Capítulo 2

Uma Introdução ao Combate Automático às *Fake News* em Redes Sociais Virtuais

Paulo Márcio Souza Freire, Ronaldo Ribeiro Goldschmidt

Abstract

Combating Fake News (i.e., false news intentionally spread) is not a recent problem. However, its complexity has increased mainly due to the growth of volume and speed of news dissemination provided by the virtual social networks. In this scenario, computational approaches are becoming essential devices to combat this type of news. Thus, this Chapter presents a conceptual and practical introduction to the main computational approaches to combat Fake News, besides some comments on related areas and recent research on this theme.

Resumo

O problema de combater Fake News (ie., notícias falsas veículadas de forma intencional) não é recente. Contudo, sua complexidade vem aumentando em função do crescimento do volume e da velocidade de divulgação de notícias proporcionado pelas redes sociais virtuais. Diante deste cenário, abordagens computacionais que possam auxiliar no combate automático deste tipo de notícia estão se tornando cada vez mais necessárias. Assim sendo, o presente Capítulo apresenta uma introdução conceitual e prática às principais abordagens computacionais de combate às Fake News, além de comentar sobre áreas e pesquisas recentes relacionadas a este tema.

2.1. Considerações Iniciais

Historicamente, a publicação de notícias estava restrita à mídia tradicional, como rádio, TV, jornais e revistas impressos. Com o surgimento das redes sociais virtuais de fácil acesso e baixo custo (também conhecidas como, simplesmente, redes sociais), as pessoas vêm, a cada dia, aumentando o consumo de notícias *on-line*, em vez daquelas fornecidas pelos canais tradicionais [Vosoughi et al. 2017].

Apesar de seus benefícios, as redes sociais permitem que qualquer pessoa, independentemente de sua credibilidade, divulgue (publique/propague) notícias com intenso poder de espalhamento [Shu et al. 2017a][Wang et al. 2018a]. Portanto, as redes sociais amplificaram um problema antigo: a disseminação de notícias falsas [Conroy et al. 2015] [Zhang et al. 2018]. Este problema abrange uma questão ainda mais difícil: *Fake News* que é a divulgação de uma notícia falsa de forma intencional [Shu et al. 2017a]. Este espalhamento de notícias propositalmente falsas costuma ser prejudicial, pois uma inverdade deliberada tende a ser melhor elaborada e, portanto, mais eficaz em seu objetivo principal, que é influenciar na mudança de opinião. A proliferação de *Fake News*, geralmente, afeta não apenas a integridade jornalística, mas também perturba as áreas social, política, econômica, cultural e da segurança [Wang 2017][Mustafaraj and Metaxas 2017].

Como materialização do poder de influência deste tipo de notícia, pode-se destacar que, somente nos Estados Unidos da América (EUA), mais de sessenta e dois porcento dos adultos recorrem às redes sociais para receberem notícias. Como consequência deste elevado percentual, alguns casos relevantes, ocorridos em 2016, podem ser destacados [Farajtabar et al. 2017]:

• Nos três meses finais das eleições presidenciais, as notícias falsas publicadas no *Facebook*, que favoreceram qualquer um dos dois candidatos, foram compartilhadas 37 milhões de vezes;

• Uma análise do *Buzzfeed News* [1] mostra que, a partir das 20 principais notícias falsas sobre as eleições, criadas por sites fraudulentos, foram geradas quase 1,5 milhões de atividades de engajamento de usuários no *Facebook*;

• Um homem, carregando um rifle AR-15, aterrorizou os frequentadores de uma pizzaria na capital *Washington*, porque ele havia lido uma notícia falsa *on-line*, afirmando que o referido estabelecimento usava crianças jovens como escravas sexuais.

Inclusive, casos relacionados às *Fake News* não se limitam aos EUA. Em 2018, na Índia, após notícias falsas terem, supostamente, levado a linchamentos, o *WhatsApp* anunciou um limitador para a quantidade de encaminhamentos de mensagem [2]. Portanto, há um apelo urgente para desenvolver estratégias efetivas para mitigar o impacto deste tipo de notícia falsa.

Nos últimos anos, tanto a academia quanto a indústria estudam como combater *Fake News* nas redes sociais [Flintham et al. 2018] [Wang et al. 2018a] [Zhou et al. 2019] [Campan et al. 2017] [Kshetri and Voas 2017]. Este combate apresenta-se como não trivial, tanto pelo volume de publicações quanto pela velocidade das suas respectivas propagações. Assim, o emprego de apordagens computacionais, devido à sua maior velocidade de atuação, vem se destacando no combate às *Fake News* nas redes sociais [Ruchansky et al. 2017].

Baseado nesta necessidade computacional, o presente Capítulo provê uma introdução ao referido combate através da seguinte estrutura: A Seção 2.2 apresenta diferentes definições para o termo *Fake News*, assim como aborda o comportamento disseminativo deste tipo de notícia nas redes sociais. Um levantamento sobre os trabalhos relacionados

[1] https://www.buzzfeed.com/

[2] BBC News Brasil - https://www.bbc.com

é realizado na Seção 2.3. A Seção 2.4, por sua vez, realiza um estudo de caso onde aplica detecção automática baseada na reputação do usuário via *Crowd Signals*. Por fim, na Seção 2.5, são abordados os problemas em aberto.

2.2. Fundamentos

Como a utilização do termo *Fake News* é relativamente recente, a sua caracterização se faz necessária. Para tal, são agrupadas as diferentes definições para *Fake News*, assim como são categorizadas as razões que levam ao seu comportamento disseminativo nas redes sociais.

2.2.1. Definição de *Fake News*

Apesar da originalidade da expressão, as *Fake News* não surgiram com o uso das redes sociais. Haja vista que, mesmo com as mídias tradicionais, já existiam pessoas que, por diferentes razões, divulgavam notícias falsas de forma proposital [Golbeck et al. 2018]. Independente do surgimento, devido à contemporaneidade do termo, *Fake News* apresenta diversas definições que podem ser organizadas em dois grupos.

O primeiro grupo considera que o aspecto proposital é fundamental, pois define as *Fake News* como publicações intencionalmente e verificadamente falsas [Shu et al. 2017a] [Mustafaraj and Metaxas 2017] [Reis et al. 2019] [Zhou et al. 2019] [Campan et al. 2017] [Flintham et al. 2018] [Wang et al. 2018a] [Zhou and Zafarani 2018] [Conroy et al. 2015]. Para enfatizar a diferença entre uma notícia falsa e uma intencionalmente falsa, pode-se utilizar dois termos denominados *misinformation* e *disinformation* [Golbeck et al. 2018] [Campan et al. 2017]. Enquanto *misinformation* corresponde às notícias falsas publicadas pela falta da informação verdadeira, a *disinformation* diz respeito às notícias falsas divulgadas com algum propósito. Com base nestas correspondências, é possível caracterizar *Fake News* como sendo uma *disinformation* [Kshetri and Voas 2017]. Cabe ressaltar que, apesar de pertencente ao primeiro grupo, o trabalho [Zhou and Zafarani 2018] é ainda mais específico em sua definição, pois só considera *Fake News* quando a notícia intencionalmente falsa é divulgada por uma agência de notícias. Ademais, ainda de acordo com este primeiro grupo, existem outras áreas que, apesar de não abordarem a questão do combate às *Fake News*, apresentam relação com as notícias intencionalmente falsas. Algumas destas áreas se encontram descritas abaixo:

• Classificação de Rumores (*Rumor Classification*) - Rumor é uma informação em circulação cuja veracidade não foi verificada no momento da publicação. Um rumor pode ser classificado como verdadeiro, falso ou ainda não verificado [Shu et al. 2017a] [Liu and Xu 2016] [Vosoughi et al. 2017] [Ma et al. 2015]. Portanto, uma *Notícia* não verificada antes da publicação é um *Rumor*, que pode ser caracterizado como *Fake News* a partir do momento que seja identificado como falso e intencional. A tarefa mais relacionada com o combate às *Fake News* é a classificação da veracidade dos rumores;

• Descoberta da Verdade (*Truth Discovery*) - é a descoberta da verdade de fatos conflitantes entre diferentes fontes [Shu et al. 2017a] [Li et al. 2015]. Assim, uma mesma *Notícia* pode conter afirmações diferentes (distintas opiniões), onde as intencionalmente falsas podem ser caracterizadas como *Fake News*. Assim, o combate às *Fake News* pode se beneficiar da Descoberta da Verdade para determinar a veracidade das afirmações;

• Detecção de Iscas de Cliques (*Clickbait Detection*) – procura identificar, nas páginas *Web*, as chamadas iscas de cliques que, praticamente, forçam o usuário a selecionar a opção apresentada. Neste caso, o corpo do texto (*bodytext*) dos artigos é, frequentemente, pobre em relação ao seu cabeçalho (*headlines*). Esta discrepância pode ser encontrada não só em *Clickbait*, como também em *Fake News*. Sendo assim, o *Clickbait* pode ser usado como um indicador de *Fake News* [Shu et al. 2017a];

• Detecção de Bots (*Bot Detection*) – procura identificar o envio automático de informações nas redes sociais por meio de robôs [Braz and Goldschmidt 2017]. Estes envios podem potencializar tanto a publicação quanto a respectiva propagação da *Fake News* [Wang et al. 2018a] [Nasim et al. 2018] [Ferrara et al. 2016];

• Checagem de fatos (*Fact Checking*) - são *Websites* ou *Frameworks* responsáveis pela verificação, normalmente realizada com a ajuda de especialistas, da veracidade de fatos divulgados em redes sociais [Ciampaglia et al. 2015] [Ruchansky et al. 2017] [Sethi 2017] [Vo and Lee 2018]. Inclusive, existem abordagens voltadas para a seleção automática de notícias a serem enviadas para a referida checagem [Kim et al. 2018] [Tschiatschek et al. 2018]. A verificação da verdade dos fatos pode ser utilizada na tarefa de detecção de *FakeNews* [Cazalens et al. 2018], assim como na criação de *datasets*;

• Sistemas de Reputação (*Reputation System*) - são sistemas que buscam determinar o nível de confiança em redes sociais baseados na obtenção de graus de reputação [Vavilis et al. 2014] [Hendrikx et al. 2015] [Seo J. 2013] [Deng et al. 2014] [Sherchan et al. 2013]. A determinação de graus de reputação dos usuários pode ser utilizada na tarefa de identificação das *Fake News*.

O segundo grupo, entretanto, tem uma definição mais genérica. Para este segmento, as *Fake News* são todas as notícias falsas, independente da sua natureza intencional [Sharma et al. 2019] [Castelo et al. 2019] [Ajao et al. 2019]. Inclusive, consideram-se como *Fake News* outros tipos de notícia, como, por exemplo, Rumor.

Este Capítulo adota a definição do primeiro grupo, consequentemente considera *Fake News* como sendo, somente, uma notícia intencionalmente falsa. A principal razão para esta escolha é que uma notícia propositalmente divulgada tende a ser mais bem elaborada, podendo, assim, causar mais malefícios aos usuários das redes sociais.

