Combining compact news representations generated using DistilBERT and topological features to classify fake news

Carlos Abel Córdova Sáenz[1], Marcelo Dias[1,2], and Karin Becker[1]

[1]Universidade Federal do Rio Grande do Sul, Brazil
{cacsaenz, marcelo.dias, karin.becker}@inf.ufrgs.br
[2]Instituto Federal de Educação Ciência e Tecnologia Sul-rio-grandense, Brazil
marcelodias@ifsul.edu

Abstract. Fake news (FN) have affected people's lives in unimaginable ways. The automatic classification of FN is a vital tool to prevent their dissemination and support fact-checking. Related work has shown that FN spread faster, deeper, and more broadly than the truth on social media. Besides, deep learning has produced state-of-the-art solutions in this field, mainly based on textual attributes. In this paper, we propose initial experiments to combine compact representations of the textual news properties generated using DistilBERT, with topological metrics extracted from the social propagation network. Using a dataset related to politics and five distinct classification algorithms, our results are encouraging. Regarding the textual attributes, we reached results comparable to state-of-the-art solutions using only the news title and contents, which is useful for FN early detection. The topological attributes were not as effective, but the promising results encourage the investigation of alternative architectures for their combination.

CCS Concepts: • **Computing methodologies** → **Machine learning**.

Keywords: distilBERT, fake news, fake news classification, topological features

1. INTRODUÇÃO

O fenômeno das *fake news* se acentuou na última década, afetando vários aspectos da vida cotidiana, incluindo política, saúde, educação, entre outros. As redes sociais têm papel ativo neste contexto, pois os mesmos mecanismos de democratização da informação são usados para propagar inverdades. Os efeitos de um boato ou notícia falsa podem ser trágicos, comprometendo a democracia em nível mundial ou afetando a vida das pessoas de formas inimagináveis [Wang 2017].

Atualmente, não há um consenso quanto ao conceito de *fake news* [Zhou and Zafarani 2020], podendo ser definido de forma ampla ou estrita [Shu et al. 2017]. Na interpretação ampla, são consideradas notícias, afirmações, discursos ou postagens em redes sociais, de informação falsa relacionada a figuras públicas e organizações. Esta vertente inclui também trabalhos para a detecção de rumores, sátiras e bots [Bondielli and Marcelloni 2019].

Na definição estrita, adotada por este trabalho, *Fake News* (FN) referem-se a artigos jornalísticos falsos cuja veracidade pode ser verificada e que foram publicados intencionalmente para enganar o consumidor da notícia [Shu et al. 2017]. O conceito enfatiza a autenticidade e a intenção, além de indicar que FN são similares a notícias que seguiram o protocolo jornalístico, dificultando a sua identificação pelos receptores destas.

Identificar informação enganosa não é fácil para humanos [Zhou and Zafarani 2020], e o potencial danoso é tão relevante que muitas iniciativas de checagem de fatos vem sendo desenvolvidas. Tais iniciativas são dirigidas tanto por grupos da grande mídia jornalística individualmente (e.g. Washing-

Esta pesquisa é parcialmente apoiada pelo CNPq (proceso: 131178/2020-2).

ton Post e CNN nos Estados Unidos - EUA, e Folha de São Paulo e Estado de São Paulo no Brasil) ou em consórcio (e.g. Projeto Comprova[1]), quanto por grupos jornalísticos de menor repercussão (e.g. PolitiFact[2], Agência Lupa[3]). Porém, a quantidade em que são produzidas, a velocidade de sua disseminação e a complexidade em se realizar checagem de fatos manualmente levam à necessidade de mecanismos automáticos para o combate às *fake news* [Reis et al. 2019].

A Detecção de FN é a tarefa que tem por objetivo identificar se uma notícia é falsa ou verdadeira. Trabalhos centrados na tarefa de classificação de notícias foram desenvolvidos utilizando abordagens de aprendizado supervisionado de máquina [Bondielli and Marcelloni 2019; Zhou and Zafarani 2020]. Tais abordagens são baseadas no treinamento de classificadores usando dados rotulados e se diferenciam essencialmente pelos algoritmos de aprendizado utilizados (aprendizado raso ou profundo), e pelas *features* exploradas na tarefa, as quais se dividem em *features* relacionadas às notícias propriamente ditas, e ao contexto social de propagação da notícia [Shu et al. 2019].

