MINERAÇÃO DE DADOS APLICADA À IDENTIFICAÇÃO DE NOTÍCIAS FALSAS

Marcos Paulo Moraes[1] (Aluno)**, Anderson Cordeiro Charles**[1] (Co-orientador)**, Jonice Oliveira**[1] (Orientadora)

[1]Universidade Federal do Rio Janeiro (UFRJ) Rio de Janeiro – RJ – Brasil

marcospaulo.moraes@ufrj.br, andersoncordeironf@gmail.com, jonice@dcc.ufrj.br

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.*

Resumo. *Notícias falsas existem desde o século VI. Porém, com o avanço da mídias sociais e do acesso à internet, elas se tornaram um problema maior. Devido à rápida disseminação em mídias sociais e aplicativos de mensagens instantâneas, notícias falsas podem alcançar mais pessoas em menos tempo e influenciar diretamente os processos democráticos, criar ou expandir crises sociais, alavancando problemas de segurança que às vezes levam a fins trágicos. Com o intuito de gerar um método rápido e automatizado de identificação de notícias falsas, este estudo realizou uma análise dessas notícias, escritas em português, a partir de um corpus e outras fontes de dados verificadas. Validamos estudos anteriores e adicionamos novas variáveis para ajudar na identificação de notícias falsas.*

1. Introdução

Historiadores afirmam que notícias falsas existem desde o século VI, com o intuito de arruinar a reputação de pessoas no poder. Já, por exemplo, entre os séculos 14 e 18, tais notícias se tornaram fontes de dinheiro a partir de chantagem, para que não houvesse difamação dos alvos a partir de mentiras ou fatos distorcidos (Victor, 2017).

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 (Rossi, 2014).

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 sociopolíticas 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.

Apesar do aumento de ferramentas de fact checking no Brasil, como exemplos, Fato ou Fake[1], E-farsas[2], Boatos.org[3] e a Agência Lupa[4], 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.

O objetivo geral deste trabalho é realizar a análise textual de notícias falsas escritas na língua portuguesa a partir de bases de notícias coletadas entre 2015 e 2019. Visando a identificação de padrões em sua escrita para auxiliar iniciativas de combate à disseminação de desinformação.

Como objetivos específicos:

- Identificar variáveis como o uso de classes gramaticais, análise de sentimento e n-gramas e calcular métricas (percentual médio, mediana, desvio padrão e distribuição) para identificação de padrões que diferenciem as classes de notícias;
- Aplicar algoritmos de aprendizado de máquina para a identificação de notícias falsas, criando um método automatizado que auxilie a checagem de notícias quanto a sua veracidade, considerando o português como língua de escrita.

2. Trabalhos relacionados

Fake news influenciam processos democráticos ao redor do mundo, como a eleição de Trump (Allcott, 2017). Há, também, estudos de análise de texto em português em redes sociais (Stiilpen, 2016). 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.

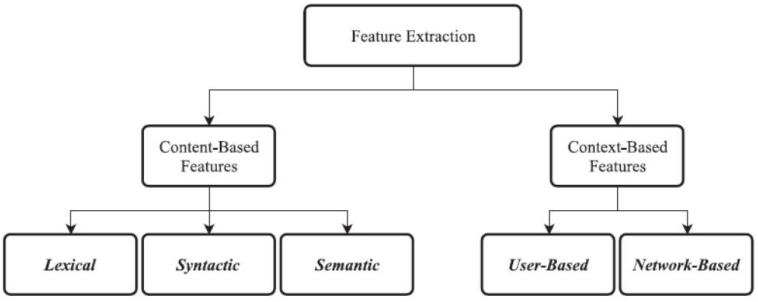

Figura 1: Diferentes tipos de extração de atributos usados na literatura para detecção de fake news (A. Bondielli and F. Marcelloni / Information Sciences 497 (2019))

[1] https://g1.globo.com/fato-ou-fake/
[2] https://www.e-farsas.com/
[3] https://www.boatos.org
[4] https://piaui.folha.uol.com.br/lupa/

Em pesquisa de Bondielli e Marcelloni, 2019, 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, 2017, conduziram estudo para detectar conteúdo falso compartilhado nas mídias sociais devido ao aumento de consumo de notícias nestas plataformas.