2.2.2. Comportamento disseminativo das *Fake News*

A disseminação e, consequente, divulgação de uma notícia se inicia pela sua publicação e provável propagação na rede social (Efeito de Câmara de Eco) [Shu et al. 2017a]. Desta forma, é importante destacar o momento no qual uma notícia pode ser caracterizada como *Fake News*. Basicamente, uma notícia intencionalmente falsa pode surgir de duas formas. A primeira é quando a *Fake News* é iniciada na rede social por meio da sua publicação e, posteriormente, potencializada pela sua possível propagação. A segunda é quando uma notícia não *fake* é publicada, porém se tornar *fake*, a partir do seu espalhamento, de acordo com as contribuições intencionalmente falsas feitas durante a sua propagação.

Independente do momento de criação, a recente proliferação de notícias falsas e mal-intencionadas nas redes sociais tem sido uma fonte de preocupação generalizada. Esta apreensão se deve pelo seu poder de espalhamento e, consequente, influência na

sociedade [Flintham et al. 2018]. As razões que potencializam a divulgação das *Fake News* nas redes sociais podem ser divididas em quatro categorias. A primeira tem relação com poder de influência ocasionado pelos fatores inerentes ao ser humano, dentre eles podemos destacar que as pessoas [Shu et al. 2017a]:

• Preferem receber informações que confirmem as suas opiniões sem, necessariamente, verificarem a veracidade da notícia;

• Tendem a aceitar as informações não pela análise da verdade, mas pela relação de ganhos e perdas que a notícia pode trazer para elas;

• Tendem a avaliar as informações não pela busca da veracidade, pois acabam acompanhando a aceitação dos outros.

A segunda categoria é a carência de legislação punitiva, sendo uma das alegações para tal fato é que as referidas leis poderiam cercear a liberdade de expressão. A terceira categoria está vinculada ao potencial ganho financeiro com a divulgação de determinadas notícias [Kshetri and Voas 2017] nas redes sociais. Já a quarta categoria advém da facilidade de criação de contas nas redes sociais [Conroy et al. 2015]. Um aspecto importante inerente à esta facilidade é a criação de contas digitais maliciosas por meio de divulgadores de natureza humana e/ou computacional [Shu et al. 2017a]. Estes divulgadores subdividem-se em:

• *Bot* - robôs responsáveis por divulgar *Fake News*;

• Humano - pessoas (*trolls*) intencionadas em disseminar *Fake News*;

• *Cyborg* - mecanismos híbridos (Humano/*Bot*) que divulgam *Fake News*.

Ainda se tratando da facilidade de divulgação de notícias intencionalmente falsas nas redes sociais, uma das formas mais simples de criar uma *Fake News* é se infiltrar em uma comunidade de pessoas engajadas em discutir um determinado assunto. Portanto, segundo [Mustafaraj and Metaxas 2017], devem ser realizados os seguintes passos: Criar um domínio falso (*website*), criar contas anônimas, identificar comunidades e usuários interessados em um determinado assunto, contaminar estes usuários com a notícia falsa e, finalmente, incentivar a discussão para que a *Fake News* seja espalhada.

2.3. Trabalhos Relacionados

Para apresentar os trabalhos vinculados ao combate automático às *Fake News* nas redes sociais é proposto e, em seguida, aplicado um modelo comparativo que viabilize uma distinção entre abordagens computacionais. Desta forma, as abordagens, juntamente com o seu respectivo *dataset*, são enquadrados no citado modelo.

2.3.1. Proposta de Modelo Comparativo

O combate às *Fake News* em redes sociais, por meio de abordagens computacionais, possui uma variedade de aspectos que podem ser considerados. Com o objetivo de facilitar a comparação e a consequente classificação das referidas abordagens, tais aspectos são categorizados na Figura 2.1. As próximas subseções detalham cada um destes aspectos.

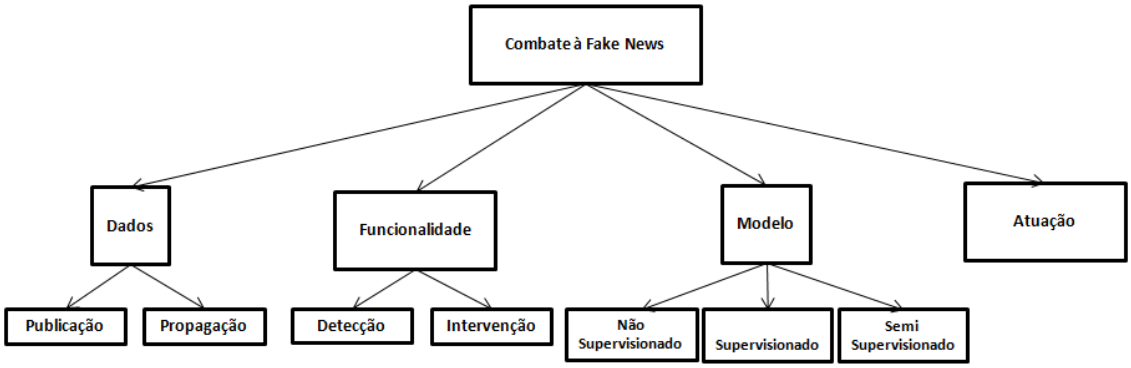

Figura 2.1: Aspectos considerados em Abordagens de Combate Automático às *Fake News*

2.3.1.1. Dados

Aspecto relacionado aos dados que podem ser utilizados pelas abordagens computacionais de combate às *Fake News*. Este aspecto subdivide-se em dados obtidos a partir da *Publicação* da notícia, como também aqueles associados com a sua *Propagação*.

Os dados de *Publicação* representam as informações inerentes ao surgimento da notícia na rede social. Estes dados podem ser classificados em *Notícia, Usuário, Assunto e Temporalidade*. No que diz respeito à Notícia, a abordagem pode ser capaz de analisar dados oriundos da publicação a partir de diferentes tipos de *Mídia* (*Texto, Áudio e Imagem*). Independente da *Mídia*, a análise do *Conteúdo* pode ser realizada de forma *Léxica, Sintática, Semântica e Legibilidade*. Com relação ao *Usuário* publicador, a abordagem pode identificar diferentes *Tipos*, tais como: humano, *bot* ou *cyborg*. Pode-se analisar também dados referentes ao *Perfil* do usuário na rede social, tais como: identificação e idade. Outro aspecto relevante está relacionado à *Reputação* do publicador, que pode estar vinculada à sua capacidade em identificar ou publicar *Fake News*. A abordagem pode também utilizar o *Assunto* abordado no momento da publicação. Assim, é possível tratar Especificidades, tais como: relacionamento entre assuntos, assuntos controversos ou análise de tópicos. Outro aspecto leva em consideração a *Relevância* do assunto publicado, haja vista que assuntos em voga motivam a criação de *Fake News*. A variação das características de uma notícia com o passar do tempo, torna a *Temporalidade* mais um relevante recurso para a identificação de *Fake News*.

Os dados de *Propagação* representam as informações obtidas após a publicação, consequentemente, aquelas inerentes às contribuições devido ao espalhamento da notícia na rede social (ex: curtida/*like*, comentário/*reply* ou compartilhar/*retweet*). Portanto, estes dados podem ser classificados em *Contribuição, Usuário, Assunto, Temporalidade e Rede*. No que diz respeito à *Contribuição*, *Usuário*, *Assunto* e *Temporalidade* a abordagem pode ser capaz de analisar os dados oriundos da *Propagação*, a partir dos mesmos aspectos anteriormente citados na *Publicação*. Ademais, as informações relacionadas à *Rede* criada, a partir da propagação da notícia, possibilitam não só a identificação de uma *Fake News* como uma possível atuação contra a mesma.

2.3.1.2. Funcionalidade

Além dos dados coletados, as abordagens automáticas de combate às *Fake News* podem, basicamente, possuir duas funcionalidades: *Detecção e Intervenção*.

A *Detecção* automática da *Fake News* pode ser, basicamente, um problema de classificação binária onde dada uma rede social $\mathscr{G}$, uma notícia a e um conjunto de postagens (publicações/propagações) $\mathscr{P}$, relacionadas à a, são espalhadas através da $\mathscr{G}$ por um conjunto de usuários U em um intervalo de tempo t. Assim o referido classificador binário $\mathscr{F}$ deve, aprendendo a partir dos dados, prever se a é uma *fake news* ou não, como formalmente indicado na equação 1. Uma outra forma é a utilização de técnicas mais subjetivas que definam a probabilidade, peso ou pertinência de uma notícia a ser *fake*.

$$\mathscr{F}(\mathscr{G},a,\mathscr{P},U,t) = \begin{cases} 1, & \text{se } a \text{ é uma } \textit{fake news}; \\ 0, & \text{caso contrário.} \end{cases} \tag{1}$$

Independente da forma, para que uma notícia a possa ser detectada como *Fake News* é necessária a realização de duas subfuncionalidades: *Autenticidade e Intencionalidade* [Janze and Risius 2017] [Vosoughi et al. 2017]. A *Autenticidade* analisa se a notícia é verdadeira ou falsa, enquanto que a *Intencionalidade* busca determinar a intenção dos divulgadores em ludibriar os receptores. Esta *Intencionalidade* pode ser mensurada como pontuação, peso ou score e obtida, por exemplo, por intermédio da análise de sentimentos que a notícia disponibiliza, pela associação entre usuários, assim como pelas características de perfil, tipo e reputação (credibilidade/confiança) dos divulgadores.

Já a *Intervenção* automática procura atacar as *Fake News*, nas redes sociais, de forma proativa ou reativa [Shu et al. 2017a][Farajtabar et al. 2017]. A intervenção reativa busca combater os efeitos da notícia a partir do momento da sua detecção como notícia propositalmente falsa. Por outro lado, a intervenção proativa tenta atuar antes mesmo da referida detecção, agindo então como uma forma de prevenção. Além disso, a tarefa de intervenção pode ser dividida em dois segmentos: *o Bloqueio e a Mitigação*. O *Bloqueio* atua de forma reativa. Na sua forma mais branda, o bloqueio interrompe a propagação da notícia e/ou a atuação do(s) usuário(s) responsáveis. Uma outra forma mais incisiva seria remover a(s) notícia(s) e/ou o(s) usuário(s) divulgador(es). Já a *Mitigação* pode agir de forma reativa ou proativa buscando enfraquecer as consequências causadas pela *Fake News*. Na reatividade, a mitigação pode, por exemplo, imunizar os usuários provendo notícias verdadeiras [Farajtabar et al. 2017]. Uma forma de proatividade na mitigação é prover alertas, mesmo que a notícia ainda não tenha sido detectada com propositalmente falsa. Estes alertas podem estar relacionados com o nível de reputação da fonte (usuário) ou sobre o assunto estar relacionado com outras *Fake News* já identificadas.

Independente da funcionalidade da abordagem, a coleta dos dados inerentes à divulgação da notícia se faz necessária para subsidiar a detecção e a intervenção das notícias intencionalmente falsas. Assim, tanto a coleta de dados na rede social quanto as tarefas de detecção e intervenção são fases iterativas, conforme ilustra a Figura 2.2. Cabe ressaltar que quanto mais cedo acontecer a detecção e a intervenção da *Fake News*, os impactos negativos desta notícia tendem a ser menores.

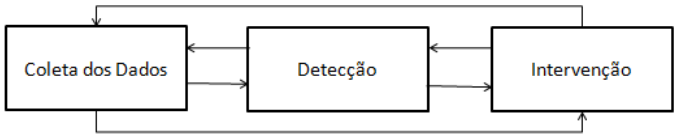

Figura 2.2: Fluxo do processo de combate às Fake News

2.3.1.3. Modelo

Quando a solução é por aprendizado de máquina, pode-se utilizar modelos computacionais para, a partir dos dados coletados, detectar as *Fake News*. Estes modelos são categorizados em *Não Supervisionado, Semi-Supervisionado e Supervisionado*.