Features extraídas da notícia (e.g. título, texto e imagem) permitem uma detecção precoce da notícia falsa (*early detection*), i.e. antes que ela se espalhe, já que não dependem da disseminação da notícia em redes sociais. Porém, esta abordagem costuma limitar as soluções ao domínio dos dados de treino utilizados na construção dos modelos preditivos. Um estudo [Reis et al. 2019] argumenta que *features* relacionadas à fonte da notícias e ao engajamento gerado na sua propagação são as mais discriminatórias na classificação de FN. Uma proposta de classificação de meios de comunicação propagadores de FN baseada na topologia da rede de propagação é apresentada em [Pierri et al. 2020]. Propostas de *features* representando o contexto social incluem perfis difusores de notícias em redes sociais [Shu et al. 019a], comportamento social (e.g. *likes*) [Bauskar et al. 2019] e padrões de propagação [Shu et al. 2020].

Os trabalhos relacionados que exploram atributos textuais e Aprendizado Profundo [Shu et al. 2019; Zhou et al. 2020] na classificação de FN têm reportado os melhores resultados. Uma nova tendência em processamento de linguagem natural (PLN) é criar modelos através da transferência de aprendizado de representações de linguagens codificadas usando quantidades massivas de dados, tais como BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al. 2019]. Outra oportunidade é verificar se a abordagem topológica para verificar veículos de comunicação difusores de FN proposta em [Pierri et al. 2020] pode contribuir à classificação de FN independentemente da fonte.

O objetivo deste trabalho é experimentar o uso combinado de conteúdo textual da notícia e da topologia das redes de difusão de notícias para a classificação de FN. Mais especificamente, propomos o uso de DistilBERT [Sanh et al. 2019], uma versão mais leve de BERT, para gerar *features* que representem as notícias de forma compacta. Como contexto social, propomos representar por métricas topológicas (e.g. métricas de conexão entre nodos e de agrupamentos) as propriedades da rede de difusão de cada notícia no Twitter, considerando tweets, retweets e menções. Nossos experimentos foram realizados usando um dataset relacionado a política disponível no FakeNewsNet[4] (FNN) [Shu et al. 2018], usando distintos algoritmos de classificação.

As principais contribuições deste trabalho são: a) experimentos com o uso de Aprendizado Profundo baseado em *fine-tuning* de modelos de representação de linguagens (DistilBERT), ainda pouco exploradas na classificação de FN; b) avaliação da contribuição de *features* topológicas de redes de propagação como atributos representativos de engajamento social, previamente restrita à identificação de veículos de comunicação propagadores de notícias falsas [Pierri et al. 2020]. Nossos resultados foram promissores, mostrando que a classificação de FN baseadas apenas no título e conteúdo da notícia alcança resultados próximos ao estado da arte [Shu et al. 2019; Zhou et al. 2020], que considera também o texto das postagens de propagação. O valor das *features* topológicas foi observado somente em alguns dos algoritmos experimentados, mas os resultados promissores motivam a investigação de arquiteturas alternativas de combinação.

[1]https://projetocomprova.com.br/
[2]https://www.politifact.com/
[3]https://piaui.folha.uol.com.br/lupa/
[4]https://github.com/KaiDMML/FakeNewsNet

O restante deste trabalho está organizado como segue. A Seção 2 apresenta trabalhos relacionados à classificação de FN, destacando as principais propostas que utilizaram o FNN. A Seção 3 descreve a proposta de combinação de *features* extraídas das notícias e de métricas topológicas de difusão das mesmas para a classificação de FN. A Seção 4 detalha os experimentos realizados. A Seção 5 apresenta as conclusões e aponta as pesquisas futuras.

2. TRABALHOS RELACIONADOS

A exploração de FN no contexto de eleições, central na vitória de Donald Trump em 2016 nos EUA e seguida em outros países, motivou um expressivo interesse no tema. *Surveys* como [Zhou et al. 2020; Bondielli and Marcelloni 2019; Shu et al. 2017] contribuem com um arcabouço conceitual e a compilação de importantes trabalhos na área.

