MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

MARCELO PEREIRA DE SOUZA

DETECÇÃO DE FAKE NEWS NO IDIOMA PORTUGUÊS: UM MÉTODO BASEADO EM ANÁLISE DE SENTIMENTOS E CARACTERÍSTICAS LINGUÍSTICAS

RIO DE JANEIRO
2021

MARCELO PEREIRA DE SOUZA

DETECÇÃO DE FAKE NEWS NO IDIOMA PORTUGUÊS: UM MÉTODO BASEADO EM ANÁLISE DE SENTIMENTOS E CARACTERÍSTICAS LINGUÍSTICAS

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Ronaldo Ribeiro Gosldschmidt, D.Sc.

Rio de Janeiro

2021

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80 – Praia Vermelha
Rio de Janeiro – RJ CEP: 22290-270

Souza, Marcelo Pereira de.
DETECÇÃO DE FAKE NEWS NO IDIOMA PORTUGUÊS: UM MÉTODO BASEADO EM ANÁLISE DE SENTIMENTOS E CARACTERÍSTICAS LINGUÍSTICAS / Marcelo Pereira de Souza. – Rio de Janeiro, 2021.
99 f.

Orientador(es): Ronaldo Ribeiro Gosldschmidt.

Dissertação (mestrado) – Instituto Militar de Engenharia, Sistemas e Computação, 2021.

1. APRENDIZADO DE MÁQUINA. 2. DETECÇÃO DE FAKE NEWS. 3. IA. 4. ANÁLISE DE EMOÇÕES. 5. ANÁLISE DE SENTIMENTOS. 6. LINGUÍSTICA. i. Gosldschmidt, Ronaldo Ribeiro (orient.) ii. Título

MARCELO PEREIRA DE SOUZA

DETECÇÃO DE FAKE NEWS NO IDIOMA PORTUGUÊS: UM MÉTODO BASEADO EM ANÁLISE DE SENTIMENTOS E CARACTERÍSTICAS LINGUÍSTICAS

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Ronaldo Gosldschmidt.

Aprovado em Rio de Janeiro, 09 de fevereiro de 2021, pela seguinte banca examinadora:

Prof. **Ronaldo Ribeiro Goldschmidt** - D.Sc do IME - Presidente

Prof. **Prof. Julio Cesar Duarte –** D.Sc do IME

Prof. **Gustavo Paiva Guedes e Silva** – D.Sc do CEFET-RJ

Rio de Janeiro

2021

Este trabalho é dedicado a minha filha Erika,
para inspirá-la a sempre continuar estudando.

AGRADECIMENTOS

Os agradecimentos principais são direcionados primeiramente a minha namorada Ana Cristina que, com sua paixão pela ciência, todo o tempo me incentivou a continuar. Com paciência e atenção ajudou nas revisões, vibrou com as conquistas e consolou pelos revezes, abrindo mão de noites de sono e finais de semana sem sair.

Aos professores Gustavo Paiva Guedes do CEFET-RJ, Jonice Oliveira da UFRJ e Karin Becker da UFRGS e seus respectivos alunos que contribuíram fornecendo informações e detalhes adicionais sobre seus trabalhos que foram referências para este.

Ao meu orientador Prof. Ronaldo Goldshmidt pela paciência, experiência e conhecimento em todas as etapas do processo e ao Prof. Paulo Freire pela assistência na tarefa de orientação e revisão, cujo trabalho foi inspiração para mim.

Agradecimentos especiais são direcionados ao Instituto Militar de Engenharia, que sempre admirei como referência de excelência de instituição de ensino e que me deixou honrado ao me admitir no programa de mestrado.

Finalmente agradeço à Petrobras, empresa em que trabalho, pela importância que dá à formação de seus funcionários.

