Supervised Learning for Misinformation Detection in WhatsApp

Julio C. S. Reis
Universidade Federal de Viçosa
Viçosa, Minas Gerais, Brasil
jreis@ufv.br

Fabrício Benevenuto
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
fabricio@dcc.ufmg.br

ABSTRACT

WhatsApp created a new channel for smartphone users to consume and share news. The easiness to create groups of people that partake similar interests and share content has made WhatsApp prone to abuse by misinformation campaigns. Although fact-checking is very effective for detecting misinformation, it cannot keep up with the sheer volume of information that is now generated online. In this context, we investigate the potential of automatic approaches based on supervised machine learning as a support tool to help fact-checkers identify misinformation shared through images on WhatsApp. Our results show that the predictive performance of the investigated approaches has a useful degree of discriminative power to detect misinformation. Finally, we discussed how WhatsApp misinformation detection approaches can be used in practice, highlighting challenges and opportunities.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Applied computing** → *Sociology*.

KEYWORDS

Supervised Learning, WhatsApp, Misinformation Detection

ACM Reference Format:
Julio C. S. Reis and Fabrício Benevenuto. 2021. Supervised Learning for Misinformation Detection in WhatsApp. In *Brazilian Symposium on Multimedia and the Web (WebMedia '21), November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3470482.3479641

1 INTRODUÇÃO

O WhatsApp, um aplicativo de mensagens instantâneas que conecta mais de dois bilhões de usuários em todo o mundo[1], mudou drasticamente a maneira como as pessoas consomem e compartilham notícias. Uma pesquisa recente divulgada pelo *Reuters Institute* mostrou que em países como Brasil, Malásia e África do Sul, o WhatsApp já se tornou a principal plataforma para discussão e compartilhamento de notícias [21]. No entanto, a facilidade para criação de grupos pelos usuários bem como a simplicidade para compartilhamento de informações tornaram o WhatsApp uma plataforma vulnerável para abusos por campanhas de desinformação [18]. Além disso, o uso frequente do WhatsApp para socialização entre usuários também favorece esse cenário. O *Reuters Institute* [21] também mostrou em sua pesquisa que 76% dos usuários do WhatsApp participam de grupos de discussão nesta plataforma. Os autores mostraram ainda que 58% destes usuários participam de grupos que incluem pessoas desconhecidas. Particularmente, isso é evidente no Brasil, onde o WhatsApp foi apontado como uma das principais mídias para divulgação de desinformação durante as eleições de 2018[2], consolidando-se como uma plataforma bastante utilizada para distribuição de mensagens políticas em massa[3].

Neste contexto, uma maneira eficiente de se detectar desinformação disseminada em plataformas digitais como o WhatsApp é a checagem de fatos, ou seja, a realização de uma avaliação da veracidade de uma notícia ou afirmação [34]. Esta tarefa, normalmente realizada por jornalistas especializados, verifica a exatidão das informações comparando-as com uma ou mais fontes confiáveis [20]. No entanto, a checagem de fatos é um processo árduo que normalmente demanda uma análise bastante detalhada para apoiar o veredito fornecido pelos checadores de fatos [34]. Consequentemente, a checagem de fatos tradicional não consegue acompanhar o enorme volume de informações que são gerados diariamente no ambiente online [7]. Especificamente em relação ao WhatsApp, ainda há um desafio adicional: os canais de comunicação estabelecidos são descentralizados e em sua maioria privados, por padrão. Assim, não é possível descobrir facilmente quais os assuntos são discutidos neste ambiente, e os checadores de fatos geralmente precisam de apoio para identificar potenciais demandas de verificação. Nesse cenário, as soluções automáticas para detecção de desinformação em plataformas digitais podem ser usadas como uma ferramenta auxiliar à checagem de fatos apoiando, por exemplo, o processo de identificação de um conteúdo com maior probabilidade de ser falso ou ainda um conteúdo que precise ser checado, porém ainda deixando a decisão final (i.e., o veredito) para um especialista no término deste processo.

Dada a relevância do problema da desinformação em plataformas digitais, surgiram vários esforços de pesquisa com o objetivo de compreender o fenômeno para proposição de soluções. De forma geral, esses trabalhos identificam padrões e atributos típicos deste tipo de conteúdo para propor abordagens automáticas que sejam capazes de detectar desinformação [8, 31, 35, 37]. Apesar da inegável importância dos esforços existentes nesta direção, eles são, em sua maioria, trabalhos concorrentes que identificam padrões recorrentes da desinformação no contexto político americano ou que, isoladamente, propõem atributos para o treinamento de classificadores

[1] https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately

[2] https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html

[3] https://www.bbc.com/news/technology-45956557

WebMedia '21, November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil

ACM ISBN 978-1-4503-8609-8/21/11…$15.00
https://doi.org/10.1145/3470482.3479641

a partir de ideias que não foram testadas em conjunto. Assim, é difícil avaliar o potencial prático que abordagens supervisionadas treinadas a partir de atributos propostos em estudos recentes têm para detectar desinformação, principalmente no cenário político brasileiro.