Boa parte dos trabalhos recaem no uso de aprendizado de máquina supervisionado para classificação de FN, seja através de algoritmos tradicionais ou de aprendizado profundo. Segundo [Shu et al. 2019], os dois grandes grupos de atributos utilizados para Detecção de FN são *features* extraídas a partir do conteúdo da notícia ou do contexto social. O primeiro envolve características textuais extraídas do título ou texto das notícias (e.g. n-gramas), atributos derivados (e.g. características linguísticas, emoções) ou obtidas a partir de imagens publicadas junto à notícia. O segundo envolve propriedades extraídas do perfil do usuário, padrões de interação social ou de propagação da notícia. Um estudo [Reis et al. 2019] avalia a contribuição de distintos tipos de *features* para a classificação de FN, concluindo que todas contribuem de forma discriminatória, mas que algumas podem ser mais úteis, entre elas as extraídas do engajamento social gerado pela notícia. Usando 5 diferentes algoritmos, reporta resultados de medida-F variando entre 0,75 e 0,81 nos datasets testados.

Abordagens supervisionadas requerem dados rotulados para treino [Zhou and Zafarani 2020], e diversos esforços focaram na construção de conjuntos de dados para este fim, tais como [Wang 2017; Shu et al. 2018]. O presente trabalho faz uso do dataset Politifact, um dos disponíveis no repositório FakeNewsNet (FNN) [Shu et al. 2018]. Para respeitar a política de privacidade do Twitter, o FNN disponibiliza um programa que automatiza o download de dados de notícias (título, texto, URLs das imagens, etc.) e de informações relacionadas ao contexto social (*tweets*, *retweets*, perfis de usuários, *timelines* de usuários, seguidores e seguidos pelos usuários que tuitaram a respeito da notícia).

Os *datasets* disponíveis no FNN permitem a classificação de FN utilizando o conteúdo da notícia, o contexto social ou sua combinação. É possível encontrar mais de uma dezena de propostas publicadas utilizando o dataset Politifact. Destacamos as abordagens dEFEND [Shu et al. 2019], com a melhor medida-F (0,92) usando atributos textuais extraídos de notícias e de postagens, e SAFE [Zhou et al. 2020], com o melhor resultado reportado usando apenas o conteúdo da notícia (0,89).

dEFEND utiliza o conteúdo textual das notícias e dos *tweets* que as mencionaram. Propõe o uso de *encoders* para extrair *features* deste conteúdo e mecanismos de co-atenção (notícia e comentários) com o objetivo de melhorar o desempenho de classificação e de selecionar sentenças que justificam a classificação realizada (*explainability*). Já SAFE explora a similaridade entre o conteúdo textual (título e texto) e imagens da notícia usando redes convolucionais. Estes trabalhos evidenciam o papel relevante do uso de Aprendizado Profundo aplicado ao conteúdo textual das notícias. O aprendizado por transferência a partir do ajuste fino de modelos pré-treinados de representações de linguagens tornou-se uma solução prevalente em PLN, e BERT [Devlin et al. 2019] apresentou soluções estado da arte para diferentes tarefas. DistilBERT [Sanh et al. 2019], baseado em destilação de conhecimento, permite o que este ajuste fino seja baseado em representações de linguagem bem mais compactas, a um custo computacional bem menor.

Para mitigar a dependência do conteúdo do corpus de treino, outras propostas sobre o Politifact propõem o uso do contexto social, tais como perfil do usuário [Shu et al. 019a], reações [Bauskar et al. 2019] e padrões de propagação das notícias [Shu et al. 2020]. Estes trabalhos confirmaram a importância de *features* sobre engajamento social na classificação de notícias [Reis et al. 2019], mas um desempenho melhor é obtido quando combinadas a *features* textuais de notícias [Shu et al.

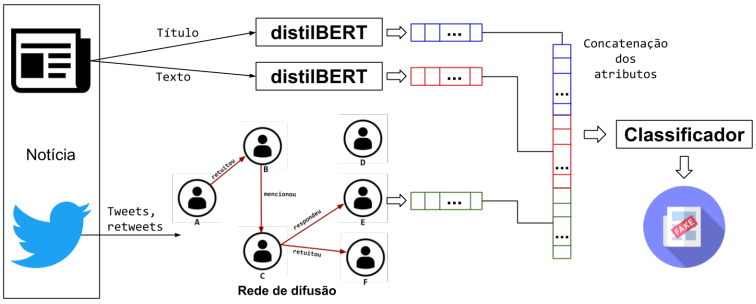

Fig. 1. Arquitetura proposta para a classificação de Fake News

019a; Shu et al. 2020]. Um estudo [Pierri et al. 2020] mostra resultados promissores para classificação automática de veículos de comunicação como *mainstream* ou de desinformação baseado exclusivamente em métricas topológicas da rede de difusão, as quais possuem a vantagem de ser mais difíceis de serem simuladas através de robôs. Esses mesmos atributos são explorados no presente trabalho para a classificação de notícias.