No modelo *Não Supervisionado* são categorizadas as técnicas que normalmente levam mais tempo para realizar a identificação, porém, como não necessitam de rótulos, podem utilizar *datasets* mais simples [Shu et al. 2017a].

Os modelos *Supervisionados* são lentos na fase do treinamento, entretanto, tendem a ser mais rápidos do que os não supervisionados no momento da sua utilização na identificação das *Fake News*. Devido à necessidade de treinamento, os modelos supervisionados precisam de *datasets* mais completos [Shu et al. 2017a].

O modelo *Semi-Supervisionado* procura realizar a tarefa de identificação da *Fake News*, em redes sociais, de uma forma híbrida que busque utilizar, tanto as técnicas supervisionadas quanto não supervisionadas. Esta abordagem pode utilizar *datasets* mais simples do que aqueles manipulados pelos modelos supervisionados, porém mais complexos do que os utilizados pelos não supervisionados [Shu et al. 2017a].

2.3.1.4. Atuação

As abordagens computacionais que visam o combate às *Fake News*, independemente dos dados coletados, funcionalidade e modelo utilizados, podem ter diferentes formas de atuação.

Uma das possibilidades de atuação está associada à localização física do combate dentro da rede social. Uma abordagem *Centralizada* encontra-se, fisicamente, em um único ponto da rede social. Portanto, todas as tarefas relacionadas com a detecção/intervenção da *Fake News* são executadas em um mesmo local.

Por outro lado, uma abordagem *Descentralizada* encontra-se, fisicamente, espalhada na rede social. Assim, esta forma de atuação possibilita, inclusive, uma execução paralela e/ou ditribuida [Wu and Liu 2018] no combate às *Fake News*.

2.3.2. Revisão dos Trabalhos Relacionados

Nesta Subseção são apresentados alguns trabalhos relacionados ao combate automático às *Fake News* em redes sociais. Para tal, foram realizadas buscas onde as principais fontes de

consulta foram os artigos [Zhou and Zafarani 2019] [Reis et al. 2019] [Shu et al. 2017a] [Conroy et al. 2015] [Zhou et al. 2019] [Zhou and Zafarani 2018] [Sharma et al. 2019].

Para um melhor entendimento, os citados trabalhos são identificados e enquadrados no Modelo Comparativo tratado na Subseção 2.3.1, conforme mostram as Tabelas 2.1, 2.2 e 2.3. Cabe ressaltar que, nestas três Tabelas, as células não preenchidas indicam a não utilização do respectivo aspecto no trabalho correspondente.

Além disso, os referidos trabalhos são brevemente descritos, podendo seus detalhes serem consultados através das respectivas referências:

T1) *A Topic-Agnostic Approach for Identifying Fake News Pages* [Castelo et al. 2019]: O trabalho propõe um *topic-agnostic* (TAG) classificador que usa dados linquísticos e *Web-Markup* (padrões de layout das páginas) para detectar Fake News. Assim, ao invés de usar o *bag of words*, o trabalho explora as topic-agnostic, incluindo características morfológicas, psicológicas e de legibilidade que são comuns em *Fake News*. O trabalho propõe que páginas com *Fake News* normalmente têm inclinação sensacionalista, assim como a ocorrência de termos, tais como: *"Just in" e "Read this"*. Foram utilizados 3 classificadores *Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Random Forest (RF)*. Comparou o TAG com os resultados obtidos em [Pérez-Rosas et al. 2018] (T2), separando-os ano a ano (2013 até 2018);

T2) *Automatic Detection of Fake News* [Pérez-Rosas et al. 2018]: Este trabalho cria uma ferramenta de detecção de *Fake News* por classificação com *Support Vector Machines* (SVM), combinando informações léxicas, sintáticas, semânticas e de legibilidade. O presente trabalho compara os resultados com a detecção humana;

T3) *Automatic Detection of Fake News on Social Media Platforms* [Janze and Risius 2017]: Este artigo implementa a detecção com os classificadores binários *Logistic Regression, Support Vector Machines (SVM), Decision Tree, Random Forest e Extreme Gradient Boosting*. O referido trabalho compara os resultados entre os classificadores, onde os melhores resultados foram alcançados com SVM;

T4) *Automatically Identifying Fake News in Popular Twitter Threads* [Buntain and Golbeck 2017]: O trabalho apresenta um método para detecção de *Fake News* no *Twiter* que acumula, ao longo do tempo, as características de rede, usuário e conteúdo para gerar uma regressão linear. Assim, a abordagem realiza a sua análise, levando em consideração os aspectos temporais relacionados à notícia. O artigo avalia os resultados nos *datasets PHEME (Twitter* para rumor), *CredBank (Twitter)* e *BuzzFeed News Fact-Checking Dataset (Facebook)* que precisaram ser alinhados com as mesmas características e rótulos. Os resultados apontam que o *dataset CredBank* foi o mais indicado para a detecção automática de *Fake News* praticada;

T5) *Beyond News Contents: The Role of Social Context for Fake News Detection* [Shu et al. 2019b]: Este artigo explora as correlações da postura da notícia, o bias e engajamento do usuário. Assim, é apresentado um Tri-Relacionamento (TriFN) onde tanto informações partidárias quanto níveis de confiança do usuário podem ser utilizados para detecção de *Fake News*. Além disso, os usuários tendem a formar relacionamentos com pessoas afins que podem aumentar o espalhamento das *Fake News*. Esta abordagem compara os seus resultados com outros trabalhos, como [Rubin et al. 2015] (T23);

T6) *CIMTDetect: A Community Infused Matrix-Tensor Coupled Factorization Based Method for Fake News Detection* [Gupta et al. 2018]: Através da modelagem de Câmara de Ecos, o trabalho representa uma notícia como um *3-mode tensor* <News, User, Community> e propõe um método baseado em *tensor factorization*. Além disso, apresenta uma extensão deste método com a junção de modelos que utilizam o conteúdo da notícia através de um *framework coupled matrix-tensor factorization*. Este artigo usou o algoritmo de detecção da comunidade *Girvan-Newman* para identificar, na rede social, comunidades representativas de câmaras de eco. Os seus resultados são comparados com métodos que utilizam o classificador SVM, porém com diferentes formas de análise de conteúdo (ex. N-Gram). Os dois métodos propostos *CITDetect (community-infused tensor information) e CIMTDetect (community-infused tensor information + conteúdo da notícia)* utilizam o classificador SVM;

T7) *Combining Neural, Statistical and External Features for Fake News Stance Identification* [Bhatt et al. 2018]: Neste estudo a ferramenta, desenvolvida para o primeiro desafio (FNC-1) [3], não tem o objetivo final de detectar se a notícia é *Fake News*. Nesta abordagem, as notícias são classificadas de acordo com a relação existente entre a manchete e o corpo do texto. Portanto os possíveis rótulos são *Agree* - o texto do corpo concorda com a manchete, *Disagree* - o texto do corpo discorda da manchete, *Discuss* - o texto do corpo discute a mesma afirmação que o título, mas não toma uma posição ou *Unrelated* - o texto do corpo discute uma alegação que difere do título. A ferramenta combina as abordagens neural e estatística com recursos externos. Para isto, a solução implementa um modelo profundo recorrente (*Neural Embedding*), um modelo ponderado de características estatísticas (*n-gram bag-of-words*) e recursos externos criados à mão com a ajuda de uma heurística de engenharia de recursos. Por fim, usando uma camada de rede neural profunda, todas as referidas abordagens são combinadas. Os resultados foram comparados com as demais ferramentas participantes do referido desafio;

T8) *CSI: A Hybrid Deep Model for Fake News Detection* [Ruchansky et al. 2017]: O trabalho procura melhorar a acurácia na detecção de *Fake News* por meio de um modelo híbrido de rede neural profunda chamado CSI. Este modelo utiliza três características: o texto da notícia, a resposta do usuário que recebeu a notícia e o usuário fonte da notícia. O CSI trabalha com o comportamento temporal dos usuários e da notícia. Este modelo se divide em três partes: *Capture, Score e Integrate*. O primeiro módulo é baseado no texto e na resposta, por meio de uma rede neural recorrente (LSTM) para capturar um padrão temporal de atividades do usuário sobre a notícia e a representação *Doc2Vec*. O segundo usa uma rede neural para aprender as características da fonte, baseado nas interações dos usuários, gerando um score por meio de um grafo. Os dois módulos são integrados com o terceiro para caracterizar ou não a notícia como *Fake News*. O trabalho propõe a sua utilização em diferentes domínios, inclusive, em bancos de dados. Os resultados foram comparados com técnicas criadas para detecção de rumores;

T9) *DistrustRank: Spotting False News Domains* [Woloszyn and Nejdl 2018]: Esta solução propõe uma estratégia de aprendizagem semi-supervisionada para separar automaticamente notícias falsas a partir de fontes não confiáveis de notícias. O trabalho utiliza como fonte *experts* de portais de checagem de fatos para classificar manualmente as no-

[3] http://www.fakenewschallenge.org/

tícias. A partir disto, é criado um grafo de pesos com os *ranks* de confiança sobre os *sites* e as arestas representam a similaridade dos mesmos. A pesquisa computa a centralidade, utilizando o *PageRank* em busca de uma similaridade entre os *sites* não confiáveis. O resultado da análise é a classificação em *Trust* ou *Distrust* para a fonte da notícia. O trabalho verificou que a semelhança entre os sites de notícias falsas é estatisticamente superior aos sites de notícias verdadeiras. Esta abordagem cita e compara os seus resultados com outros trabalhos a partir do mesmo *dataset*;

T10) *EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection* [Wang et al. 2018b]: O artigo aponta que a maioria das abordagens existentes aprendem a detectar *Fake News* a partir de características específicas do evento, consequentemente, não podem ser transferidas para outros eventos ainda não aplicados. Assim, este trabalho desenvolveu um *framework*, de ponta a ponta, denominado *EANN*, que pode derivar características invariantes de um evento para outro. Desta forma, propõe uma detecção de *Fake News* para eventos recém-chegados. Isso consiste de três componentes principais: o extrator de características multimodais para texto e imagem (rede neural Convolucional), o detector de *Fake News (fully connected layer com softmax)* e o discriminador de eventos (rede neural) que é o responsável por remover as características específicas do evento e manter as características compartilháveis entre os eventos para poder rotulá-los. Assim, o *framework* mede as características não similares entre diferentes eventos e remove-os para capturar as características invariantes entre eventos. Para avaliar seus resultados, realizou testes com técnicas de identificação de texto e imagem, porém utilizadas em trabalhos não ligados à detecção de *Fake News*;

T11) *Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks* [Liu and BrookWu 2018]: O artigo propõe um modelo para detecção precoce de Fake News através da classificação dos caminhos de propagação da notícia. O referido trabalho modela o caminho de propagação de cada notícia como uma série temporal multivariada, na qual cada tupla é um vetor numérico que representa as características do usuário empenhado em espalhar a notícia. Para tal, é construído um classificador de série temporal que incorpora redes recorrente e convolucional. Estas redes capturam as variações globais e locais das características do usuário, ao longo do caminho de propagação, para detectar *Fake News*. Esta abordagem cita e compara os seus resultados com outros trabalhos a partir do mesmo *dataset*;