Em Charles et al, 2017, propuseram a criação de uma fonte de dados para consulta de notícias falsas e verdadeiras: a Fakepedia, com o 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, 2018, analisaram 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 (Monteiro, 2018) que possibilita a identificação de alguns padrões na escrita de *fake news*.

Em seu estudo, Stiilpen e Merschmann, 2016, 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 exclamação, 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 neste trabalho baseia-se nos seguintes passos: coleta de dados a partir de bases disponibilizadas a partir de outros trabalhos, 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 (Kotsiantis, 2007).

O estudo fez uso de duas bases de notícias: a Fakepedia (Charles, 2018), desenvolvida como ferramenta de crowdsourcing para validação de notícias que reúne as verificações realizadas por diversas agências de fact checking no Brasil, e disponibiliza uma API para realização de consultas em seu banco de dados; e o corpus do projeto Fake.br (Monteiro, 2018), que possui notícias verdadeiras e falsas, manualmente verificadas, criado por grupos de pesquisa da USP e UFSCar.

3.1. Arquitetura

Para realizar a análise, utilizou-se de mecanismos para processamento de linguagem natural, como tokenização para identificação das classes gramaticais em uso, remoção de pontuação e aplicação de lowercase no texto para cálculo de palavras mais utilizadas no texto. Fez-se uso de uma plataforma de indexação de documentos e busca para ter uma base unificada e auxiliar nas consultas aos dados necessárias para análise.

Utilizou-se a linguagem Python, assim como bibliotecas específicas para processamento de linguagem natural (NLTK e spaCy) e aplicação de algoritmos para aprendizado de máquina (scikit-learn). Para fazer uso das facilidades de uma base unificada, utilizou-se ferramentas da stack ELK: Elasticsearch e Kibana (Figura 2). O Elasticsearch é um motor de busca, onde são indexados os dados coletados e depois enriquecidos com outras informações calculadas durante o processamento, gerando uma base única de consulta. Já o Kibana é um plugin de visualização de dados que consulta, de forma nativa, o Elasticsearch, fornecendo recursos de visualização para análise a partir dos dados indexados.

Como a arquitetura 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.

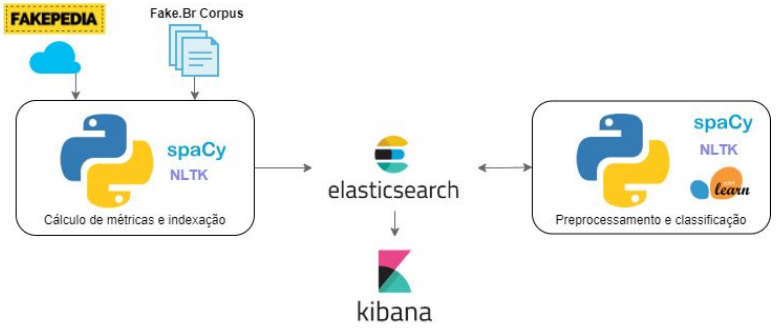

Figura 2: Esquema de arquitetura de desenvolvimento

3.1. 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 variáveis que auxiliem na classificação de notícias, utilizou-se os seguintes dados presentes no corpus: texto e categoria (política, sociedade, TV & celebridades, ciência & tecnologia e religião) da notícia. Informações como link de acesso à matéria, título (quando disponível) e data de publicação podem ser utilizados posteriormente para validações temporais, de fonte ou falsa conexão, 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, entre falsas e verdadeiras, totalizando 12062 notícias para o estudo.