O presente trabalho se diferencia dos relacionados em classificação de FN ao combinar DistilBERT para processamento do conteúdo textual das notícias, com atributos topológicos extraídos da rede de difusão da notícia.

3. EXTRAÇÃO DE *FEATURES* A PARTIR DAS NOTÍCIAS E REDE DE DIFUSÃO PARA A CLASSIFICAÇÃO DE FAKE NEWS

A proposta de classificação de FN avaliada neste trabalho combina: a) representações de notícias baseadas no ajuste fino de modelos de representações de linguagens usando DistilBERT [Sanh et al. 2019], e b) características do contexto social usando métricas topológicas das redes utilizadas na difusão de notícias falsas. Ao contrário de [Shu et al. 2019; Zhou et al. 2020], extraímos *features* textuais apenas do título e/ou texto da notícia. Incluímos o contexto social de forma diferenciada, representado por métricas que caracterizam a topologia da rede social de difusão da notícia, i.e. tweets e retweets onde a URL da notícia esteja presente. A inclusão de propriedades da topologia da rede de difusão permite ter um conjunto de *features* independentes do domínio da notícia, e de difícil reprodução artificial usando robôs.

A Figura 1 esboça a abordagem proposta. Para combinar os dois tipos de *features*, concatenamos os vetores representando cada aspecto em um vetor único, utilizado como entrada de um algoritmo de classificação tradicional.

3.1 Features Textuais da Notícias

Em nosso trabalho utilizamos DistilBERT como um encoder para criar uma representação compacta da notícia. Desta maneira, o título e o texto da notícia, que são originalmente dados não estruturados, são transformados em uma outra representação, estruturada: vetores de números flutuantes que sumarizam o conteúdo textual. DistilBERT é treinado sobre um modelo BERT (bert-base-uncased), resultando em um modelo mais compacto, que pode ser ajustado a custo computacional bem menor, ao mesmo tempo que preserva a quase totalidade das propriedades de compreensão da linguagem do modelo original. Utilizamos em nossos experimentos tanto o título, a notícia, quanto a combinação de ambos, para verificar as propriedades mais relevantes da notícia na detecção de FN. Nossa proposta difere de [Zhou et al. 2020], que busca padrões locais de complexidade crescente usando convoluções, e de [Shu et al. 2019] que usa encoders que alinham notícias e postagens associadas.

Usamos a biblioteca *transformers*[5], que inclui DistilBERT. Sobre o texto bruto do título e do

[5]https://huggingface.co/transformers/

conteúdo da notícia, utilizamos funções para os processos de *tokenization*, *padding* e *masking*. Então, extraímos a representação vetorial dos mesmos com DistilBERT na forma base e em minúsculas (*'distilbert-base-uncased'*).

3.2 Features da rede de difusão

Transpusemos para o contexto de difusão de uma notícia a abordagem topológica proposta em [Pierri et al. 2020] para classificação de veículos de comunicação mainstream ou de desinformação. Para cada notícia, criamos uma rede de difusão usando os tweets e retweets onde a respectiva URL está presente. Nessa rede, os nodos representam usuários que (re)tuitaram a notícia, responderam a esses tweets, ou são mencionados nos mesmos. Pares de nodos são conectados por arestas dirigidas e não ponderadas, sempre que o usuário representado pela origem tenha retuitado, respondido ou mencionado o nodo destino. No exemplo da Figura 1, dado um conjunto de tweets com a URL relativa à notícia i, o usuário C foi mencionado por B, respondeu a um tweet de de E, e retuitou um tweet de F. o usuário D não interagiu com outros usuários no contexto desta difusão. Assim, esta rede representa a forma como as pessoas disseminaram a notícia no Twitter.