T12) *Evaluating Machine Learning Algorithms for Fake News Detection* [Gilda 2017]: Este artigo explora técnicas de linguagem natural para a detecção de *Fake News*. O trabalho aplicou *term frequency-inverse document frequency (TF-IDF) de bigrams e probabilistic context free grammar (PCFG)* para um conjunto de 11.000 artigos em um *dataset* obtido pela *Signal Media* [4] e uma lista de fontes da *OpenSources.com* [5]. Este *dataset* foi testado com os algoritmos de classificação *Support Vector Machines, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees e Random Forests*. Os modelos com melhor desempenho foram os *Stochastic Gradient Descent*, treinados apenas no conjunto de recursos do TF-IDF;

[4] https://research.signal-ai.com/newsir16/signal-dataset.html

[5] http://www.opensources.co

T13) *FActCheck: Keeping Activation of Fake News at Check* [Srivastava et al. 2018]: Esta abordagem de Intervenção sobre *Fake News* propõe uma melhoria na abordagem *competing cascades*, onde os *AFC (algoritmos polynomial time greedy) e RAFC (fast graph-pruning)* procuram escolher quais usuários têm maior poder de mitigação. Assim, os usuários com maior capacidade de influência na rede social realizam a mitigação através da divulgação de notícias alternativas *(Real News)*;

T14) *Fake News Detection in Social Networks via Crowd Signals* [Tschiatschek et al. 2018]: A ferramenta desenvolvida trabalha na detecção e consequente intervenção de *Fake News*. Esta solução possui um algoritmo, chamado de *Detective* que usa inferência Bayesiana para detectar *Fake News* a partir de *Crowd Signals*. Este *Crowd* é formado pela opinião dos usuários sobre a notícia, juntamente com a sua capacidade em opinar corretamente. O objetivo é detectar, de forma antecipada, a *Fake News* e bloqueá-la. Os resultados foram comparados a partir de variações na própria abordagem, sendo as mesmas denominadas pelo artigo como Opt, Oracle, Fixed-CM e No-Learn;

T15) *Fake News Mitigation Via Point Process Based Intervention* [Farajtabar et al. 2017]: Neste artigo, o enfoque está na intervenção de *Fake News*. A proposta é intervir, mitigando a notícia falsa, fornecendo recompensas na forma de notícias verdadeiras para quem recebeu a *Fake News*. A nível de influência da *Fake News* e a respectiva mitigação são quantificadas por contadores. O modelo utilizado foi baseado em *least-squares temporal difference learning* (LSTD). Um dos experimentos foi real, com a criação de cinco contas no *Twitter*;

T16) *FakeNewsTracker: A Tool for Fake News Collection, Detection, and Visualization* [Shu et al. 2019a]: Apresenta o FakeNewsTracker, um sistema para detecção de notícias falsas. O FakeNewsTracker pode coletar, automaticamente, dados para notícias e contexto social. Este tratalho propõe um *framework end to end* para realizar a coleta de dados, a detecção das *Fake News* e a visualização dos resultados. Esta pesquisa usa *autoencoders* para aprender o conteúdo de notícias e *RNN* para capturar o padrão temporal dos usuários de acordo com o seu engajamento com a notícia. O trabalho compara os seus resultados, internamente, a partir de variações do próprio *FakeNewsTracker*, onde são considerados somente o conteúdo da notícia ou o contexto social. Além disso, os resultados são comparados também com *Support Vector Machine, Logistic Regression and Naive Bayes*;

T17) *Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection* [Wang 2017]: Além de propor um *dataset*, cria uma técnica de detecção de *Fake News* híbrida, usando redes neurais convolucionais (CNNs) para analisar, não somente textos, mas também os dados do usuário. O artigo obteve os melhores resultados ao ser comparado com os de outros três detectores implementados com *Logistic Regression Classifier* (LR), *Support Vector Machine Classifier* (SVM) e *bi-directional long short-term memory* (Bi-LSTMs);

T18) *Neural User Response Generator: Fake News Detection with Collective User Intelligence* [Qian et al. 2018]: O trabalho enfatiza a rápida propagação das *Fake News* nas redes sociais e, portanto, destaca a importância da sua detecção nos estágios iniciais, onde considera que apenas o texto da notícia está disponível. Tal afirmação se baseia no fato de que informações adicionais, como respostas dos usuários e padrões de

propagação, podem ser obtidas somente após a notícia se espalhar. Contudo, como as respostas propagadas podem ajudar na tarefa de detecção, os autores propõe um *Two-Level Convolutional Neural Network with User Response Generator (TCNN-URG)* onde o TCNN captura a semântica do texto da notícia e o URG cria um modelo generativo de resposta dos usuários propagadores. O URG, a partir de respostas históricas, é treinado para aprender como os usuários respondem às notícias publicadas, gerando respostas de usuários para ajudar a TCNN na detecção da *Fake News*. Esta abordagem cita e compara os seus resultados com outros trabalhos a partir do mesmo *dataset*;

T19) *Ranking-based Method for News Stance Detection* [Zhang et al. 2018]: Mais uma pesquisa relacionada ao primeiro desafio (FNC-1). A solução do artigo é criada a partir de uma rede neural *Multi-Layer Perceptron*. Os resultados foram comparados com as demais ferramentas participantes do referido desafio;

T20) *Real-time Detection of Content Polluters in Partially Observable Twitter Networks* [Nasim et al. 2018]: Esta pesquisa procura encontrar um tipo específico de *bots*, chamados de poluidores de conteúdo, para poder distinguir notícias verdadeiras de *Fake News*. Segundo o artigo, o estado da arte de detecção de *bots*, normalmente, necessita de um histórico completo da rede. Assim, o trabalho propõe uma abordagem baseada em informações parciais onde, ao invés de mapear um grafo com seguidores e seguidos, utiliza um grafo com a (dupla de Usuário) x (Evento). Esta dupla é obtida a partir do momento em que o par tenha *tweetado* no mesmo dia do evento. Desta forma, os dados são clusterizados para que os usuários possam ser classificados como *bots* pela análise dos respectivos perfis e a frequência dos *tweets*. Os resultados do trabalho foram comparados com os obtidos por uma ferramenta citada pelo artigo, denominada de *Truthy*;

T21) *Sentiment Aware Fake News Detection on Online Social Networks* [Ajao et al. 2019]: O trabalho se aplica tanto a *Fake News* como Rumor. Assim, o artigo propõe a hipótese de que existe uma relação entre mensagens falsas ou rumores com os sentimentos dos textos. Foram utilizados dois modelos para extrair os escores de emoção (positividade, negatividade ou neutralidade) do texto: *Latent Semantic Analysis (LSA) e Latent Dirichlet Allocation (LDA)*. O objetivo foi desenvolver um classificador que utilize os escores de sentimento. Assim, utilizando classificadores distintos, compara os resultados a partir da abordagem proposta com sentimentos;

T22) *This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News* [Horne and Adali 2017]: Detecção por meio da análise do texto. Este trabalho usa um classificador SVM e compara os seus resultados entre detecção de *Fake News, Real News* e Sátira. Este estudo determinou que as *Fake News* são mais próximas das Sátiras do que as notícias Reais;

T23) *Towards News Verification: Deception Detection Methods for News Discourse* [Rubin et al. 2015]: O trabalho propõe a ferramenta RST-SVM que analisa a notícia para extrair o estilo por meio da combinação do *Rhetorical Structure Theory (RST) e Vector Space Modeling (VSM)* para Clusterização. A detecção da notícia como enganosa ou real foi feita por meio de um classificador SVM. Os resultados obtidos foram comparados com a detecção humana;

T24) *Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate* [Wu and Liu 2018]: Este trabalho busca a detecção de *Fake News*, pela modelagem da propagação da notícia através da mineração de grafos em Florestas. Segundo o artigo, classificar notícias pelo seu conteúdo é muito difícil, devido à atual similaridade entre as divulgações *fake* e não *fake*. Em contra partida, as *Fake News* tendem a ter as mesmas fontes e sequências. O trabalho propõe a ferramenta paralelizável chamada TraceMiner que utiliza *Recurrent Neural Networks* (LSTM-RNNs), para classificar o caminho de propagação das mensagens no *Twitter*. O artigo comparou os seus resultados por meio de técnicas de análise de conteúdo criadas com SVM e XGBoost;

T25) *Weakly Supervised Learning for Fake News Detection on Twitter* [Helmstetter and Paulheim 2018]: Neste estudo, como existe uma dificuldade em conseguir um grande volume de dados para análise *(datasets)*, os *tweets* são rotulados, automaticamente, durante a coleta, de acordo com a confiança na sua fonte. Assim é criado um *dataset*, denominado de *Large-scale Training Dataset*, onde cada tweet de uma fonte confiável é rotulado como uma notícia real, assim como, cada tweet de uma fonte não confiável é rotulado como uma *Fake News*. Esperasse que neste *dataset* a classe de notícias reais contenha apenas uma quantidade negligenciável de ruído, pois fontes confiáveis raramente divulgam *Fake News*. Também é criado um segundo *dataset*, denominado de *Small-scale Evaluation Dataset*, possuindo *tweets* rotulados manualmente, como *fake* e não *fake*, a partir do site PolitiFact [6]. O objetivo principal do trabalho é treinar um classificador, a partir do primeiro *dataset*, para aplicá-lo no segundo *dataset*. Portanto, este classificador, apesar de ter sido treinado em um *dataset* desenvolvido a partir da confiança, é utilizado para detectar *tweets fakes* e não *fakes* no segundo *dataset*. Portanto, o artigo considera que o classificador foi treinado e avaliado com alvos distintos *(weakly supervised learning, mais especificamente, learning with inaccurate supervision)*. Para a referida detecção foram levadas em consideração as características do usuário (ex: engajamento e qtd seguidores), do tweet (ex: dia da semana, hora e texto), do tópico (assunto) e do sentimento. Como algoritmos de aprendizado foram usados o *Naive Bayes*, Árvores de Decisão, *Support Vector Machines* (SVM) e Redes Neurais. Além disso, foram usados dois *ensemble methods* conhecidos como *Random Forest e XGBoost*. Os resultados foram comparados, utilizando diferentes combinações para os classificadores;

T26) *XFake: Explainable Fake News Detector with Visualizations* [Yang et al. 2019]: O detector *XFake* é composto por 3 *frameworks: MIMIC, ATTN e PERT*. O *MIMIC* é construído para análise de atributos (ex. contexto da notícia e publicador) por meio de uma *deep neural network*. O *ATTN* é para análise semântica através de *pre-trained word embedding*, rede neural convolucional e *self-attention mechanism*. O *PERT* é para análise linguística utilizando um classificador *XGBoost*. A ferramenta, além de realizar as predições, também possui um módulo de interface para prover os usuários de explicações sobre as predições. O *XFake* é implementado em Python e deployed em *FLASK* com *front-end em HTML*. Para comparar seus resultados, o trabalho utilizou mão-de-obra humana realizada pela Amazon Mechanical Turk [7].