Diante disso, neste trabalho nós exploramos um conjunto de dados do WhatsApp construído durante a última eleição presidencial brasileira - a saber, 2018, um período bastante marcado pela disseminação de desinformação[4]. Primeiro, nós pesquisamos trabalhos relacionados e recentes como uma tentativa de implementar atributos propostos na literatura para a detecção de desinformação. No total nós implementamos 181 atributos para detecção de desinformação, incluindo alguns propostos neste trabalho. Destacamos que, explorar esses atributos e seu potencial discriminatório é também uma tentativa de entendermos melhor características únicas de desinformação disseminadas no WhatsApp no contexto político brasileiro. Em seguida, nós avaliamos o desempenho de previsão de abordagens clássicas de aprendizado de máquina supervisionado na realização desta tarefa. Nossos resultados experimentais revelam que os atributos implementados combinados com classificadores existentes possuem um grau útil de poder discriminativo para a tarefa de detecção de desinformação, podendo ser utilizados por checadores de fatos em eleições futuras, como uma ferramenta auxiliar no processo de identificação de um conteúdo com maior probabilidade de ser falso.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta definições importantes enquanto a Seção 3 discute trabalhos relacionados. Em seguida, na Seção 4, é apresentada a metodologia experimental adotada no trabalho, incluindo uma breve descrição do conjunto de dados utilizado e dos atributos implementados para detecção de desinformação. Depois, os resultados experimentais são apresentados na Seção 5. Por fim, a Seção 6 conclui o trabalho e apresenta direcionamentos para trabalhos futuros.

2 DEFINIÇÕES

Desinformação é um tópico que ainda carece de uma definição clara que seja universalmente aceita. Diante disso, apresentamos brevemente as definições relacionadas ao termo que foram utilizadas neste trabalho.

Definição 2.1. ***(Desinformação)*** *"Uma notícia ou mensagem publicada e propagada pela mídia, contendo informações falsas, independentemente dos meios e motivos que a embasam" [30].*

Definição 2.2. ***(Detecção de Desinformação.)*** *Dada uma notícia não rotulada $a \in \mathcal{A}$, um modelo para detecção de desinformação atribui uma pontuação $S(a) \in [0, 1]$ indicando até que ponto se acredita que a contenha desinformação. Por exemplo, se $S(a') > S(a)$, de acordo com o modelo é mais provável que a' contenha desinformação em comparação com a. Neste cenário, um limite τ pode ser definido de forma que a função de previsão $F : \mathcal{A} \rightarrow \{$desinformação, conteúdo não verificado$\}$[5] seja:*

$$F(a) = \begin{cases} \textit{desinformação} & \textit{if } S(a) > \tau, \\ \textit{conteúdo não verificado} & \textit{caso contrário.} \end{cases}$$

3 TRABALHOS RELACIONADOS

De forma geral, existem dois tipos de esforços que investigam o problema da desinformação em plataformas digitais. O primeiro grupo está focado em prover uma melhor compreensão do fenômeno[13, 36]. Por exemplo, Vosoughi *et al.* [36] mostra que a desinformação tende a se espalhar mais rapidamente do que informações verdadeiras disseminadas em mídias sociais. Por outro lado, Lazer *et al.* [13] ressalta que este problema, devido à sua complexidade, deve ser abordado de forma interdisciplinar.

O segundo grupo de esforços existentes compreende aqueles que propõem soluções para o problema ou fornecem insights sobre como detectar desinformação disseminada em plataformas digitais. Particularmente, esses trabalhos discutem padrões típicos de desinformação que podem ser usados como atributos para treinamento de classificadores para detecção deste tipo de conteúdo. Por exemplo, Pérez-Rosas *et al.* [22] apresenta um conjunto de experimentos envolvendo aprendizagem de máquina com o objetivo de construir detectores de desinformação precisos com base em atributos linguísticos de uma notícia. De forma similar, Volkova *et al.* [35] constrói modelos linguísticos para classificar um conjunto de notícias como suspeitas ou confiáveis.

Em suma, a maioria dos esforços existentes neste espaço são trabalhos simultâneos que usam dados específicos e conjuntos de atributos propostos para treinar classificadores sem fornecer diretrizes claras sobre quais atributos são úteis para detectar desinformação. Neste contexto, este trabalho faz um levantamento desses estudos existentes que abordam o tema aqui explorado, identificando os principais atributos propostos para essa tarefa com o objetivo de testar o poder discriminativo dos mesmos de forma combinada. Além disso, embora existam outras iniciativas que explorem o fenômeno da desinformação no contexto brasileiro [14, 15, 19], até onde sabemos, este é o primeiro trabalho que investiga o potencial de abordagens supervisionadas para detectar desinformação disseminada por meio de imagens no WhatsApp.

4 METODOLOGIA

Nesta seção descrevemos o conjunto de dados utilizado neste trabalho bem como o processo de extração de atributos para detecção de desinformação disseminada no WhatsApp. Por fim, uma análise de importância dos atributos implementados é apresentada.

4.1 Conjunto de Dados do WhatsApp

O WhatsApp é um aplicativo de mensagens instantâneas bastante popular no Brasil [21], porém ainda é pouco explorado em pesquisas. Devido ao grande volume de informações que são compartilhadas diariamente neste aplicativo e à sua natureza fechada (adoção de criptografia de ponta-a-ponta), obter dados relevantes desta plataforma por si só já é um desafio. É difícil rastrear os dados compartilhados neste ambiente e prover informações adicionais relacionadas

[4] https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html

[5] Conforme também discutido na Seção 4.1, para o conjunto de dados do WhatsApp utilizado neste trabalho nós temos instâncias rotuladas como "desinformação" e informações que não foram verificadas por checadores de fatos. Assim, nós nos referimos a este último grupo como "conteúdo não verificado" (ao invés de "não contém desinformação", por exemplo), uma vez que a veracidade do mesmo não foi necessariamente verificada por uma agência de checagem de fatos.

aos mesmos. Assim, para investigar o potencial de abordagens automáticas de aprendizado de máquina na tarefa de identificar desinformação disseminada no WhatsApp nós exploramos o conjunto de dados construído em [27]. Este conjunto de dados compreende mensagens disseminadas por meio de imagens no WhatsApp durante o período eleitoral brasileiro de 2018. Ele contém mensagens coletadas de um grande número de grupos públicos com foco em política bem como metadados relativos à divulgação das mesmas, incluindo grupos, identificadores anonimizados dos usuários responsáveis por cada uma das postagens, e informações de data e hora de divulgação das mensagens.