Utilizou-se o teste de Qui-Quadrado, teste estatístico para avaliar a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno. O teste foi realizado para validar o balanceamento de acordo com a proporção de notícias falsas e verdadeiras do estudo. Esse balanceamento é importante para os passos de classificação via aprendizado de máquina implementados no estudo.

Visando balanceamento da base e para se ter uma melhor acurácia com textos completos e com conteúdo suficiente, foram filtradas notícias com menos de 70 palavras. Ainda foram filtradas as notícias falsas com quantidade de palavras menor que 120 com o intuito de deixar a base mais proporcional. Além disso, também é necessária outra forma de saneamento na base, que são as notícias repetidas, existentes tanto na Fakepedia quanto do corpus Fake.br. Como as bases foram geradas por estudos diferentes, podendo ser alimentadas das mesmas fontes de notícias, foi necessário a validação de notícias pertencentes às duas bases.

Os dados provenientes do corpus Fake.Br são categorizados em política, sociedade & cotidiano, TV & celebridades, ciência & tecnologia e religião. Utilizou-se as mesmas categorias nas notícias coletadas através da Fakepedia para manter a padronização. A categorização foi realizada de forma manual.

Aplicados os filtros e saneamento, a quantidade de notícias ficou em 8776 notícias e temos uma melhor proporção entre notícias falsas (51%) e verdadeiras (49%).

4. 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 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 há 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.

Utilizou-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.

5. 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 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.

Tabela 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

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.

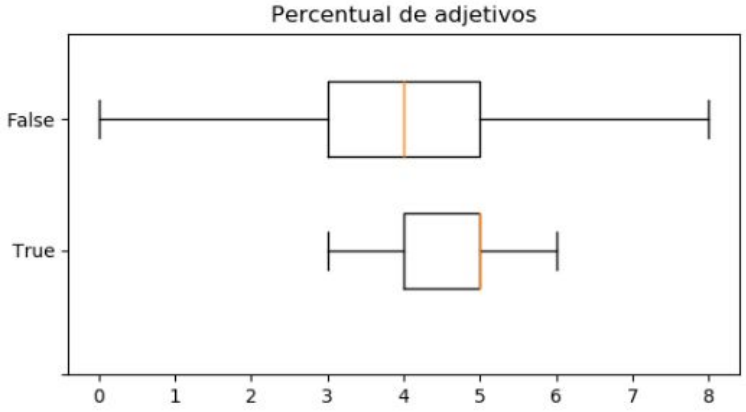

Figura 4: Boxplot com a distribuição de percentuais de adjetivos entre notícias falsas e verdadeiras.

Após análises nos textos, foi criado um classificador com aprendizado de máquina supervisionado para identificar se as 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 com python.

Na classificação, foram contabilizados quatro tipos de resultados:

- 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. Aplicando a classificação com as métricas calculadas, o Gaussian Naive Bayes mostrou acurácia de 84%, o SVM obteve 89% quando executado com seus parâmetros padrão e 94% quando utilizados os parâmetros C=10 e gamma=0.001,

otimizados pelo módulo GridSearchCV, e o AdaBoostClassifier atingiu 93% novamente, tabela 2.

Tabela 2: Resultados dos algoritmos

Algoritmo	Acurácia com texto vetorizado (%)	Acurácia com métricas (%)
AdaBoost	93	93
Árvore de decisão	92	92
Naive Bayes	82	84
SVM	93	94

A classificação utilizando somente as métricas teve tempo de execução muito inferior ao da classificação com os textos vetorizados, mesmo obtendo resultados próximos. Isso se deve ao fato de que a vetorização dos textos geram vetores extremamente grandes, 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 3.

Tabela 3: Resultados da melhor classificação (SVM) com as métricas calculadas

		Valor calculado	
		Verdadeiro	Falso
Valor esperado	Verdadeiro	851	54
	Falso	36	818

6. Discussão

A análise de textos compartilhados em redes sociais em estudos de Stiilpen, 2016, obteve 81% de acurácia para classificação de textos em saúde ou tecnologia, 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 podem remover características inerentes às notícias, como o uso de exclamações. Neste trabalho, os textos não sofrem qualquer alteração em sua escrita.