[6] https://www.politifact.com/

[7] https://www.mturk.com/

Tabela 2.1: Comparação entre abordagens - Dados de Publicação

Id	Dados							
	Publicação							
	Notícia		Usuário			Assunto		Temporalidade
	Mídia (Texto, Áudio e Imagem)	Conteúdo (Léxica, Sintática, Semântica e Legibilidade)	Tipo (Humano, Bot e Cyborg)	Perfil	Reputação	Especificidades	Relevância	
T1	Texto	Léxica e Semântica						
T2	Texto	Léxica, Sintática, Semântica e Legibilidade						
T3	Texto e Imagem	Léxica						
T4	Texto	Léxica e Semântica		X				X
T5	Texto	Léxica e Semântica		X	X			
T6	Texto	Léxica		X				
T7	Texto	Semântica						
T8	Texto	Léxica e Semântica		X	X			X
T9					X	Assuntos controversos		
T10	Texto e Imagem	Léxica						
T11				X				X
T12	Texto	Léxica e Semântica						
T13								
T14					X			X
T15								
T16	Texto	Léxica e Semântica		X				X
T17	Texto	Léxica e Semântica		X		Relaciona Assuntos		
T18	Texto	Semântica						
T19	Texto	Semântica						
T20			Bot	X				X
T21	Texto	Léxica e Semântica						
T22	Texto	Léxica e Sintática						
T23	Texto	Semântica						
T24					X			
T25	Texto	Léxica e Semântica		X		Análise dos Tópicos		X
T26	Texto	Léxica e Semântica		X		Análise dos Tópicos		

Tabela 2.2: Comparação entre abordagens - Dados de Propagação

Id	Dados								
	Propagação								
	Contribuição		Usuário			Assunto		Temporalidade	Rede
	Mídia (Texto, Áudio e Imagem)	Conteúdo (Léxica, Sintática, Semântica e Legibilidade)	Tipo (Humano, Bot e Cyborg)	Perfil	Reputação	Especificidades	Relevância		
T1									
T2									
T3	Texto	Léxica							X
T4	Texto	Léxica e Semântica		X				X	X
T5				X	X				X
T6				X					X
T7									
T8	Texto	Léxica e Semântica		X	X			X	X
T9									
T10									
T11				X				X	X
T12									
T13									X
T14					X			X	X
T15									X
T16				X				X	X
T17									
T18	Texto	Semântica							
T19									
T20			Bot	X				X	X
T21	Texto	Léxica e Semântica							
T22									
T23									
T24					X				X
T25	Texto	Léxica e Semântica							
T26									

Tabela 2.3: Comparação entre abordagens - Modelo, Funcionalidade e Atuação

Id	Modelo			Funcionalidade				Atuação (Centralizada ou Descentralizada)
				Detecção		Intervenção		
	Não Supervisionado	Semi Supervisionado	Supervisionado	Autenticidade	Intencionalidade	Bloqueio (Reativa)	Mitigação (Proativa e Reativa)	
T1			X	X				Centralizada
T2			X	X				Centralizada
T3			X	X				Centralizada
T4			X	X	Análise das características dos usuários			Centralizada
T5		X		X	Pontuação de credibilidade para os usuários			Centralizada
T6			X	X				Centralizada
T7			X	X				Centralizada
T8			X	X	Score para os usuários			Centralizada
T9		X		X	Atribui pesos de confiança aos websites			Centralizada
T10			X	X				Centralizada
T11			X	X				Centralizada
T12			X	X				Centralizada
T13							Reativa	Centralizada
T14			X	X		X		Centralizada
T15							Reativa	Centralizada
T16		X		X				Centralizada
T17			X	X	Associação com o usuário			Centralizada
T18			X	X				Centralizada
T19			X	X				Centralizada
T20		X		X	identificação de bots			Centralizada
T21			X	X	Análise de Sentimentos			Centralizada
T22			X	X				Centralizada
T23		X		X				Centralizada
T24			X	X	Relação entre os usuários			Centralizada (pode ser paralelizada)
T25			X	X	Análise de Sentimentos			Centralizada
T26			X	X				Centralizada

2.3.3. Datasets

Apesar da relevância do problema de combate às *Fake News* nas redes sociais, os *datasets* que contêm dados reais ainda estão raramente disponíveis para download. Como consequência, a maioria das pesquisas relacionadas ao combate às *Fake News* adaptou *datasets* originalmente criados para investigar outros problemas em redes sociais, como divulgação de *Rumor*. Esses datasets adaptados, geralmente, não contêm informações importantes para a detecção de *Fake News*, como rótulos *fake* / não *fake*. Além disso, a maioria desses *datasets*, adaptados ou originalmente criados para detecção de *Fake News*, não descrevem a propagação das notícias nas redes sociais, como uma mesma notícia divulgada por vários usuários e várias notícias divulgadas por um mesmo usuário. Assim, não há um consenso sobre os *datasets* de referência para este problema [Shu et al. 2017a]. Outro fator complicador para a criação de *datasets* é a carência de informação, proveniente das redes sociais, para combate às *Fake News*. Tal carência acontece pois, muitas vezes estas informações são apagadas, impossibilitando a sua análise [Mustafaraj and Metaxas 2017].

Independente das informações fornecidas pelo *dataset*, cabe salientar as diferentes formas pelas quais os referidos dados são disponibilizados:

- No Dataset: a informação está armazenada na própria base de dados;
- Link para o dado: a informação não está armazenada na base de dados, mas o

dataset disponibiliza um link direto para o dado específico;

• Link para a notícia: Nesta caso, o *dataset* simplesmente disponibiliza o link para a notícia. Assim, se faz necessário o acesso à notícia original para a retirada das informações desejadas.

Com o objetivo de apresentar alguns *datasets*, a Tabela 2.4 relaciona os trabalhos apresentados na Subseção 2.3.2 com os seus respectivos *datasets*. Em seguida, a Tabela 2.5 enquadra estes repositórios de acordo com os dados fornecidos por cada um deles. Este enquadramento é realizado no Modelo Comparativo, restrito ao aspecto *Dados*, tratado na Subseção 2.3.1.1. Cabe ressaltar que, na Tabela 2.5, as células não preenchidas indicam o não fornecimento do respectivo dado no *dataset* correspondente.

Além disso, os referidos *datasets* são brevemente descritos, podendo seus detalhes serem consultados através das respectivas referências:

D1) *BS Detector* [Shu et al. 2017a]: Este *dataset* é coletado de uma extensão de *browser* chamada *BS Detector* que foi desenvolvido para checagem da veracidade de notícias. Os rótulos existentes são *"Fake news", "Satire", "Extreme bias", "Conspiracy theory", "Rumor mill", "State news", "Junk science", "Hate group"e "Clickbait"*;

D2) *BuzzFace* [Santia and Williams 2018]: Este repositório foi criado pela equipe do BuzzFeed. Ele contém 2.282 artigos rotulados como *"Mostly true", "Mixture of true and false", "Mostly false"e "No factual content"*;

D3) *BuzzFeedNews (2016-10-facebookfact-check modificado)*
[Janze and Risius 2017]: Conjunto de dados criado a partir do *BuzzFeedNews (2016-10-facebookfact-check)* (D4), contudo os artigos são rotulados com *"Fake"e "Non-Fake"*;

D4) *BuzzFeedNews (2016-10-facebookfact-check)* [Shu et al. 2017a]: Este *dataset* compreende as notícias, do *Facebook*, oriundas de nove agências para a eleição presidencial americana de 2016. Os eventos e artigos ligados foram checados por jornalistas do *BuzzFeed*. Ele contém 1.627 artigos rotulados como *"Mostly true", "Mixture of true and false", "Mostly false"e "No factual content"*;

D5) *Celebrity* [Pérez-Rosas et al. 2018]: Este *dataset* fornece os dados da notícia para análise de texto. As notícias verdadeiras e falsas foram retiradas da *Web*, sendo relacionadas com assuntos de celebridades;

D6) *CredBank* [Shu et al. 2017a]: Conjunto de dados criado a partir do cruzamento de várias fontes, com aproximadamente 60 milhões de *tweets*, que cobrem 96 dias, iniciados em outubro de 2015. Todos os *tweets* são relacionados com mais de 1.000 eventos de notícias. Cada evento foi avaliado por 30 anotadores da *Amazon Mechanical Turk*. Os rótulos existentes são *"[-2]Certainly inaccurate", "[-1]Probably inaccurate", "[0] Uncertain (doubtful)", "[+1] Probably accurate"e "[+2] Certainly accurate"*;

D7) *DataSet Emergent* [Zhang et al. 2018][Bhatt et al. 2018]: Neste repositório, as notícias são rotuladas como *"Agree"* (o texto do corpo concorda com a manchete), *"Disagree"* (o texto do corpo discorda da manchete), *"Discuss"* (o texto do corpo discute a mesma afirmação que o título, mas não toma uma posição) e *"Unrelated"* (o texto do corpo discute uma alegação que difere do título). Esta base faz parte do primeiro desafio (FNC-1) e foi criado a partir do *dataset* para detecção de rumor chamado *Emergent*;

D8) *DistrustRank Datasets* [Woloszyn and Nejdl 2018]: Foram desenvolvidos dois *datasets*. O primeiro, gerado com sites confiáveis, por meio do *SimilarWeb* [8], tem 502 domínios e 396.422 *URLs* de notícias. O segundo, obtido com sites não confiáveis, através do *Wikipedia's list of prominent Fake News* [9], possui 47 dominios e 37.320 *URLs* de notícias;

D9) *Facebook para Detective* [Tschiatschek et al. 2018]: Repositório que considera os círculos sociais do *Facebook*, consistindo de 4.039 usuários (nós) e 88.234 arestas;

D10) *Fake News vs Satire* [Golbeck et al. 2018]: *DataSet* para diferenciar *Fake News* e Sátiras onde as notícias são codificadas manualmente. A base, oriunda de diversas fontes, é composta por 283 relatos rotulados como *Fake News* e 203 como *Satirical*. Estes relatos são compostos pelo título, texto e um link para cada artigo;

D11) *FakeNewsAMT* [Pérez-Rosas et al. 2018]: As notícias falsas e legítimas são fornecidas em duas pastas separadas. Cada pasta contém 40 notícias de seis domínios diferentes: tecnologia, educação, negócios, esportes, política e entretenimento;

D12) *FakeNewsData1* [Horne and Adali 2017]: São dois *datasets* onde o primeiro contém notícias rotuladas como *Fake e Real* retiradas a partir do *BuzzFeed*. Já o segundo contém notícias políticas rotuladas como *Real, Fake* e Sátira obtidas, randomicamente, durante as eleições americanas de 2016;

D13) *FakeNewsNet1* [Shu et al. 2017a] [Shu et al. 2019b] [Sharma et al. 2019] [Gupta et al. 2018] [Shu et al. 2019a]: Esta base de dados, coletada do Twitter, fornece 211 notícias *Fake* e 211 notícias *Real*, rotuladas a partir do *BuzzFeed* e *PolitiFact*;

D14) *FakeNewsNet2* [Shu et al. 2018][Sharma et al. 2019]: Esta base de dados, coletada do Twitter, fornece 6.480 notícias *Fake* e 17.441 notícias *Real*, rotuladas a partir do *GossipCop* [10] e *PolitiFact*;

D15) *Kaggle* [11]: Este conjunto de dados contém texto e metadados de 244 sites, totalizando 12.999 postagens. Os dados foram extraídos usando a *API webhose.io*. Cada site foi rotulado de acordo com o *BS Detector*, sendo que as fontes de dados sem rótulo foram categorizadas como "Bs";