A Tabela 1 apresenta uma visão geral do conjunto de dados do WhatsApp explorado neste trabalho. Ele é composto por 4.524 imagens distintas[6] compartilhadas no WhatsApp entre agosto e outubro de 2018. Essas imagens foram disseminadas em 414 grupos únicos e foram compartilhadas por 17.465 usuários distintos.

No entanto, para fazer contribuições relevantes no campo de detecção de desinformação é importante obter um conjunto de dados de alta qualidade que tenha sido rotulado por anotadores (i.e., jornalistas) com experiência de domínio. Assim, inspirados pela abordagem proposta em [27] nós propusemos uma estratégia que fosse capaz de identificar, dentre o conjunto de mensagens analisado, aquelas que continham desinformação. De forma geral, nós utilizamos a API do *Google Vision*[7] para obter as páginas Web nas quais cada uma das imagens também apareceram. Para cada resultado, nós verificamos, automaticamente, se a página retornada era de uma agência de checagem de fatos brasileira[8]. Caso positivo, a imagem foi rotulada automaticamente conforme veredito fornecido pela agência de checagem. Essa informação foi recuperada a partir de um parser do html da referida página.

Para garantir a qualidade dos dados, nós verificamos manualmente se todas as imagens contendo desinformação (que foram rotuladas automaticamente pela metodologia proposta apresentada anteriormente) correspondiam àquelas presentes em nossos dados do WhatsApp. Nosso conjunto de dados final é composto 135 imagens distintas que foram rotuladas como "desinformação"[9]. Em suma, o nosso conjunto de dados é composto por notícias disseminadas no WhatsApp por meio de imagens que foram rotuladas como "desinformação" (i.e., 135 instâncias) e notícias não checadas (i.e., 4389 instâncias), uma vez que para este último grupo a veracidade de seu conteúdo não foi necessariamente verificada.

A seguir, apresentamos uma breve descrição dos atributos implementados a partir desse conjunto de dados e mostramos que eles podem ser úteis para distinguir desinformação de conteúdo não verificado.

[6]É válido ressaltar que optamos por filtrar apenas mensagens que divulgam informações por meio de imagens uma vez que esforços anteriores mostraram que as imagens são o tipo de conteúdo de mídia mais frequente do WhatsApp, bem como uma importante fonte de desinformação [27].

[7]https://cloud.google.com/vision

[8]Para este trabalho foram consideradas as seguintes agências de checagem de fatos: aosfatos.org, boatos.org, g1.globo.com/e-ou-nao-e/, piaui.folha.uol.com.br/lupa/, e-farsas.com, veja.abril.com.br/blog/me-engana-que-eu-posto/.

[9]Os dados rotulados explorados neste trabalho estão publicamente disponíveis no seguinte link: http://doi.org/10.5281/zenodo.3779157 [24].

4.2 Extração de Atributos

Trabalhos anteriores mostraram que os atributos utilizados para detecção de desinformação podem ser divididos em 3 grupos principais [25]: *(i)* atributos extraídos do *conteúdo* (ex.: propriedades textuais e de imagem); *(ii)* atributos extraídos da *fonte* da informação (ex.: informações do editor, de quem publicou a informação, aspectos de credibilidade); e *(iii)* atributos extraídos do *ambiente*, que geralmente envolvem dinâmicas de propagação em plataformas digitais e na Web como um todo. A Tabela 2 apresenta um resumo dos principais atributos para identificação de desinformação que implementamos neste trabalho. Esses atributos estão organizados conforme categorização apresentada. No total, nós computamos 181 atributos para detecção de desinformação. Particularmente, além dos atributos comumente explorados na realização desta tarefa por trabalhos anteriores, nós propusemos novos atributos que foram extraídos a partir de propriedades da imagem, estrutura semântica do conteúdo bem como atributos específicos acerca da fonte da informação e novas medidas de propagação interna e externa ao WhatsApp. Esses atributos propostos, até onde sabemos, ainda não foram explorados na identificação de desinformação disseminada nesta plataforma digital. A seguir, nós descrevemos sucintamente os atributos implementados em cada uma das categorias acima mencionadas.

4.2.1 Conteúdo. Os atributos extraídos a partir do conteúdo envolvem não apenas a notícia, mas também seu título, imagens e qualquer mensagem de endosso associada. Aqui, como consideramos apenas notícias divulgadas no WhatsApp por meio de imagens, inicialmente nós utilizamos a API do *Google Vision* para extrair várias informações associadas à elas, como por exemplo, **propriedades da imagem**, que incluem rótulos da imagem, número de cores e objetos associados à elas bem como informações relativas à presença ou não de faces. Também usamos como atributos as probabilidades de conteúdo explícito (ex.: adulto, paródia, médico, violência, etc) fornecidas por esta mesma API.

Além disso, nós extraímos o texto das imagens usando o reconhecimento óptico de caracteres (OCR) também fornecido pela API do *Google Vision*. A partir disso, nós calculamos diversos atributos textuais para detecção de desinformação propostos em trabalhos anteriores, como **atributos sintáticos** (ex.: métricas textuais em nível de sentença, incluindo número de palavras e sílabas por frase, indicadores de qualidade do texto a partir de métricas de legibilidade), e **atributos lexicais** como caracteres e medidas em nível de palavra (ex.: classe gramatical, uso de pontuações, etc), dentre outros.