"Quando o deixei, eu era só o aprendiz,
agora eu sou o mestre."
(Darth Vader a Obi-Wan Kenobi)

RESUMO

Com a popularização da Internet no Brasil, a divulgação de notícias através das mídias digitais tem proporcionado mais acesso à informação. Apesar dos benefícios, as mídias digitais potencializaram um problema antigo, a disseminação intencional de notícias falsas: as denominadas Fake News. Diante desse cenário, destacam-se as abordagens linguísticas para detecção automática de Fake News que utilizam informações que podem ser extraídas diretamente do texto da notícia. Baseados nessas abordagens, foram propostos métodos que utilizam a classificação gramatical e a análise de sentimentos presentes na escrita das notícias em língua portuguesa. Entretanto, até onde foi possível observar, a análise de sentimentos presente nesses trabalhos se limita a utilização da polaridade (i.e. positiva, neutra ou negativa) existente no texto. Tendo como base essa limitação, este estudo propõe um método estendido que, além da classificação gramatical e da análise de sentimentos por polaridade, utiliza a análise da emoção (i.e. raiva, tristeza e etc) e avalia a contribuição de cada elemento na identificação de Fake News. O método estendido apresentou resultados promissores em dados experimentais, obtendo acurácia superior a 92% na identificação de Fake News no domínio da língua portuguesa. No geral, superou o método original em 1,4% ao acrescentar a classificação de emoções.

Palavras-chave: APRENDIZADO DE MÁQUINA. DETECÇÃO DE FAKE NEWS. IA. ANÁLISE DE EMOÇÕES. ANÁLISE DE SENTIMENTOS. LINGUÍSTICA.

ABSTRACT

In the last decades, the dissemination of news through digital media has increased the information accessibility previously offered by traditional channels. Despite their benefits, digital media have exacerbated an old problem: the spread of Fake News, (i.e., false News intentionally published). Faced with this scenario, the linguistic approaches to automatic Fake News detection use information that can be directly extracted from the News' text. Several methods based on these approaches use grammatical classification and sentiment analysis over News writing in Portuguese. However, as far as it was possible to observe in the related literature, these methods are limited to the identification of polarity sentiment (i.e., positive, neutral or negative) existing in the text. Hence, this study proposes an extended method that, in addition to the grammatical classification and polarity based sentiment analysis, also uses the analysis of emotions (i.e., anger, sadness, etc.) to detect Fake News written in Portuguese. The extended method showed promising results in experimental data, obtaining accuracy greater than 92%. In average, the proposed method overcame polarity and gramatical classification based methods in 1.4 percentage points.

Keywords: fake news. machine learning. ai. emotion. sentiment. linguistic.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AB	AdaBoost
CSR	Conjunto de Símbolos Relevantes
DEM	Discrete Emotion Models
DiEM	Dimmensional Emotion Models
FNE	Nome do método proposto
GB	Gradient Boost
KNIME	Konstanz Information Miner
KNN	K-Nearest Neigbour
LIWC	Linguistic Inquiry Word Counting
MDDN	Meios Digitais de Divulgação de Notícias
NB	Naive Bayes
PLN	Processamento de Linguagem Natural
POS	Parts of Speech Tagging
SVM	Support Vector Machine

LISTA DE SÍMBOLOS

α	Função de classificação gramatical
$C_{n.t}$	Matriz com a frequência de termos relevantes
$E_{n.t}$	Matriz da frequência dos termos com emoção associada
L	Léxico
n	Uma notícia individual com seus atributos
N	Conjunto de notícias
Ne_n	Conjunto de atributos estruturados de uma notícia
Ne	Conjunto de atributos estruturados de todo o conjunto de notícias
p	Termo ou palavra
P	Polaridade de um termo ou palavra
t	Texto da notícia
$T_{n.t}$	Matriz com a frequência de tokens em uma classe
$TK_{n.t}$	Conjunto de *tokens* de um texto
CG	Conjunto de Classes Gramaticais

SUMÁRIO

1 INTRODUÇÃO

Os Meios Digitais de Divulgação de Notícias (MDDN) vêm assumindo crescente importância na forma como as pessoas se comunicam e consomem informações (3). MDDNs são as plataformas de redes sociais, os aplicativos de trocas de mensagens ou os sites de notícias online, que vêm se multiplicando[1] de forma acelerada e tomando lugar das mídias tradicionais. Apesar de facilitar o acesso às informações, os MDDN potencializam a circulação de boatos, rumores e todo tipo de notícia falsa, quer seja pela dificuldade de verificação da veracidade dos fatos narrados, quer seja pela rapidez com que as notícias se espalham pelas redes digitais. A desinformação causada por este tipo de informação falsa tornou-se uma preocupação mundial, pelo seu potencial em prejudicar comunidades ou pessoas, criar caos, gerar prejuízos financeiros ou vantagens políticas (4).