D16) *KV* [Dong et al. 2014]: Nesta base as notícias têm sujeito, predicado e objeto. Cada notícia tem um rótulo que indica a probabilidade da mesma ser verdadeira. A ferramenta, por meio de uma fusão de conhecimentos, cria um grafo relacionando o sujeito com o objeto para medir a quantidade de interações e, assim, gerar automaticamente o *dataset*;

D17) *Large-scale Training Dataset e Small-scale Evaluation Dataset* [Helmstetter and Paulheim 2018]: No *Large-scale Training Dataset* cada tweet de uma fonte confiável é rotulado como notícia real e cada tweet de um uma fonte não confiável é rotulado como uma *Fake News*. As 46 fontes confiávies e 65 não confiáveis foram obtidas através de pesquisas em sites e os *tweets* foram coletados a partir destas fontes. No total,

[8]https://www.similarweb.com/top-websites/category/News-and-media

[9]https://en.wikipedia.org/wiki/List-of-fake-News-websites

[10]https://https://www.gossipcop.com/

[11]https://www.kaggle.com/datasets

foram coletados 401.414 exemplos, nos quais 110.787 (27,6 porcento) foram rotulados como *Fake News*, enquanto 290.627 (24,4 porcento) foram rotulados como *Real News*. O *Small-scale Evaluation Dataset* contém 116 *tweets* rotulados manualmente e obtidos no *PolitiFact*;

D18) *LIAR* [Wang 2017]: Esta base de dados é coletada do *PolitiFact*. Ele inclui 12.836 notícias rotuladas manualmente como *"Pants-fire", "False", "Barely-true", "Half-true", "Mostly true"e "True"*. Cabe salientar que os dados referentes ao usuário se resumem ao nome do autor da postagem;

D19) *PoliticalNews* [Castelo et al. 2019]: Para criar o dataset foram usados os sites *Politifact, BuzzFeed, OpenSources.co e Alexa's top 500 news* [12]. O resultado foi um *dataset* com 14.240 páginas de notícias sendo 7.136 páginas vindas de 79 sites não confiáveis e 7.104 vindos de 58 sites confiáveis;

D20) *PolitiFact para XFake* [Yang et al. 2019]: Repositório criado a partir do site *PolictFact* com 5.104 notícias contendo os atributos *Subject, Context, Speaker, Targeting e Statement*. As notícias foram rotuladas como *True e False*;

D21) *RST-SVM Dataset* [Rubin et al. 2015]: Esta base de dados foi criada a partir de codificadores, usando notícias do *Bluff the Listener* [13]. Este repositório consiste de 144 notícias selecionadas, aleatoriamente, de 2010 até 2014;

D22) *Signal Media para Evaluating Machine Learning Algorithms for Fake News Detection* [Gilda 2017]: *Dataset* rotulado com *"Fake"* ou *"Não fake"* criado a partir de uma base de notícias da *Signal Media* e uma lista do repositório de confiança de fontes *OpenSources.co*. O citado *dataset* contém 11.051 artigos, sendo 3.217 categorizados com falsos;

D23) *Soc-LiveJournal* [Srivastava et al. 2018]: Este repositório não rotulado contém uma rede de relacionamentos formada por 4.847.571 nós e 68.475.391 arestas;

D24) *Twitter e Sina Weibo para CSI* [Ruchansky et al. 2017]: *Dataset* criado com 2.811 artigos rotulados como *"Fake"* e 2.845 como *"True"*. A citada base de dados foi obtida a partir do repositório, para detecção de rumores, gerado no artigo [Ma et al. 2016];

D25) *Twitter e Sina Weibo para EANN* [Ruchansky et al. 2017]: A base de dados foi criada a partir de dois *datasets* não originários de *Fake News*. O primeiro repositório foi obtido a partir do *Sina Weibo* contendo 4.749 notícias com rótulos adaptados para *fake* e 4.779 para real, além de 9.528 imagens. O segundo repositório foi obtido a partir do *Twitter* contendo 7.898 notícias com rótulos adaptados para *fake* e 6.026 para real, além de 514 imagens;

D26) *Twitter e Sina Weibo para Early Detection Through Propagation Path* [Liu and BrookWu 2018]: Este repositório foi criado a partir de três *datasets* usados para detecção de rumores. O primeiro, oriundo da rede social *Weibo*, com os rótulos *"rumor (fake)" e "otherwise (true)"*. Já os outros dois *datasets*, obtidos do *Twitter*, são rotulados como *"fake", "true", "unverified" e "non-rumor (debunking of fake)"*. As características dos usuários foram obtidas por meio de pesquisas realizadas nas respectivas redes sociais.

[12] https://www.alexa.com/topsites/category/News

[13] https://www.npr.org/bluff-the-listener

D27) *Twitter e Sina Weibo para TCNN-URG* [Qian et al. 2018]: Base de dados que utilizou dois *datasets*. O primeiro *dataset* foi obtido, automaticamente, a partir do *Sina Weibo*. Já o segundo *dataset* foi gerado por um processo manual de coleta de dados. Para tal, foram selecionadas notícias em sites avaliados como confiáveis (*The Guardian* [14]) e notoriamente falsos. Com as URLs de todas as notícias coletadas, pesquisas foram realizadas no *Twitter* para cada uma das notícias classificadas como falsas ou reais.

D28) *Twitter para Automatically Identifying Fake News*
[Buntain and Golbeck 2017]: Base de dados que utilizou os *datasets* PHEME (rumor no *Twitter*), CredBank (credibilidade no *Twitter*) e *BuzzFeed News Fact-Checking Dataset* (Checagem de fatos no *Facebook*). Os três *datasets* precisaram ser alinhados com as mesmas características e rótulos;

D29) *Twitter para Content Polluters* [Nasim et al. 2018]: Repositório de dados criado para detecção de *bots*. Este *dataset*, obtido a partir do *Twitter*, foi rotulado manualmente como "Bot"ou "Não Bot";

D30) *Twitter para Mitigation via Point Process* [Farajtabar et al. 2017]: Este trabalho realizou experimentos com contas reais no Twitter e com uma base de dados sintética onde, entre *N* nós, foi assumido que 20 nós criaram *Fake News* e outros 20 nós divulgaram notícias verdadeiras;

D31) *Twitter para TraceMiner* [Wu and Liu 2018]: Conjunto de dados gerado pela coleta de informações do *Twitter* com rotulação a partir do site de checagem de fatos *Snopes* [15]. Nesta base, os rótulos atribuidos são *"Real news"ou "Fake news"*;

D32) *Twitter Trec* [Srivastava et al. 2018]: Conjunto de dados gerado pela coleta de informações do *Twitter*, sem rotulação, contendo uma rede de relacionamentos formada por 3.919.215 nós e 5.399.949 arestas.

Tabela 2.4: Trabalhos x Datasets

Id	DataSet
T1	FakeNewsAMT (D11), Celebrity (D5) e PoliticalNews (D19)
T2	FakeNewsAMT (D11) e Celebrity (D5)
T3	BuzzFeedNews (2016-10-facebookfact-check modificado) (D3)
T4	Twitter para Automatically Identifying Fake News (D28)
T5	FakeNewsNet1 (D13)
T6	FakeNewsNet1 (D13)
T7	DataSet Emergent (D7)
T8	Twitter e Sina Weibo para CSI (D24)
T9	DistrustRank Datasets (D8)
T10	Twitter e Sina Weibo para EANN (D25)
T11	Twitter e Sina Weibo para Early Detection Through Propagation Path (D26)
T12	Signal Media para Evaluating Machine Learning Algorithms for Fake News Detection (D22)
T13	Soc-LiveJournal (D23) e Twitter Trec (D32)
T14	Facebook para Detective (D9)
T15	Twitter para Mitigation via Point Process (D30)
T16	FakeNewsNet1 (D13)
T17	LIAR (D18)
T18	Twitter e Sina Weibo para TCNN-URG (D27)
T19	DataSet Emergent (D7)
T20	Twitter para Content Polluters (D29)
T21	Pheme (dataset para Rumor)
T22	FakeNewsData1 (D12)
T23	RST-SVM Dataset (D21)
T24	Twitter para TraceMiner (D31)
T25	Large-scale Training Dataset e Small-scale Evaluation Dataset (D17)
T26	PolitiFact para XFake (D20)

[14]https://www.theguardian.com/

[15]https://www.snopes.com

Tabela 2.5: Comparação entre Datasets

Id	Dados									URL
	Publicação				Propagação					
	Notícia			Usuário	Contribuição			Usuário	Rede	
	Texto	Áudio	Imagem		Texto	Áudio	Imagem			
D1	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	Link para notícia	Link para notícia	Link para notícia	Link para notícia	https://github.com/thiagovas/bs-detector-dataset
D2	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	https://github.com/gsantia/BuzzFace
D3	No Dataset		Link para imagem	No Dataset	No Dataset			Link para notícia	Link para notícia	
D4	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	Link para notícia	Link para notícia	Link para notícia	Link para notícia	https://github.com/BuzzFeedNews/ 2016-10-facebook-fact-check
D5	No Dataset			No Dataset						http://lit.eecs.umich.edu/downloads.html#undefined
D6	No Dataset			No Dataset				No Dataset	No Dataset	http://compsocial.github.io/CREDBANK-data/
D7	No Dataset			No Dataset						https://github.com/FakeNewsChallenge/fnc-1
D8	Link para notícia	Link para notícia	Link para notícia	No Dataset						
D9				No Dataset				No Dataset		
D10	No Dataset	Link para notícia	Link para notícia	No Dataset						https://github.com/jgolbeck/fakenews
D11	No Dataset			No Dataset						http://lit.eecs.umich.edu/downloads.html#undefined
D12	No Dataset									https://github.com/BenjaminDHorne/fakenewsdata1/ blob/master/Horne2017_FakeNewsData.zip
D13	No Dataset		Link para imagem	No Dataset				No Dataset	No Dataset	https://github.com/KaiDMML/FakeNewsNet
D14	No Dataset		Link para notícia	No Dataset				No Dataset	No Dataset	https://github.com/KaiDMML/FakeNewsNet
D15	No Dataset		Link para imagem	No Dataset						https://www.kaggle.com/mrisdal/fake-news/data
D16	No Dataset			No Dataset						
D17	No Dataset			No Dataset						http://dws.informatik.uni-mannheim.de/en/ research/twitter-fake-news-detection
D18	No Dataset			No Dataset						https://github.com/nishitpatel01/Fake_News_Detection/ tree/master/liar_dataset ou https://www.cs.ucsb.edu/~william/software.html
D19	No Dataset			No Dataset						https://osf.io/ez5q4/
D20	No Dataset			No Dataset						
D21	No Dataset			No Dataset						
D22	No Dataset			No Dataset						
D23									No Dataset	https://snap.stanford.edu/data/soc-LiveJournal1.html
D24	No Dataset			No Dataset	No Dataset			No Dataset	No Dataset	https://github.com/majingCUHK/Rumor_RvNN ou http://alt.qcri.org/~wgao/data/rumdect.zip
D25	No Dataset		No Dataset	No Dataset						
D26				No Dataset				No Dataset	No Dataset	Twitter 15 e 16 (https://www.dropbox.com/s/7ewzdrbelpmrnxu/ rumdetect2017.zip?dl=0) e Weibo(http://alt.qcri.org/~wgao/data/rumdect.zip)
D27	No Dataset			No Dataset	No Dataset					False (https://drive.google.com/open? id=1WRoRVV9j4CStMFKDwP7DVGAHFJZX4t5a) e True(https://drive.google.com/open? id=1JgbW4suN2yWHx65P4QU8HkrrB3OMHsuo)
D28	No Dataset			No Dataset				No Dataset	No Dataset	
D29				No Dataset				No Dataset	No Dataset	
D30									No Dataset	
D31				No Dataset				No Dataset	No Dataset	
D32	No Dataset			No Dataset				No Dataset	No Dataset	https://trec.nist.gov/data/tweets/

2.4. Estudo de Caso em Detecção Automática de *Fake News*

Dentre as abordagens de detecção de *Fake News* apresentadas na Subseção 2.3.2, destacam-se as baseadas na reputação do usuário. Uma das razões para tal destaque é a não necessidade da utilização do conteúdo das notícias. Haja vista que a atual similaridade entre as notícias *fake* e não *fake* [Liu and BrookWu 2018] tem dificultado a detecção de *Fake News* por conteúdo.