Também usamos a versão 2015 do *Linguistic Inquiry and Word Count* (LIWC) [32] para extrair e analisar a distribuição de **atributos psicolinguísticos**[10] do texto incluído e associado a uma notícia divulgada por meio de uma imagem no WhatsApp. Desde a sua concepção, o LIWC tem sido amplamente utilizado para uma série de tarefas diferentes, incluindo caracterização do discurso de ódio [10] em plataformas de mídia social e detecção de notícias falsas [25]. Em seguida, nós extraímos aspectos da **estrutura semântica** do texto,

[10]LIWC é um programa de análise de texto que categoriza palavras em categorias derivadas de gramática e psicologia [5].

Tabela 1: Visão geral do conjunto de dados do WhatsApp.

	#Usuários	#Grupos	Imagens Únicas	#Desinformação	Período
Brasil	17.465	414	4.524	135	2018/08 - 2018/10

Tabela 2: Visão geral dos atributos implementados para detecção de desinformação.

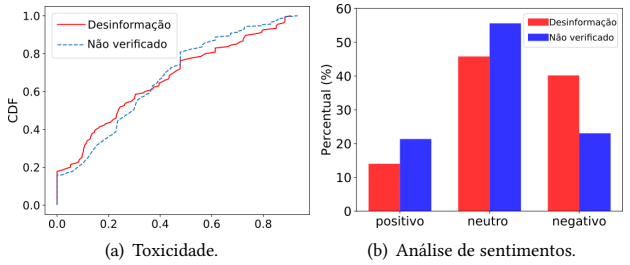
Extraído do(a)...	Grupo de Atributos	Descrição Geral (Exemplos)	Total
Conteúdo (cont)	Propriedades da Imagem (prop img)	Número de rostos em uma imagem, rótulos, cores, objetos, etc	9‡
	Atributos Sintáticos (sint)	Atributos em nível de sentença, indicadores de qualidade do texto (ex.: métricas de legibilidade), etc	31
	Atributos Lexicais (lexi)	Atributos em nível de caracteres e palavras, incluindo número de palavras, pronomes, verbos, indicadores do uso de hashtags, pontuações, etc	49
	Atributos Psicolinguísticos (psico)	Sinais adicionais de linguagem persuasiva, como raiva, tristeza, etc. e indicadores de linguagem tendenciosa	38
	Estrutura Semântica (seman)	Rótulos, informações contextuais, medição de toxicidade do texto	8⋆‡
	Subjetividade (subj)	Medidas de subjetividade e análise de sentimentos	4
Fonte (font)	Editor (edit)	Usuário responsável pelo primeiro compartilhamento da mensagem, e grupos onde a mensagem foi disseminada	5⋆‡
	Viés (vies)	Alinhamento político (ex.: direita, esquerda, centro)	3
Ambiente (amb) (WhatsApp e Web)	Engajamento Interno (WhatsApp) (eng int)	Número de compartilhamentos, número de usuários distintos que postaram uma mesma mensagem, e número de grupos distintos onde uma mesma mensagem foi postada	3‡
	Propagação Externa (Web) (prop ext)	Informações relativas ao espalhamento de uma imagem fora do WhatsApp (na Web)	5‡
	Padrões Temporais (temp)	A taxa na qual compartilhamentos são feitos internamente na plataforma para diferentes janelas de tempo	26

A presença de ⋆ indica a existência de atributos categóricos no referido grupo, enquanto a presença de ‡ destaca a existência de atributos novos ou que não foram previamente explorados na tarefa de detecção de desinformação em plataformas digitais e especificamente no contexto do WhatsApp.

incluindo informações contextuais e indicadores de toxicidade[11]. Por último, com base na relação entre os aspectos de popularidade e subjetividade das notícias [23] também computamos medidas de **subjetividade** e análise de sentimentos[12].

Como ilustração do potencial desses atributos na tarefa de identificação de desinformação disseminada no WhatsApp, a Figura 1(a) mostra a distribuição cumulativa (CDF) da pontuação de toxicidade em cada conjunto de mensagens, ou seja, desinformação e conteúdo não verificado. No geral, podemos observar que as imagens contendo desinformação tendem a ser ligeiramente mais tóxicas do que o conteúdo não verificado. Especificamente, cerca de 30% das imagens contendo desinformação possuem toxicidade superior a 0,5, enquanto menos de 20% do conteúdo não verificado apresentam tal pontuação de toxicidade. Ademais, cerca de 10% das imagens contendo desinformação apresentam um conteúdo altamente tóxico com uma pontuação acima de 0,8. Por fim, a Figura 1(b) apresenta as porcentagens de mensagens positivas, neutras e negativas para desinformação e conteúdo não verificado disseminado no WhatsApp. É interessante notar que, apesar do grande volume de mensagens neutras em ambos os grupos de mensagens, os resultados sugerem um viés mais forte para um discurso mais negativo relacionado à desinformação divulgada no WhatsApp, corroborando resultados identificados em estudos anteriores [26].

De forma geral, nossas análises acerca de atributos extraídos a partir do conteúdo revelam que imagens contendo desinformação disseminadas no WhatsApp apresentam características peculiares que podem ser úteis para identificá-las.



(a) Toxicidade. (b) Análise de sentimentos.

Figura 1: Distribuição cumulativa (CDF) da pontuação de toxicidade e sentimento associados às mensagens disseminadas no WhatsApp.

