trabalho é identificar padrões na escrita das fake news, utilizaram-se os seguintes dados presentes no corpus: texto e título (quando disponível) da notícia. Informações como link de acesso à matéria, categoria (política, sociedade, TV & celebridades, ciência & tecnologia e religião) e data de publicação podem ser utilizados posteriormente para validações temporais, de fonte ou padrões de escrita entre diferentes categorias, por exemplo.

Além dos dados do corpus Fake.br, também foram utilizados os dados pro provenientes da Fakepedia. Que possui em sua base 4858 notícias, onde:

- 3825 são falsas;
- 1033 notícias verdadeiras.

Dos dados extraídos da Fakepedia, foram utilizados o texto da notícia e o título. Para se ter uma melhor acurácia com textos completos e com conteúdo suficiente, foram filtradas notícias com menos de 70 palavras, por não possuírem texto com tamanho satisfatório para análise. Foram filtradas notícias falsas com menos que 120 para proporcionalizar a base. Com os filtros, a quantidade de notícias ficou em 8776 notícias e temos uma melhor proporção entre notícias falsas e verdadeiras (figura 3).

3.3 Implementação

Após indexar os dados de cada notícia das diferentes fontes na base única elasticsearch, pode-se realizar os cálculos necessários para o estudo. Os primeiros campos calculados foram a partir da biblioteca spaCy. Fazendo uso de funcionalidades para português, foi possível

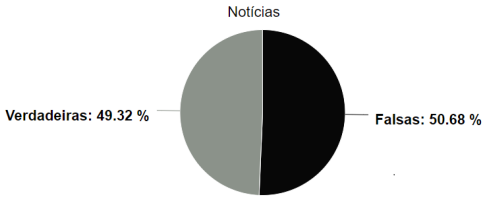

Figure 3: Proporção de notícias verdadeiras e falsas.

calcular o POS (Part Of Speech) Tagging dos textos, onde cada token é classificado em substantivo, adjetivo, verbo, verbo auxiliar, advérbio, etc.

Os campos foram indexados no elasticsearch para posterior consulta e análise. Além dos valores literais, também foi computado o percentual de cada classe gramatical em relação ao total de palavras. Essa informação é relevante, pois existem textos com tamanhos bem diferentes, o que gera a necessidade de normalização desses valores. Ainda foram criados os campos: quantidade e percentual de caracteres maiúsculos e pontos de exclamação e sentimento no texto.

Para o cálculo de sentimento do texto, foi utilizado o léxico SentiLex (um léxico de sentimento especificamente concebido para a análise de sentimento e opinião sobre entidades humanas em textos redigidos em português), com polaridade (negativo, neutro, positivo) para cada verbete presente. Esse cálculo foi realizado de forma simples: para cada token na notícia, é verificada a sua polarização no léxico, obtendo, no fim, a polaridade final do texto. Essa abordagem pode criar erros que serão discutidos nos resultados e trabalhos futuros. A partir das análises realizadas, foram criados dois classificadores de notícias: um utilizando os textos das notícias de forma vetorizada e outro utilizando somente as métricas calculadas.

Utilizaram-se algoritmos conhecidos para a classificação de texto, como o Multinomial Naive Bayes e SVM, aplicados em trabalhos relacionados e o AdaBoost, com alto desempenho, para comparar sua acurácia. A parametrização dos algoritmos considerou o melhor resultado a partir de execuções via GridSearchCV. Utilizando somente as métricas como percentual de cada classe gramatical, sentimento e quantidade maiúsculas e exclamações nos textos, sem considerar o texto em si (ou seja, sem os textos vetorizados), foram executados os algoritmos Naive Bayes (Gaussian, para dados contínuos), SVM e AdaBoost. A partir dessa execução é possível comparar com o classificador que considera o texto também. As configurações dos algoritmos para aprendizado de máquina são 60% da base como dados de treinamento, 20% da base como dados de teste e 20% da base como dados de validação.