Um dos principais trabalhos que utilizam a reputação dos usuários para detectar *Fake News* é [Tschiatschek et al. 2018]. Este trabalho se sobressai, pois a reputação do usuário não é obtida por meio do perfil do usuário na rede social, tendo como base, a dificuldade em se obter tais informações, normalmente de cunho sigiloso [Shu et al. 2017b]. Assim, [Tschiatschek et al. 2018] obtém a reputação do usuário a partir da sua capacidade em sinalizar as notícias. Em resumo, por meio de uma funcionalidade disponível na rede social, o usuário pode opinar se as notícias visualizadas são *fake* ou não. Assim, a reputação do usuário é expressa em função do seu histórico de acertos e erros em suas opiniões. Desta forma, [Tschiatschek et al. 2018] propõe um método chamado *Detective* que classifica uma notícia, como *fake* ou não, a partir de *Crowd Signals*. Este *Crowd* é formado pelas opiniões dos usuários, juntamente com as suas respectivas reputações. Este método, baseado em *Crowd Signals*, em essência, utiliza um classificador bayesiano binário, cujos conceitos básicos e fundamentos são descritos abaixo.

Dada uma rede social $\mathscr{G}$, a entrada do *Detective* contém os seguintes elementos:

um intervalo de tempo t (por exemplo, um dia), um conjunto de usuários U de $\mathscr{G}$, um *dataset* D com notícias rotuladas e uma notícia específica a ser analisada a.

D contém notícias com dois tipos de rótulos: o real e o sinalizado pelo usuário. O rótulo real e o seu valor são indicados pelas variáveis $Y^*(x)$ e $y^*(x)$, respectivamente, onde $y^*(x)$ pertence a $\{f, \overline{f}\}$ onde $y^*(x) = f$ (resp. $y^*(x) = \overline{f}$) significa que uma notícia x é *fake* (resp. não *fake*). Denotado por uma variável $Y_u(x)$, o rótulo sinalizado é aquele atribuído por um usuário u para uma notícia x. Seu valor $y_u(x)$ pertence a $\{f, \overline{f}\}$ onde $y_u(x) = f$ (resp. $y_u(x) = \overline{f}$) significa que u sinalizou x como *fake* (resp. não *fake*). É importante notar que, diferente de outras notícias, $y^*(a)$ é desconhecido e deve ser previsto pelo *Detective*.

Inicialmente, *Detective* aplica as funções $\pi^t(a)$ e $\psi^t(a)$ ao D. Enquanto o primeiro retorna o conjunto de usuários que viram a notícia a no final da época t, o último retorna o conjunto completo de usuários que sinalizaram a como *fake* no final de t.

O *Detective* pode assumir que não há abstinência na sinalização e, para cada usuário $u \in \pi^t(a)$, calcula $\theta_{u,\overline{f}}$ e $\theta_{u,f}$, considerando as notícias sinalizadas por u antes de t. Assim, $\theta_{u,\overline{f}}$ (resp. $\theta_{u,f}$) é a probabilidade de u sinalizar uma notícia x como não *fake* (resp. *fake*), dado que x é realmente não *fake* (resp. fake). Em ambos os casos, o cálculo da probabilidade é limitado ao conjunto de notícias revisadas por u antes de t. Assim, para cada usuário u, *Detective* representa a observada atividade de sinalização de u pela correspondente matriz $\mathscr{M}_u$, genericamente definida como segue.

$$\begin{vmatrix} \theta_{u,\overline{f}} & 1 - \theta_{u,f} \\ 1 - \theta_{u,\overline{f}} & \theta_{u,f} \end{vmatrix}$$

onde:

- $\theta_{u,\overline{f}} = P(Y_u(x) = \overline{f} \mid Y^*(x) = \overline{f})$
- 1 - $\theta_{u,\overline{f}} = P(Y_u(x) = f \mid Y^*(x) = \overline{f})$
- $\theta_{u,f} = P(Y_u(x) = f \mid Y^*(x) = f)$
- 1 - $\theta_{u,f} = P(Y_u(x) = \overline{f} \mid Y^*(x) = f)$

Por fim, seguindo uma abordagem Bayesiana, o *Detective* usa as equações 2 e 3 para calcular as probabilidades de a ser *fake* e não *fake*, respectivamente. Sendo que ω (resp. $1 - \omega$) é a probabilidade a priori de que qualquer notícia seja *fake* (resp. não *fake*). Ambas as equações consideram a capacidade dos usuários de acertar, assim como de errar suas opiniões de acordo com seu voto. Assim, o *Detective* pode se beneficiar quando os usuários acertarem ou errarem, mesmo que eles mostrem incapacidade [Freeman 2017] ou má intenção ao avaliar as notícias. A classe correspondente à maior probabilidade é a opinião do *Detective* sobre a e, portanto, sua saída.

$$P(Y^*(a) = f) = \omega \cdot \prod_{u \in \psi^t(a)} \theta_{u,f} \cdot \prod_{u \in \pi^t(a) \setminus \psi^t(a)} (1 - \theta_{u,f}) \tag{2}$$

$$P(Y^*(a) = \overline{f}) = (1 - \omega) \cdot \prod_{u \in \psi^t(a)} (1 - \theta_{u,\overline{f}}) \cdot \prod_{u \in \pi^t(a) \setminus \psi^t(a)} \theta_{u,\overline{f}} \tag{3}$$

Para este estudo de caso, a metodologia experimental utilizada foi semelhante à metodologia ótima seguida por [Tschiatschek et al. 2018]. As probabilidades θ foram aleatoriamente designadas aos usuários, criando três grupos: *bom* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0.9$), *indiferente* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0.5$) e *spammer* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0,1$). Assim, como realizado em [Tschiatschek et al. 2018], foi assumido que nenhum usuário se absteve de dar sua opinião sobre as notícias a serem analisadas. Embora não esteja claramente indicado em [Tschiatschek et al. 2018], também foi assumido que cada usuário deveria sinalizar aleatoriamente uma notícia de acordo com a probabilidade atribuída ao seu grupo. Por exemplo: dada uma notícia *a* para ser analisada por *u*, um *bom* usuário. De acordo com a configuração definida para *bons* usuários, *u* deve acertar ou errar para *a* com probabilidades de 90% e 10%, respectivamente. Além disso, foi usado o método da roleta para decidir se cada usuário deve acertar ou errar o rótulo real de uma notícia. Embora pouco realista, essa metodologia leva aos resultados de maior precisão produzidos pelo *Detective*.

Para a execução dos nossos experimentos, foram escolhidos os *datasets* BuzzFeed e PolitiFact, ambos pertencentes ao repositório FakeNewsNet1 (D13), descrito na Subseção 2.3.3. Nossa escolha foi guiada por três razões principais. Primeiro, estes *datasets* foram criados para o específico propósito de detecção de *Fake News* e contêm, para cada notícia, seu rótulo real, ou seja, a indicação de que a notícia é *fake* ou não. Em segundo lugar, eles descrevem a propagação das notícias nas redes sociais. Por fim, eles foram usados e disponibilizados por publicações recentes e relevantes [Sharma et al. 2019] [Shu et al. 2017a] [Shu et al. 2019b] [Shu et al. 2019a] [Gupta et al. 2018]. A Tabela 2.6 fornece uma visão estatística geral dos *datasets* escolhidos.

Tabela 2.6: *Datasets* usados nos Experimentos

Dataset	Não *Fake News*	*Fake News*	Usuários	Média de usuários por notícia
BuzzFeed	91	91	15257	125,16
PolitiFact	120	120	23865	136,63

Para se obter os resultados do método *Detective* foi utilizada a acurácia como métrica de desempenho. A Tabela 2.7 ilustra o resultado de uma rodada de execução do experimento.

Tabela 2.7: Acurácia do método de detecção de *Fake News*

Método	BuzzFeed	PolitiFact
Detective	**0.9890**	**0.9791**

2.5. Problemas em Aberto

O combate automático às *Fake News* em redes sociais é uma nova e emergente área de pesquisa que, mesmo com estudos já realizados, ainda carece de maior aprofundamento científico. Desta forma, os seguintes problemas são descritos como áreas ainda fertéis para o desenvolvimento de novos trabalhos:

• Carência de *datasets* que forneçam, de forma suficiente, os diferentes dados necessários para combater as *Fake News* em redes sociais;

• Trabalhos que levem em consideração aspectos temporais do ciclo de vida da *Fake News* e que, consequentemente, possam intervir mais rapidamente;

• Estudos que analisem o aspecto intencional, assim não se limitam a verificar a autenticidade (veracidade) das notícias;

• Extração de características a partir de imagem e/ou áudio, portanto não se limitando as análises de texto;

• Métodos que abordem características baseadas na rede que representa a propagação da notícia. Neste caso, inclusive, podem ser aplicadas técnicas baseadas em grafos;

• Pesquisas que, ao invés de realizarem uma classificação binária, utilizem probabilidades e/ou pertinências na detecção. Esta linha de trabalho se baseia no fato de que, normalmente, a *Fake News* é uma mistura de afirmações falsas e verdadeiras;

• Utilização de um comitê de classificadores para determinar se uma notícia é *fake*. Desta forma, pode-se agregar diferentes técnicas de classificação na detecção;

• Utilização de modelos não supervisionados ou semi-supervisionados devido à carência de *datasets* rotulados que possuam variedade de dados;

• Estudo sobre o comportamento distinto da *Fake News* em diferentes comunidades (escolar, trabalho e etc) e/ou redes sociais *(Weibo, WhatsApp e etc)*. Isto se deve pela possível mudança de comportamento das notícias de acordo com o meio;

• Classificar os usuários de *Fake News* com o objetivo de identificar o seu tipo *(humanos, bots e cyborgs)*. Isto se deve pela possível alteração de comportamento das notícias propositalmente falsas de acordo com o seu tipo de usuário divulgador;

• Trabalhos relacionados à intervenção de *Fake News*, tanto para bloqueio quanto para mitigação. Haja vista que o combate às *Fake News* não se limita à detecção, sendo necessária, também, a intervenção sobre a mesma;

• Abordagens que atuem descentralizadas na rede. Esta atuação se destaca, pois quanto mais rápido e extensivo for o combate, menor serão os efeitos nocivos da notícia;

• Abordagens que utilizem o assunto para a análise da notícia, pois assuntos relevantes, normalmente, motivam a criação de notícias intencionalmente falsas;

• Pesquisas que levem em consideração a reputação dos usuários, pois usuários com baixa reputação tendem a ser potenciais divulgadores de *Fake News*.