4.2.2 Fonte. Primeiro, consideramos o identificador (único) anonimizado do usuário que compartilhou uma notícia pela primeira vez no WhatsApp como um atributo categórico. De maneira similar, capturamos o primeiro grupo do WhatsApp em que a mensagem foi postada. Em uma análise preliminar, descobrimos que apenas 10 entre os mais de 17 mil usuários únicos foram responsáveis pela primeira postagem de 23% das imagens que continham desinformação, e que 9 de 414 grupos concentram quase metade (44%) das primeiras aparições dessas imagens que contém desinformação. Presumimos que essas estatísticas fornecem informações valiosas que sejam capazes de capturar uma possível indicação de ação maliciosa e orquestrada para espalhamento intencional de desinformação.

Além disso, esforços anteriores mostram que há uma correlação entre a polarização política e a disseminação de desinformação [29]. Assim, nós inferimos os **vieses** políticos dos grupos do WhatsApp de acordo com a seguinte estratégia: (1) analisamos automaticamente a descrição do grupo (ou seja, o nome do grupo) para verificar

[11] A *Perspective API*, disponível em https://www.perspectiveapi.com utiliza modelos de aprendizado de máquina para quantificar até que ponto um texto pode ser percebido como "tóxico" e tem sido explorada por um número significativo de trabalhos recentes [11].

[12] Para a análise de sentimentos do conteúdo disseminado no WhatsApp, nós usamos uma versão em português do método SentiStrength [33], disponível em sentistrength.wlv.ac.uk.

se havia alguma informação disponível sobre o seu viés político. Nesse caso, rotulamos o grupo como "direita ", "esquerda " ou "centro".

Por exemplo, o grupo "# BOLSONARO PRESIDENTE" foi rotulado como "direita", uma vez que Jair Bolsonaro, então candidato à presidência em 2018, é um partidário de direita. Para os casos em que a descrição do grupo não forneceu qualquer indicação de seu alinhamento político, nós (2) inspecionamos manualmente o conteúdo do grupo, ou seja, o viés das mensagens nele postadas, a fim de inferir o viés político do mesmo. Esta estratégia foi usada em estudos anteriores para quantificar os vieses de uma determinada fonte [3, 28]. Para os casos em que o conteúdo de ambos os vieses (i.e. direita e esquerda) foi compartilhado em um mesmo grupo, nós o rotulamos como "centro". Em uma análise inicial, constatamos que durante o período eleitoral brasileiro de 2018 os grupos de direita foram mais ativos na disseminação de conteúdo no WhatsApp. Além disso, uma imagem postada nesses grupos tem mais probabilidade de estar associada a uma história falsa do que uma mensagem postada em grupo com alinhamento político diferente de "direita", potencialmente devido ao desequilíbrio entre desinformação e o conteúdo não verificado. Esse resultado corrobora estudos anteriores que mostram que grupos de direita são mais eficazes no uso da ferramenta de mídia social para divulgação de notícias, desinformação e opiniões [4].

4.2.3 Ambiente. Alguns recursos podem ser extraídos do ambiente, como métricas de engajamento do usuário e estatísticas relativas à dinâmica de propagação do conteúdo. Assim, neste trabalho, calculamos algumas medidas de **engajamento interno** ao WhatsApp como por exemplo, o número de usuários distintos que postaram a mesma notícia por meio de uma imagem na plataforma, o número de grupos distintos em que a mesma notícia foi postada e o número total de compartilhamentos da mesma notícia em todos os grupos analisados, tanto para desinformação quanto para mensagens com conteúdo não verificado.

A Figura 2 mostra as distribuições cumulativas (CDFs) dessas medidas. Conforme apresentado na Figura 2(a), cerca de 60% das imagens contendo desinformação foram compartilhadas por até 11 usuários, enquanto a mesma fração de conteúdo não verificado foi compartilhada por um número menor de usuários (até 5). Além disso, é possível percebermos que as imagens contendo desinformação tendem a atingir um número muito maior de grupos distintos. A Figura 2(b) indica que 40% das imagens contendo desinformação atingiram mais de 10 grupos distintos. Em contraste, apenas 20% do conteúdo não verificado foi compartilhado em mais de 8 grupos. Finalmente, de acordo com a Figura 2(c), a distinção entre desinformação e conteúdo não verificado é drástica quando se trata do número total de compartilhamentos. Aproximadamente 60% do conteúdo não verificado foi compartilhado mais de uma vez e apenas 5% deste foi compartilhado mais de 10 vezes. Por outro lado, 80% das imagens contendo desinformação foram compartilhadas mais de uma vez, e mais da metade delas foram compartilhadas mais de 10 vezes.

Adicionalmente, nós também computamos medidas de **propagação externa** ao WhatsApp, ou seja, informações sobre a divulgação dessas notícias (i.e., imagens) disseminadas no WhatsApp em páginas da Web. Para fazer isso, usamos as informações sobre sites com imagens correspondentes da API do *Google Vision*. Este serviço retorna informações sobre páginas Web que contêm imagens idênticas à uma imagem fornecida como entrada. Do conjunto de sites/domínios que publicaram as imagens presentes em nosso dado na Web, nós medimos o volume de páginas ainda disponíveis, páginas incomuns[13] e links seguros (i.e., https).

A Figura 3 apresenta a distribuição cumulativa (CDF) do número de sites e domínios únicos onde as imagens compartilhadas no WhatsApp também foram disseminadas. Primeiramente, observamos uma tendência: as imagens contendo desinformação que foram compartilhadas no WhatsApp também foram mais divulgadas na Web como um todo. Usando o teste de Kolmogorov [16], descobrimos que a diferença entre as distribuições é estatisticamente significativa (valor de $p < 0{,}05$). De maneira mais específica, aproximadamente 60% das imagens contendo desinformação foram compartilhadas mais de 100 vezes em sites distintos (Figura 3(a)), enquanto 40% do conteúdo não verificado foi compartilhado em mais de 10 sites distintos. Essas descobertas se mantém (proporcionalmente) para a nossa análise de domínios únicos apresentada na Figura 3(b).