4 RESULTADOS

Com os novos campos calculados e adicionados à base elasticsearch, a primeira métrica avaliada foi a média das classes gramaticais presentes no texto. Na tabela 1, temos as médias percentuais de algumas classes gramaticais. Todas possuem médias próximas para notícias falsas e verdadeiras. Característica que se repete com as outras classes gramaticais avaliadas. Resultados melhores foram

Table 1: Médias de variáveis

Classes	Falsas	Verdadeiras
%Substantivos	16,67	18,43
%Adjetivos	4,30	4,64
%Verbos	12,56	11,56
%Nomes próprios	10,45	10,17
#Sentimento	-4,13	-1,63
%Maiúsculas	4,15	5,92
%Exclamação	0,15	3,45

obtidos com os campos de sentimento, e percentual de letras maiúsculas e exclamações. Tais variáveis mostraram diferenças maiores entre notícias de classes distintas, como pode ser visto na tabela 1. Ambas possuem, em média, polaridade negativa. Tal resultado pode ser melhor verificado ou retificado utilizando métodos de cálculo de sentimento mais avançados, incluindo léxicos com bigramas e verificando o contexto em que a palavra está sendo utilizada. Já os percentuais de maiúsculas e pontos de exclamação nos textos mostram a tendência das fake news de chamarem a atenção, como se estivessem gritando ao leitor. Existem mais notícias falsas com letras maiúsculas e o uso de exclamações é também maior nesses textos, mostrando a maior distância entre valores encontrados no estudo.

Como a média mostrou ser próxima para as classes gramaticais, não trazendo informações que distinguissem as classes de notícias, verificou-se, então, a distribuição dessas métricas. Foram gerados gráficos do tipo boxplot com a distribuição das classes gramaticais para notícias verdadeiras e falsas. Na figura 4, vemos que os percentuais de adjetivos das notícias falsas possuem uma dispersão (desvio padrão) maior em relação às notícias verdadeiras. O mesmo comportamento se repete para as outras classes gramaticais.

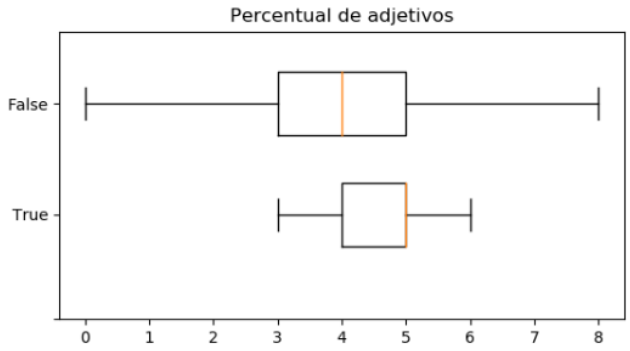

Figure 4: Boxplot com a distribuição de percentuais de adjetivos entre notícias falsas e verdadeiras.

Após análises nos textos, foram criados os classificadores para identificar se notícias são falsas ou verdadeiras. Foi escolhido o model selection da biblioteca scikit-learn para avaliar mais de um modelo de predição. Para a remoção de stopwords e lematização, foi utilizada a biblioteca NLTK via python. Na classificação, foram contabilizados quatro tipos de resultados:

Table 2: Resultados dos algoritmos

	Acurácia	
Algoritmo	Texto vetorizado (%)	Métricas (%)
Adaboost	93	93
Naive Bayes	82	84
SVM	93	94

Table 3: Resultados da classificação

		Valor calculado	
		verdadeiro	falso
Valor esperado	verdadeiro	855	43
	falso	69	792

- Verdadeiro positivo: foi detectado falso e o texto é falso;
- Verdadeiro negativo: foi detectado verdadeiro e o texto é verdadeiro;
- Falso positivo: foi detectado falso e o texto é verdadeiro;
- Falso negativo: foi detectado verdadeiro e o texto é falso.

Com os algoritmos citados anteriormente, a taxa de acerto oscilou entre 82% e 93%, para os algoritmos Multinomial Naive Bayes e AdaBoostClassifier, respectivamente, quando utilizamos o texto vetorizado na classificação. O SVM obteve 93% de acurácia, o que se mostrou melhor do que o resultado de estudos anteriores para classificação de textos escritos na língua portuguesa.

Aplicando a classificação com as métricas calculadas, o Gaussian Naive Bayes mostrou acurácia de 84%, e o AdaBoost atingiu 93% novamente. O SVM 89% parametrizado com valores default e 94% quando utilizados os parâmetros C=10 e gamma=0.001, otimizados a partir do módulo GridSearchCV. Os resultados estão na tabela 6. Nos dois classificadores, o algoritmo de boosting não obteve melhor resultado que o SVM.

A classificação utilizando somente as métricas obteve tempo de execução inferior ao da classificação com os textos vetorizados, mesmo obtendo resultados próximos. Isso se deve ao fato de que a vetorização dos textos geram matrizes esparsas, com 60 mil posições, dependendo do tamanho do dataset. Enquanto que as métricas calculadas são menos de 30. O melhor resultado da classificação é mostrado na tabela 5 abaixo.