Em seguida, calculamos um conjunto de métricas que caracteriza as propriedades topológicas de cada rede de difusão de notícia. Estas métricas representam a complexidade da rede, seu poder de propagacão e a força de conexão e coesão entre os participantes. Buscamos com isto determinar se a maneira em que os usuários interagem entre eles e formam grupos fechados, pode contribuir na detecção de FN. As métricas a serem calculadas, demonstraram em [Pierri et al. 2020] ter conseguido atingir diversos casos nas redes de difusão, como quando os usuários dentro da rede formam grupos, redes onde não existe uma mono-direcionalidade na difusão da notícia ou redes onde existe um único usuário que distribui a notícia entre todos os outros (*broadcast*). Essas métricas são as seguintes:

- Número de componentes fortemente conexos
- Tamanho do maior componente fortemente conexo
- Número de componentes fracamente conexos
- Tamanho do maior componente fracamente conexo
- Diâmetro do maior componente fracamente conexo
- Coeficiente de clusterização
- K-Core

Para calcular os atributos topológicos, carregamos as notícias, com seus tweets e retweets no banco de dados orientado a grafos *neo4j*[6]. Então, construímos e executamos uma consulta para obter a rede de difusão de cada notícia. Finalmente, calculamos cada uma das métricas das redes de difusão. Diferentes bibliotecas existem para este propósito, utilizamos *networkx*[7] neste trabalho.

4. EXPERIMENTOS

4.1 Conjunto de dados

Utilizamos o conjunto de notícias *Politifact* disponível no repositório FNN. Utilizando o programa disponibilizado pelo FNN para extrair os dados, conseguimos coletar 507 notícias verdadeiras e 385 falsas de um total de 1058 notícias disponíveis.

Para construir a rede de difusão das notícias, foram coletados também os respectivos tweets e retweets onde as notícias são referenciadas. Contudo, alguns dos tweets/retweets não puderam ser baixados por razões diversas (e.g. remoção no Twitter). Para evitar a reprodução de redes de difusão não fidedignas, desprezamos em nossos experimentos envolvendo topologias todas as notícias com problemas na coleta de tweets/retweets. O conjunto de dados com contexto social limitou-se a 304

[6]https://neo4j.com/
[7]http://networkx.github.io/

Tabela I. Conjunto de notícias do site Politifact nos experimentos

Tipo	# Notícias verdadeiras	# Notícias falsas	Total
Total no dataset original	624	432	1056
Notícias com conteúdo textual	507	385	892
Notícias com conteúdo textual e rede de difusao	304	355	659

notícias verdadeiras e 355 falsas, com seus tweets e retweets. A Tabela I contrasta a quantidade de notícias do repositório e coletadas em cada caso.

4.2 Objetivos e configuração dos experimentos

Os experimentos realizados tiveram os seguintes objetivos:

(a) Determinar qual propriedade textual da notícia mais contribui à classificação de FN quando compactada usando DistilBERT: título da notícia, texto ou ambos.
(b) Identificar se a combinação das *features* topológicas da rede de difusão da notícia agregam valor à classificação de FN.
(c) Determinar qual algoritmo de classificação, dentro de um conjunto de algoritmos candidatos, produz os melhores resultados.

Escolhemos os algoritmos Regressão Logística (LR), Floresta Aleatória (RF), vizinhos mais próximos (KNN), Máquinas de suporte vetorial (SVM) e Naïve Bayes (NB). Estes algoritmos foram explorados em [Reis et al. 2019], sendo que incluímos a regressão logística no lugar de XgBoost. Os experimentos foram desenvolvidos no ambiente *Python*, utilizando a biblioteca *scikit-learn*[8]. Executamos 10 vezes cada algoritmo sobre o mesmo conjunto de dados, treinando cada modelo segundo a técnica de *cross validation* ($k - fold = 10$). Como métrica de avaliação, utilizamos a medida-F (F1). Os resultados reportados referem-se à média dos valores obtidos no conjunto de execuções.