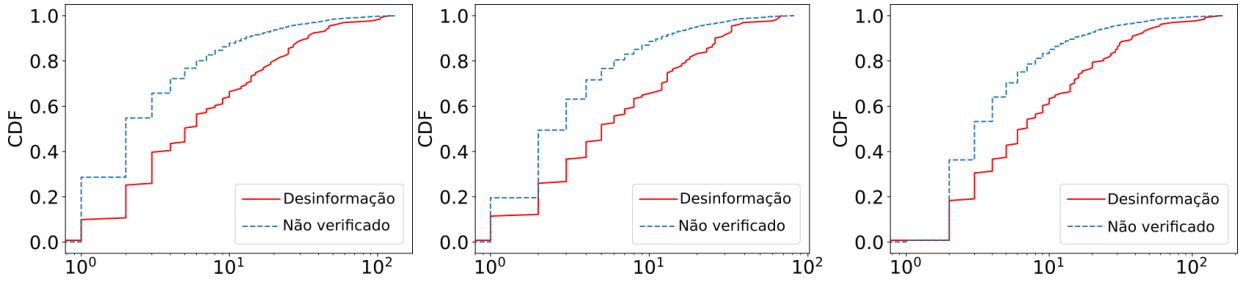
Por último, para capturar **padrões temporais** de cada notícia a partir da atividade de compartilhamento da informação no WhatsApp, calculamos a taxa em que os compartilhamentos são feitos em intervalos desde o primeiro compartilhamento (900, 1800, 2700, 3600, 7200, 14400, 28800, 57600, 86400, 172800, 259200, 345600 e 432000 segundos). No geral, descobrimos que as imagens contendo desinformação têm um alcance muito mais rápido no WhatsApp e na Web, sugerindo um comportamento viral dentro e fora do WhatsApp.

4.3 Importância dos Atributos

Nesta seção, avaliamos o poder relativo de cada de um dos atributos implementados neste trabalho em discriminar desinformação do conteúdo não verificado de acordo com o InfoGain (IG) [1]. A Tabela 3 apresenta uma relação dos top-20 atributos mais discriminativos de acordo com esta medida.

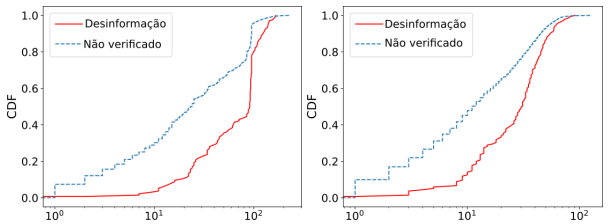
Podemos observar que os 20 atributos mais discriminativos são distribuídos entre os três grupos apresentados, ou seja, conteúdo, fonte e ambiente, ressaltando a necessidade de se explorar atributos oriundos de todos eles. Além disso, percebemos uma tendência: os atributos extraídos a partir do conteúdo (`cont`) são majoritários, seguidos dos atributos extraídos do ambiente (`amb`) e, depois, os relacionados à fonte (`font`). De forma geral, a expressividade do número de atributos extraídos do conteúdo ressalta a importância dos mesmos para a detecção de desinformação disseminada no WhatsApp por meio de imagens, principalmente se considerarmos os atributos semânticos relacionados às imagens. Por exemplo, existem várias notícias que são rotuladas como desinformação simplesmente por apresentarem informações que, em alguns casos são até verdadeiras, mas que foram disseminadas fora do contexto. Além disso, considerando a importância dos atributos de acordo com IG, podemos observar que os atributos extraídos do ambiente

[13]Para determinar os links comuns, usamos sufixos predefinidos: ".com", ".net", ".edu", ".org", ".mil", ".gov", ".br" extraídos de https://www.domain.com/blog/2018/10/30/domain-name-types/



(a) Número de usuários únicos por mensagem. (b) Número de grupos únicos por mensagem. (c) Número total de compartilhamentos.

Figura 2: Distribuições cumulativas (CDFs) do alcance de cada mensagem em termos de usuários distintos, grupos distintos e total de compartilhamentos.



(a) Número total de sites. (b) Número de domínios únicos.

Figura 3: Distribuições cumulativas (CDFs) do alcance de cada mensagem divulgada no WhatsApp e pela Web.

Tabela 3: Importância dos atributos para detecção de desinformação implementados neste trabalho.

	Top-20 atributos de acordo com o InfoGain (IG)	(%)
1	count_web_dissemination_urls (amb:prop ext)	9.30
2	web_dissem_accessible_links (amb:prop ext)	5.10
3	web_dissem_foreign_uncommon_domains (amb:prop ext)	4.30
4	acc_259200 (amb:eng int)	3.80
5	count_groups (amb:eng int)	2.90
6	sentence_info_syll_per_word (cont:sint)	2.80
7	Dic (cont:lexi)	2.10
8	Bridge (cont:seman)	2.00
9	ingest (cont:psico)	1.90
10	count_low_word (cont:lexi)	1.80
11	toxicity (cont:seman)	1.80
12	Quote (cont:lexi)	1.70
13	img_faces (img_has_faces) (amb:prop img)	1.70
14	img_count_labels (cont:prop img)	1.60
15	Sixltr (cont:lexi)	1.60
16	img_count_objects (cont:prop img)	1.60
17	political_bias_right (font:vies)	1.60
18	anger (cont:psico)	1.50
19	number (cont:lexi)	1.40
20	cogmech (cont:psico)	1.40

(amb) totalizam o maior ganho de informação (27,1%) e que os relacionados à propagação (interna e externa ao WhatsApp) aparecem entre os 5 principais, confirmando que eles podem ser, de fato, úteis para fins de detecção de desinformação disseminada neste ambiente, conforme resultados apresentados na próxima seção.