O trabalho de Monteiro, 2018, 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. Ver comparação dos resultados na tabela 10.

Tabela 10: Acurácia dos algoritmos SVM

	Stiilpen, 2016	Monteiro, 2018	Este estudo
Acurácia	81%	89%	94%

7. Conclusão

Com este trabalho, foi feita uma análise de textos de notícias escritas em português brasileiro com objetivo de analisar a estrutura gramatical das notícias falsas, fazendo um comparativo com notícias verdadeiras. Validamos estudos anteriores e adicionamos novas variáveis para ajudar na identificação de *fake news*.

Nos experimentos realizados, foi verificado que a média percentual de uso de classes gramaticais é próxima entre notícias falsas e verdadeiras, porém, a distribuição das classes gramaticais em *fake news* possui desvio padrão maior do que as mesmas métricas das notícias verdadeiras. Isso denota uma tendência de que notícias falsas possuem estilos de escrita mais diversificados, enquanto que notícias verdadeiras possuem similaridades na sua escrita independente do autor. Isso é compreensível, uma vez que notícias falsas possuem diversas fontes, enquanto as notícias verdadeiras são provenientes de poucos sites quando comparados às outras.

Além dessas características, extraídas após a aplicação das métricas, outras análises foram realizadas com relação ao estilo de escrita de mensagens verdadeiras e falsas. Percebeu-se um maior uso de pontuações de exclamação e de letras maiúsculas nas notícias falsas. Portanto, tais métricas devem ser consideradas na identificação de notícias falsas, já que a presença delas é bem maior.

A análise de sentimento mostrou que ambas classes de notícias possuem polaridade negativa, com as verdadeiras um pouco mais. Também não há diferença nos termos mais utilizados nas notícias. Essa análise deve ser revista contemplando termos de negação do léxico para inversão de sua polaridade, além da consideração de bigramas e utilizando-se de técnicas mais sofisticadas de análise de sentimento.

Como trabalhos futuros, prevê-se a aplicação de algoritmos de regras de associação, como o apriori, para identificar relações entre os percentuais de classes gramaticais. Com o mesmo algoritmo, podemos encontrar relações entre os termos mais utilizados nas duas classes de notícias, como, por exemplo, quais palavras acompanham “governo”, “Brasil” e “país”, para adicionar informações que não tenham sido verificadas com o cálculo de bigramas.

É necessário realizar o estudo considerando as demais categorias de notícias, como ciência e tecnologia, religião, etc, para verificar se os padrões continuam os mesmos ou se as diferenças se tornam mais evidentes. E 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.

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. A partir da obtenção desses textos, é possível aplicar a mesma metodologia verificada neste estudo.

8. Referências

Allcott, H., Gentzkow, (2017) M. Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-36.

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

Charles, A. C., & de Oliveira Sampaio, J. (2018) Checking fake news on web browsers: an approach using collaborative datasets. Workshop on Big Social Data and Urban Computing

Kotsiantis,S.B.; Zaharakis, I; Pintelas, P. (2007) Supervised machine learning: A review of classification techniques. p. 3-24

Monteiro R.A., Santos R.L.S., Pardo T.A.S., de Almeida T.A., Ruiz E.E.S., Vale O.A. (2018) Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In: Villavicencio A. et al. (eds) Computational Processing of the Portuguese Language. PROPOR 2018.

Rossi, M. 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. Acesso em: junho de 2019

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.

Stiilpen Junior, M., Merschmann, L. H. C. (2016) A methodology to handle social media posts in brazilian portuguese for text mining applications. In Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web (pp. 239-246. ACM.

Victor, F.. 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. Acesso em: julho 2019.