4.3 Resultados

O primeiro objetivo dos experimentos é comparar o desempenho de classificação de acordo com as propriedades da notícia. Utilizou-se assim todas as notícias recuperadas do repositório, i.e. 892 notícias. Os algoritmos foram treinados utilizando como entrada: a) o vetor correspondendo somente ao título, b) o vetor correspondendo somente ao conteúdo textual da notícia, e c) os dois vetores concatenados. Os resultados podem ser vistos na Figura 2, onde a utilização de DistilBERT sobre as duas *features* textuais (título e texto) apresenta o melhor resultado em quatro dos cinco algoritmos. Nestes algoritmos, o desempenho relacionado ao título apenas, ou texto apenas, são similares ou comparáveis. Os algoritmos com melhor performance foram: regressão logística ($F1 = 0,90$), Floresta Aleatória ($F1 = 0,89$) e SVM ($F1 = 0,88$). Estes experimentos (ver Figura 2) revelaram uma abordagem de requisitos mínimos (título e texto da notícia), que permitem uma detecção precoce de FN (*early detection*), extremamente competitiva por vários motivos. Primeiro, obteve resultado próximo ao estado da arte representado por dEFEND [Shu et al. 2019], que depende da propagação da notícia em rede social. Segundo, quando comparada à abordagem de requisito similar (SAFE) [Zhou et al. 2020], que utiliza apenas o conteúdo da notícia (textual e visual), apresenta resultados superiores aos reportados. E, por fim, destaca-se o resultado que, com o simples uso do título como *feature* para classificação, já apresenta desempenho comparável aos *baselines*.

O segundo experimento foi realizado com a intenção de avaliar os benefícios da combinação das *features* topológicas com as textuais. Por esta razão, foi realizado utilizando apenas o conjunto de notícias para os quais conseguimos os respectivos tweets e retweets (i.e. 659 notícias), necessários à construção das redes de difusão. Analisamos a combinação das *features* topológicas com a textuais considerando somente o título, o conteúdo da notícia ou ambos. Como comparação, reproduzimos os resultados de cada algoritmo utilizando DistilBERT aplicado somente às *features* textuais (título e texto) destas mesmas notícias. Os resultados obtidos são apresentados no gráfico da Figura 3.

[8]https://scikit-learn.org/

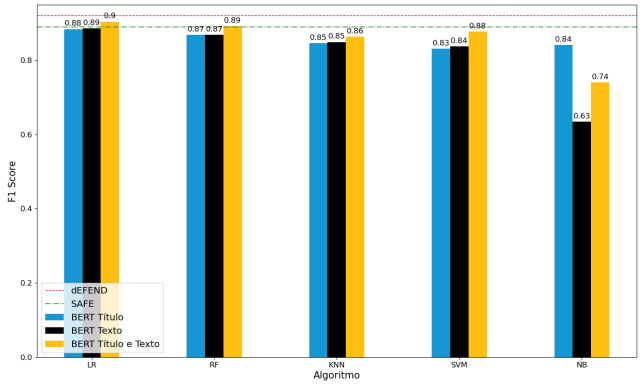

Fig. 2. F1 Score da classificação das notícias usando BERT e o título e texto, com diversos algoritmos de classificação

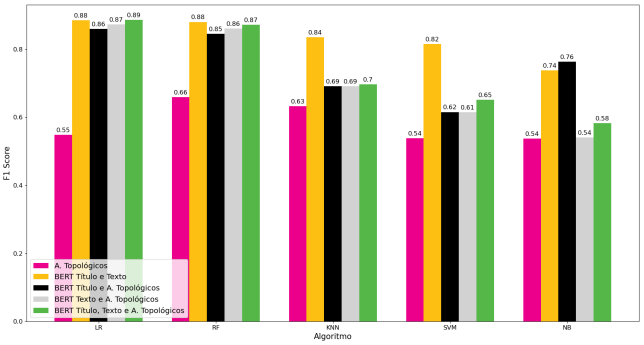

Fig. 3. F1 Score da classificação das notícias incluindo as *features* topológicas, com diversos algoritmos de classificação

É possível verificar que as *features* topológicas sozinhas conseguem classificar as notícias, mas com desempenho bem inferior (F1 variando de 0,54 a 0,66). Comparando com os resultados reportados na Figura 2, também observamos que o menor conjunto de treino afetou o desempenho da abordagem baseada em DistilBERT, que reduziu a medida-f em até 4 pontos percentuais.