5 DETECÇÃO DE DESINFORMAÇÃO NO WHATSAPP

Nesta seção, apresentamos detalhes da nossa configuração experimental, incluindo as abordagens de aprendizado de máquina que foram testadas bem como as métricas utilizadas para avaliação da eficácia das mesmas. Por fim, apresentamos e discutimos os principais resultados experimentais.

5.1 Configuração Experimental

Neste trabalho, investigamos o poder discriminativo dos 181 atributos implementados para detecção de desinformação usando vários classificadores clássicos e modernos, incluindo *k-Nearest Neighbors* (KNN) [9], *Naive Bayes* (NB) [17], *Random Forests* (RF) [2], non-linear Support Vector Machine with the Radial Basis Function (SVM) [12] e XGBoost (XGB) [6].

Além disso, para avaliar a eficácia de nossas estratégias de classificação investigadas, nós adotamos métricas comumente utilizadas em tarefas de Aprendizado de Máquina e Recuperação de Informação [1, 38]: MacroF1, que nos permite avaliar de forma adequada a performance das abordagens em nosso cenário desbalanceado, e *area under the ROC curve* (AUC), que é uma métrica para classificação binária frequentemente usada como uma medida de qualidade do desempenho dos modelos.

Finalmente, todos os experimentos foram executados com a aplicação de uma validação cruzada de 5 partições (i.e., treino e teste). Além disso, os experimentos foram replicados 50 vezes (com versões embaralhadas do conjunto de dados original) para permitir o cálculo e reporte do intervalo de confiança (95%) para todas as métricas. É válido ressaltar que todas as partições construídas mantiveram a distribuição original do conjunto de dados. De forma geral, cada classificador "aprendeu" um modelo a partir de um conjunto de dados previamente rotulado (i.e., pré-classificados) e, em seguida, nós usamos este modelo para classificar novas instâncias (ainda não vistas) como " desinformação" ou "conteúdo não verificado". Os resultados obtidos durante esta etapa do trabalho são apresentados na próxima seção.

5.2 Resultados de Classificação

A Tabela 4 apresenta os resultados experimentais obtidos com intervalos de confiança de 95% para os modelos treinados a partir de

Tabela 4: Resultados experimentais.

Classificador	MacroF1	AUC
KNN	0.95±0.00	0.63±0.03
NB	0.95±0.00	0.64±0.03
RF	0.96±0.01	0.71±0.06
SVM	0.95±0.01	0.46±0.15
XGB	**0.97±0.01**	**0.82±0.03**

todos os atributos implementados e descritos anteriormente (ver Tabela 2) considerando o nosso conjunto de dados do WhatsApp. De forma geral, podemos observar que o XGB obteve os melhores resultados, com 0,97 (± 0,01) e 0,82 (± 0,03) para MacroF1 e AUC, respectivamente.

Depois disso, como uma tentativa inicial de investigação do potencial prático dessas abordagens automáticas para detecção de desinformação disseminada no WhatsApp, nós inspecionamos a curva ROC para o melhor classificador (i.e., XGB), conforme resultado apresentado na Figura 4. Podemos observar que é possível escolher um limite para classificar corretamente a maioria das mensagens que contém desinformação (Taxa de Verdadeiros Positivos ≈ 1), enquanto classificamos erroneamente cerca de 55% do conteúdo não verificado (Taxa de Falsos Positivos ≈ 0, 55).

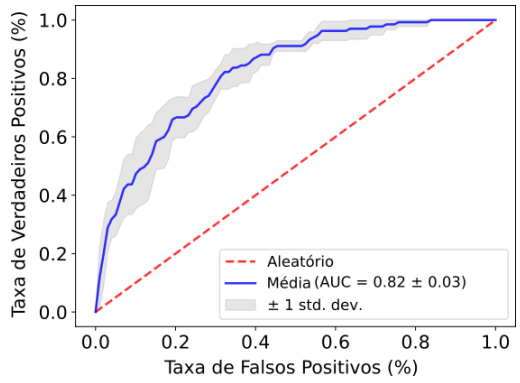


Figura 4: Curva ROC para o classificador que obteve melhores resultados - a saber XGB, usando dados do WhatsApp. É possível classificar corretamente quase todas as mensagens contendo desinformação com taxa de 55% de falsos positivos.

6 CONCLUSÃO

Neste trabalho nós investigamos o potencial de abordagens supervisionadas na tarefa de identificar desinformação disseminada por meio de imagens no WhatsApp. Para isso, nós pesquisamos trabalhos recentes e relacionados como uma tentativa de implementar todos os potenciais atributos para detecção de desinformação. Neste contexto, nós computamos 181 atributos que capturam informações relativas ao conteúdo, fonte e a dinâmica de propagação deste tipo de conteúdo dentro e fora desta plataforma digital. É válido ressaltar que neste contexto propomos atributos novos relacionados incluindo aqueles relacionados à propriedades da imagem, estrutura semântica do texto, novas medidas de propagação do conteúdo, etc, e aplicamos, ainda, alternativas que até então não tinham sido aplicadas no contexto explorado neste trabalho.