A análise de textos compartilhados em redes sociais em estudos de Stiilpen [9] obteve 81% de acurácia, porém, como a aplicação é para textos curtos, são utilizados métodos que podem interferir na análise de notícias, como a aplicação de correções gramaticais, pontuação e remoção de gírias. Tais modificações removem características inerentes à escrita das notícias, como o uso de exclamações. O trabalho de Monteiro [6] utilizou o algoritmo SVM na construção de um classificador e obteve uma acurácia de 89% utilizando as features mencionadas anteriormente como emotividade e pausalidade. Já este estudo utilizou a análise de sentimento e, como visto, também não encontrou grandes diferenças na escrita de notícias falsas e verdadeiras, com ambas tendo polaridade negativa.

5 CONCLUSÕES E TRABALHOS FUTUROS

Com este trabalho, foi possível analisar a estrutura gramatical das notícias falsas, fazendo um comparativo com notícias verdadeiras, validando estudos anteriores e adicionando novas variáveis para ajudar na sua identificação. Gerando, portanto, uma importante ferramenta para auxiliar as plataformas de fact checking.

Apesar de as médias serem iguais, a distribuição das classes gramaticais em notícias falsas possui desvio padrão maior que as notícias verdadeiras. Denotando que notícias falsas possuem estilos de escrita mais diversificados. Compreensível, uma vez que tais matérias possuem diversas fontes, já as notícias verdadeiras são provenientes de menos sites. A quantidade de exclamações também deve ser considerada na identificação de notícias falsas, já que a presença delas é maior.A análise de sentimento mostrou que ambas classes de notícias possuem polaridade negativa, com as verdadeiras um pouco mais. Essa análise deve ser revista contemplando termos de negação do léxico para inversão de sua polaridade e utilizando-se de técnicas mais sofisticadas de análise de sentimento.

Como trabalhos futuros, prevemos considerar informações referentes aos divulgadores de notícias em mídias sociais e suas redes, para realizar uma análise baseada no contexto, conforme pesquisa de Bondielli e Marcelloni [2]. Também é necessário realizar o estudo considerando as categorias de notícias, como política e sociedade, para verificarmos se as diferenças se tornam mais evidentes. Como o compartilhamento de imagens com textos é uma das principais formas de difusão de notícias falsas, também é necessário a aplicação de métodos de extração para que tais textos possam ser analisados, sendo um limitante do trabalho atual. A partir da obtenção desses textos, é possível aplicar a mesma metodologia deste estudo.

REFERENCES

[1] & Gentzkow M. Allcott, H. 2017. Social media and fake news in the 2016 election. Journal of economic perspectives. *NBER* (2017).

[2] F. Bondielli, A. & Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences* (2019).

[3] & de Oliveira Sampaio J. Charles, A. C. 2018. Checking fake news on web browsers: an approach using collaborative datasets. In *Posters at Big Social Data and Urban Computing - BiDU*. CEUR.

[4] Victoria L. Conroy, N. and Y. Chen. 2015. Automatic Deception Detection: Methods for Finding Fake News. *ASIST* (nov 2015).

[5] I; Pintelas P. Supervised machine learning Kotsiantis, S.B.; Zaharakis. 2007. Supervised machine learning: A review of classification techniques. *Informatica* (2007).

[6] Pardo T.A.S. de Almeida T.A. Ruiz E.E.S. Vale O.A. Monteiro R.A., Santos R.L.S. 2018. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. *PROPOR* (2018).

[7] Mariane Rossi. 2005. Mulher espancada após boatos em rede social morre em Guarujá, SP. http://g1.globo.com/sp/santos-regiao/noticia/2014/05/mulher-espancada-apos-boatos-em-rede-social-morre-em-guaruja-sp.html

[8] Sliva A. Wang S. Tang J. & Liu H. Shu, K. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* (2017).

[9] & Merschmann L. H. C. Stiilpen Junior, M. 2016. A methodology to handle social media posts in brazilian portuguese for text mining applications. *WebMedia* (2016).

[10] Fábio Victor. 2017. Notícias falsas existem desde o século 6, afirma historiador Robert Darnton. https://www1.folha.uol.com.br/ilustrissima/2017/02/1859726-noticias-falsas-existem-desde-o-seculo-6-afirma-historiador-robert-darnton.shtml