Os melhores resultados foram obtidos combinando as *features* topológicas com as representações vetoriais do título e do texto da notícia. O algoritmo com melhor desempenho foi de regressão logística ($F1 = 0,89$), seguido de Floresta Aleatória ($F1 = 0,87$). Estes resultados são superiores ou similares àqueles produzidos pelas *features* textuais apenas ($F1 = 0,88$). Para os demais algoritmos, os resultados foram inferiores. Mesmo combinados de forma muito simples (concatenação de vetores), observamos que estes atributos têm potencial de representação do engajamento social dos usuários a respeito da notícia. Os resultados obtidos com os algoritmos de regressão logística e random forest foram promissores, e nos motivam investigar outras formas de combinar estes dois tipos de *features* (e.g. soluções de *stacking* ou arquiteturas alternativas de Aprendizado Profundo).

Finalmente em relação aos algoritmos, os melhores resultados foram produzidos pela regressão logística, seguidos de perto pelos gerados com Floresta Aleatória. Naïve Bayes e SVM apresentaram os piores desempenhos. Os desempenhos com estes algoritmos estão consistentes com os reportados em [Reis et al. 2019].

5. CONCLUSÃO E TRABALHOS FUTUROS

O presente trabalho propôs um processo de classificação de FN baseado na extração de *features* do conteúdo de notícias (título e texto) usando DistilBERT e do grafo de difusão formado pelos disseminadores em rede social. Foram exploradas combinações destas *features* e cinco algoritmos de

classificação distintos.

Gerando *features* com o uso de DistilBERT apenas sobre os atributos textuais das notícias, conseguimos resultados comparáveis ao estado da arte [Shu et al. 2019; Zhou et al. 2020], e superiores à maioria dos trabalhos que utilizaram o conjunto de dados Politifact (e.g. [Shu et al. 2020; Papanastasiou et al. 2019]. Dentre estes trabalhos, encontram-se tanto abordagens de requisito mínimo (apenas conteúdo da notícia) [Zhou et al. 2020], que são aplicáveis em contexto de detecção precoce da notícia, quanto abordagens que extraem informações da respectiva rede de difusão [Shu et al. 2019; Shu et al. 2020; Papanastasiou et al. 2019]. Note-se que mesmo aplicada somente ao título da notícia, a abordagem proposta atinge desempenho muito bom, denotando a capacidade discriminatória desta *feature*. Por outro lado, o uso das *features* topológicas se mostrou mais limitado, apesar de apresentar potencial discriminatório em alguns algoritmos. Contudo, mostrou seu potencial e oportunidades de combinações mais sofisticadas que a simples concatenação.

Os resultados deste trabalho levam à evolução desta proposta inicial para a detecção de FN, através de: a) investigação de formas alternativas de combinação de features, tais como stacking ou aprendizado profundo; b) análise de métricas representativas da topologia de difusão de notícias (e.g. centralidades), e c) o estudo de *features* extraídas de imagens obtidas do conteúdo da notícia combinadas às já estudadas neste trabalho.

REFERÊNCIAS

Bauskar, S., Badole, V., Jain, P., and Chawla, M. Natural Language Processing based Hybrid Model for Detecting Fake News Using Content-Based Features and Social Features. *International Journal of Information Engineering and Electronic Business* 11 (4): 1–10, 2019.

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, J. Burstein, C. Doran, and T. Solorio (Eds.). pp. 4171–4186, 2019.

Papanastasiou, F., Katsimpras, G., and Paliouras, G. Tensor factorization with label information for fake news detection. *arXiv preprint arXiv:1908.03957*, 2019.

Pierri, F., Piccardi, C., and Ceri, S. Topology comparison of twitter diffusion networks effectively reveals misleading information. *Scientific Reports* 10 (1), Jan, 2020.

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. Supervised learning for fake news detection. *IEEE Intelligent Systems* 34 (2): 76–81, 2019.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. KDD '19. Association for Computing Machinery, New York, NY, USA, pp. 395–405, 2019.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286* vol. 8, 2018.

Shu, K., Mahudeswaran, D., Wang, S., and Liu, H. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. pp. 626–637, 2020.

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

Shu, K., Zhou, X., Wang, S., Zafarani, R., and Liu, H. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. pp. 436–439, 2019a.

Wang, W. Y. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

Zhou, X., Wu, J., and Zafarani, R. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981*, 2020.

Zhou, X. and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* 0 (ja), 2020.