Em seguida, nós investigamos diferentes abordagens de aprendizado de máquina supervisionado, avaliando sua eficácia na tarefa de detectar automaticamente desinformação disseminada no WhatsApp. Nossos resultados revelaram que a detecção automática pode ser usada por verificadores de fatos como uma ferramenta auxiliar no processo de identificação de conteúdo com maior probabilidade de ser falso. Particularmente, mostramos que o desempenho de previsão dos atributos implementados neste trabalho combinados com classificadores existentes apresentam um grau útil de poder discriminativo para detectar desinformação. Nossos melhores resultados de classificação podem detectar corretamente quase todas as mensagens contendo desinformação em nossos dados, enquanto classificam incorretamente cerca de 55% do conteúdo não verificado, o que já é suficiente para apoiar a tarefa de checagem de fatos.

Por fim, esperamos que este trabalho motive esforços futuros focados na proposição de soluções automáticas que apoiem a checagem de fatos no combate à desinformação, desencadeando novas contramedidas que sejam efetivas nas próximas eleições. Como trabalhos futuros, planejamos conduzir análises de erros das abordagens bem como investigar o desempenho dos modelos a partir de subconjuntos de atributos. Além disso, também pretendemos explorar outros contextos (ex.: saúde), outras plataformas (ex.: Telegram) e conduzir análises individuais de informatividade dos atributos implementados e investigar o potencial de técnicas de aprendizado de máquina mais sofisticadas (ex.: aprendizado ativo e profundo) na tarefa de detectar desinformação.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo Ministério Público de Minas Gerais (MPMG), projeto Capacidades Analíticas, CNPq, CAPES e Fapemig.

REFERENCES

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval.* Vol. 463. ACM press New York.
[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[3] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.
[4] Victor S Bursztyn and Larry Birnbaum. 2019. Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups. In *Proc. of the Int'l IEEE/ACM Conference on Advances in Social Networks Analysis and Mining (ASONAM).* 484–488.
[5] Flavio Carvalho, Rafael Rodrigues, Gabriel Santos, Pedro Cruz, Lilian Ferrari, and Gustavo Guedes. 2019. Avaliação da versão em português do LIWC Lexicon 2015 com análise de sentimentos em redes sociais. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BrasNAM).* 24–34.
[6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proc. of the Int'l ACM Conference on Knowledge Discovery and Data Mining (KDD).* 785–794.
[7] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLOS ONE* 10, 6 (2015), e0128193.
[8] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T).* 1–4.
[9] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on Information Theory* 13, 1 (1967), 21–27.
[10] Lucas H. C. de Lima, Julio S. Reis, Philipe Melo, Fabricio Murai, and Fabricio Benevenuto. 2020. Characterizing (Un)moderated Textual Data in Social Systems. In *Proc. of the IEEE/ACM Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM).*

[11] Samuel S Guimarães, Julio CS Reis, Filipe N Ribeiro, and Fabrício Benevenuto. 2020. Characterizing Toxicity on Facebook Comments in Brazil. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*. 253–260.

[12] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of the European Conference on Machine Learning (ECML)*. 137–142.

[13] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[14] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. 2019. A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections.. In *Proc. of the Int'l ACM Conference on World Wide Web (WWW) Companion*. 1013–1019.

[15] Antônio Diogo Forte Martins, Lucas Cabral, Pedro Jorge Chaves Mourão, José Maria Monteiro, and Javam Machado. 2021. Detection of Misinformation About COVID-19 in Brazilian Portuguese WhatsApp Messages. In *Proc. of the Int'l Conference on Applications of Natural Language to Information Systems (NLDB)*. 199–206.

[16] F. Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[17] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. 752, 1 (1998), 41–48.

[18] Philipe Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro OS de Melo, and Fabrício Benevenuto. 2019. Can WhatsApp Counter Misinformation by Limiting Message Forwarding?. In *Proc. of the Int'l Conference on Complex Networks and their Applications (Complex Networks)*. 372–384.

[19] Rafael A Monteiro, Roney LS Santos, Thiago AS Pardo, Tiago A de Almeida, Evandro ES Ruiz, and Oto A Vale. 2018. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In *Proc. of the Int'l Conference on Computational Processing of the Portuguese Language (PROPOR)*. 324–334.

[20] Lucas J Myslinski. 2012. Fact checking method and system. Google Patents. US Patent 8,185,448.

[21] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. 2019. Reuters Institute Digital News Report 2019. Reuters Institute for the Study of Journalism.

[22] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *Proc. of the Int'l Conference on Computational Linguistics*, 3391–3401.

[23] Julio Reis, Fabricio Benevenuto, Pedro OS de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*. 357–366.

[24] Julio CS Reis, Philipe Melo, Kiran Garimella, Jussara M Almeida, Dean Eckles, and Fabrício Benevenuto. 2020. A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and Indian Elections. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*. 903–908.

[25] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019).

[26] Gustavo Resende, Philipe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In *Proc. of the Int'l ACM Conference on Web Science (WebScience)*. 225–234.

[27] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proc. of the ACM Web Conference (WWW)*. 818–828.

[28] Filipe Ribeiro, Lucas Henrique, Fabrício Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P. Gummadi. 2018. Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*. 290–299.

[29] Manoel Horta Ribeiro, Pedro H Calais, Virgílio AF Almeida, and Wagner Meira Jr. 2017. "Everything I Disagree With is# FakeNews": Correlating Political Polarization and Spread of Misinformation. In *Proc. of the Workshop on Data Science + Journalism @KDD*.

[30] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3 (2019), 1–42.

[31] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. In *Proc. of the Workshop on Data Science for Social Good (SoGood)*.

[32] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[33] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.

[34] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proc. of the ACL Workshop on Language Technologies and Computational Social Science*. 18–22.

[35] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 647–653.

[36] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[37] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 422–426.

[38] Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information retrieval* 1, 1-2 (1999), 69–90.