2.6. Considerações Finais

Cada vez mais pessoas estão consumindo notícias das redes sociais, ao invés dos canais tradicionais. Tal tendência amplificou a disseminação de *Fake News*, isto é, as notícias falsas publicadas intencionalmente. Este tipo de notícia pode ter significativos impactos sociais negativos, por exemplo, a manipulação da opinião em larga escala.

Tendo como base os riscos que as *Fake News* trazem para a sociedade, tanto a academia quanto a indústria buscam por soluções que viabilizem o combate a este tipo de notícia nas redes sociais. Ademais, devido tanto ao volume como à velocidade de divulgação das *Fake News*, faz-se necessário o emprego de abordagens computacionais no combate às notícias intencionalmente falsas nas redes sociais.

Portanto, neste capítulo, nós exploramos o combate automático às *Fake News* em redes sociais. Para tal, revisamos a literatura existente, visando realizar um levantamento das abordagens computacionais, tendo como base a proposta de um modelo comparativo. Em seguida, um estudo de caso foi realizado, objetivando uma introdução, não só teórica, como também prática. Por fim, foram apresentados alguns problemas sobre o referido combate que ainda carecem de maior aprofundamento científico.

Referências

[Ajao et al. 2019] Ajao, O., Bhowmik, D., and Zargari, S. (2019). Sentiment aware fake news detection on online social networks. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2507–2511.

[Bhatt et al. 2018] Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., and Mittal, A. (2018). Combining neural, statistical and external features for fake news stance identification. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 1353–1357, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Braz and Goldschmidt 2017] Braz, P. and Goldschmidt, R. (2017). Um método para detecção de bots sociais baseado em redes neurais convolucionais aplicadas em mensagens textuais. In SBSeg 2017, pages 501–508. 10/11/2017.

[Buntain and Golbeck 2017] Buntain, C. and Golbeck, J. (2017). Automatically identifying fake news in popular twitter threads. In 2017 IEEE International Con on Smart Cloud (SmartCloud), pages 208–215.

[Campan et al. 2017] Campan, A., Cuzzocrea, A., and Truta, T. M. (2017). Fighting fake news spread in online social networks: Actual trends and future research directions. In 2017 IEEE International Con on Big Data (Big Data), pages 4453–4457.

[Castelo et al. 2019] Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., and Freire, J. (2019). A topic-agnostic approach for identifying fake news pages. In Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, pages 975–980, New York, NY, USA. ACM.

[Cazalens et al. 2018] Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., and Tannier, X. (2018). A content management perspective on fact-checking. In Companion

Proceedings of the The Web Con 2018, WWW '18, pages 565–574, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Ciampaglia et al. 2015] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. PLOS ONE, 1:1–13.

[Conroy et al. 2015] Conroy, N., Rubin, V., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Association for Information Science and Technology, 52:1–4.

[Deng et al. 2014] Deng, S., Huang, L., and Xu, G. (2014). Social network-based service recommendation with trust enhancement. Expert Systems with Applications, 41(18):8075 – 8084.

[Dong et al. 2014] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In ACM SIGKDD international Con on Knowledge discovery and data mining, pages 601–610.

[Farajtabar et al. 2017] Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., and Zha, H. (2017). Fake news mitigation via point process based intervention. In Proceedings of the 34th International Con on Machine Learning - Volume 70, ICML'17, pages 1097–1106. JMLR.org.

[Ferrara et al. 2016] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. Commun. ACM, 59(7):96–104.

[Flintham et al. 2018] Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., and Moran, S. (2018). Falling for fake news: Investigating the consumption of news via social media. In Proceedings of the 2018 CHI Con on Human Factors in Computing Systems, CHI '18, pages 376:1–376:10, New York, NY, USA. ACM.

[Freeman 2017] Freeman, D. M. (2017). Can you spot the fakes?: On the limitations of user feedback in online social networks. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, pages 1093–1102, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Gilda 2017] Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In 2017 IEEE 15th Student Con on Research and Development (SCOReD), pages 110–115.

[Golbeck et al. 2018] Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., Falak, W., Gieringer, C., Graney, J., Hoffman, K. M., Huth, L., Ma, Z., Jha, M., Khan, M., Kori, V., Lewis, E., Mirano, G., Mohn IV, W. T., Mussenden, S., Nelson, T. M., Mcwillie, S., Pant, A., Shetye, P., Shrestha, R., Steinheimer, A., Subramanian, A., and Visnansky, G. (2018). Fake news vs satire: A dataset and analysis. In Proceedings of

the 10th ACM Con on Web Science, WebSci '18, pages 17–21, New York, NY, USA. ACM.

[Gupta et al. 2018] Gupta, S., Thirukovalluru, R., Sinha, M., and Mannarswamy, S. (2018). Cimtdetect: A community infused matrix-tensor coupled factorization based method for fake news detection. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 278–281.

[Helmstetter and Paulheim 2018] Helmstetter, S. and Paulheim, H. (2018). Weakly supervised learning for fake news detection on twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 274–277.

[Hendrikx et al. 2015] Hendrikx, F., Bubendorfer, K., and Chard, R. (2015). Reputation systems: A survey and taxonomy. Journal of Parallel and Distributed Computing, pages 184–197.

[Horne and Adali 2017] Horne, B. D. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Association for the Advancement of Artificial Intelligence.

[Janze and Risius 2017] Janze, C. and Risius, M. (2017). Automatic detection of fake news on social media platforms. In PACIS 2017.

[Kim et al. 2018] Kim, J., Tabibian, B., Oh, A., Schölkopf, B., and Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In Proceedings of the Eleventh ACM International Con on Web Search and Data Mining, WSDM '18, pages 324–332, New York, NY, USA. ACM.

[Kshetri and Voas 2017] Kshetri, N. and Voas, J. (2017). The economics of fake news. IT Professional, 19(06):8–12.

[Li et al. 2015] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2015). A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 17:1–16.

[Liu and BrookWu 2018] Liu, Y. and BrookWu, Y. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In AAAI Con on Artificial Intelligence, pages 354–361.

[Liu and Xu 2016] Liu, Y. and Xu, S. (2016). Detecting rumors through modeling information propagation networks in a social media environment. IEEE Transactions on Computational Social Systems, 3(2):46–62.

[Ma et al. 2016] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K., and Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In International Joint Con on Artificial Intelligence.

[Ma et al. 2015] Ma, J., Gao, W., Wei, Z., Lu, Y., and Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In Proceedings of the 24th ACM International on Con on Information and Knowledge Management, CIKM '15, pages 1751–1754, New York, NY, USA. ACM.

[Mustafaraj and Metaxas 2017] Mustafaraj, E. and Metaxas, P. T. (2017). The fake news spreading plague:was it preventable? In Web Science Con, pages 236–239.

[Nasim et al. 2018] Nasim, M., Nguyen, A., Lothian, N., Cope, R., and Mitchell, L. (2018). Real-time detection of content polluters in partially observable twitter networks. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 1331–1339, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Pérez-Rosas et al. 2018] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In International Conference on Computational Linguistics, pages 3391–3401.

[Qian et al. 2018] Qian, F., Gong, C., Sharma, K., and Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. In International Joint Con on Artificial Intelligence, pages 3834–3840.

[Reis et al. 2019] Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Explainable machine learning for fake news detection. In Proceedings of the 10th ACM Conference on Web Science, WebSci '19, pages 17–26, New York, NY, USA. ACM.

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

[Rubin et al. 2015] Rubin, V. L., Conroy, N. J., and Chen, Y. (2015). Towards news verification: Deception detection methods for news discourse.

[Ruchansky et al. 2017] Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Con on Information and Knowledge Management, CIKM '17, pages 797–806, New York, NY, USA. ACM.

[Santia and Williams 2018] Santia, G. C. and Williams, J. R. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In AAAI Con on Web and Social Media, pages 531–540.

[Seo J. 2013] Seo J., Choi S., H. S. (2013). The method of trust and reputation systems based on link prediction and clustering. In IFIP International Con on Trust Management, pages 223–230.

[Sethi 2017] Sethi, R. J. (2017). Crowdsourcing the verification of fake news and alternative facts. In Proceedings of the 28th ACM Con on Hypertext and Social Media, HT '17, pages 315–316, New York, NY, USA. ACM.

[Sharma et al. 2019] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol., 10(3):21:1–21:42.

[Sherchan et al. 2013] Sherchan, W., Nepal, S., and Paris, C. (2013). A survey of trust in social networks. ACM Comput. Surv., 45(4):47:1–47:33.

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

[Shu et al. 2018] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. In arXiv.

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

[Shu et al. 2017b] Shu, K., Wang, S., and Liu, H. (2017b). Exploiting tri-relationship for fake news detection. In arXiv.

[Shu et al. 2019b] Shu, K., Wang, S., and Liu, H. (2019b). Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, pages 312–320, New York, NY, USA. ACM.

[Srivastava et al. 2018] Srivastava, A., Kannan, R., Chelmis, C., and Prasanna, V. K. (2018). Factcheck: Keeping activation of fake news at check. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '18, pages 2079–2081, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Tschiatschek et al. 2018] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 517–524, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Vavilis et al. 2014] Vavilis, S., PetkoviÄ‡, M., and Zannone, N. (2014). A reference model for reputation systems. Decision Support Systems, 61:147 – 154.

[Vo and Lee 2018] Vo, N. and Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, pages 275–284, New York, NY, USA. ACM.

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

[Wang et al. 2018a] Wang, P., Angarita, R., and Renna, I. (2018a). Is this the era of misinformation yet: Combining social bots and fake news to deceive the masses. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 1557–1561, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Wang 2017] Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

[Wang et al. 2018b] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018b). Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, pages 849–857, New York, NY, USA. ACM.

[Woloszyn and Nejdl 2018] Woloszyn, V. and Nejdl, W. (2018). Distrustrank: Spotting false news domains. In Proceedings of the 10th ACM Con on Web Science, WebSci '18, pages 221–228, New York, NY, USA. ACM.

[Wu and Liu 2018] Wu, L. and Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate. In Proceedings of the Eleventh ACM International Con on Web Search and Data Mining, WSDM '18, pages 637–645, New York, NY, USA. ACM.

[Yang et al. 2019] Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. B. (2019). Xfake: Explainable fake news detector with visualizations. In The World Wide Web Conference, WWW '19, pages 3600–3604, New York, NY, USA. ACM.

[Zhang et al. 2018] Zhang, Q., Yilmaz, E., and Liang, S. (2018). Ranking-based method for news stance detection. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 41–42, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Zhou and Zafarani 2018] Zhou, X. and Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. In arXvi.

[Zhou and Zafarani 2019] Zhou, X. and Zafarani, R. (2019). Fake news detection: An interdisciplinary research. In Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, pages 1292–1292, New York, NY, USA. ACM.

[Zhou et al. 2019] Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In Proceedings of the Twelfth ACM International Con on Web Search and Data Mining, WSDM '19, pages 836–837, New York, NY, USA. ACM.