1.1 Motivação e Contexto

Dados do *International Telecommunication Union (ITU)*, agência especializada das Organizações das Nações Unidas (ONU) para tecnologias de informação e comunicação, estimavam que, ao final do ano de 2019 mais de 53,6% da população mundial, cerca de 4,1 bilhões de pessoas, estariam conectadas a Internet[2], com acesso a informações de forma *on-line*. Relatórios especializados como *Digital News Report 2019* [1], uma realização conjunta do *Reuters Institute* e da Universidade de Oxford, indicam a contínua migração das fontes de consumo de notícias, em todo o mundo, dos veículos tradicionais para os meios digitais. Apesar disso, no mundo todo é crescente o número de pessoas preocupadas com a desinformação causada pelas *Fake News*. No Brasil 85% [1] dos brasileiros estão preocupados com informações falsas que circulam na Internet.

Outro motivo de preocupação decorre do efeito da proliferação extensiva de notícias falsas, que pode quebrar o equilíbrio do sistema de notícias, mudando a forma como as pessoas respondem às notícias verdadeiras e falsas (3). Por exemplo, no ano seguinte às eleições presidenciais de 2018 no Brasil, observou-se que o nível de confiança no que era divulgado caiu 11% [1]. Mesmo diante do surgimento de diversas agências de checagem de fatos (*Fact Checking*), instituições acreditadas e destinadas a verificar e publicar análises sobre a autenticidade de fatos veiculados nos meios de comunicação, o ceticismo em relação às noticias continuou aumentando.

Um exemplo recente vem com a atual pandemia de COVID-19 (*Coronavirus Disease-19*), em que mesmo diante dos melhores esforços as agências de checagem não

[1] http://www.digitalnewsreport.org/
[2] https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

impediram que movimentos políticos e sociais se aproveitassem do surto da doença para comunicar suas mensagens. Ataques xenofóbicos contra indivíduos e empresas asiáticas, promoções inescrupulosas de remédios prometendo a cura, falsas teorias sobre as origens da doença e todo tipo de notícias dos mais variados temas são objeto de textos falsos (5), definidos de forma generalizada como *Fake News*.

Em busca de mitigar os efeitos nocivos das *Fake News*, devido ao grande volume e a grande velocidade de propagação destas publicações nos meios digitais (6), estudos para o desenvolvimento de abordagens computacionais vêm sendo conduzidos visando incrementar a confiabilidade na identificação do que é falso e aumentar a agilidade e a rapidez na tarefa de detecção destes eventos. Cabe ressaltar que automatização é um conceito importante, pois a velocidade na detecção é característica essencial para mitigar os efeitos negativos causados por esta prática.

1.2 Caracterização do problema

A tarefa de detecção de *Fake News* é um problema para várias frentes de atuação (4). Todos os aspectos devem ser estudados e soluções devem ser apresentadas para cada um deles. A medida que novas ferramentas são criadas para a detecção e/ou intervenção, os perpetradores desta prática encontram maneiras de contornar e continuar com as atividades. Trata-se, portanto, de uma atividade contínua e que necessita a todo momento de novas metodologias e técnicas para continuarmos no enfrentamento destas ameaças.

O aumento da utilização de mídias digitais como fontes de notícias, tem como principais consequências o aumento na velocidade e no volume de proliferação das *Fake News*, assim como com o avanço de novas tecnologias (e.g. uso de robôs) contribuem para a sofisticação dos métodos utilizados para esta prática. Tudo isso contribui para um ecossistema cada vez mais complexo e que necessita de acompanhamento constante, sendo porém uma difícil tarefa se realizada sem a ajuda de mecanismos de automação. Agências de checagem de fatos, campanhas de conscientização e políticas de enfrentamento não dão conta da massa de novas notícias que circulam todos os dias.

Além disso, um dos fatores críticos para evitar os efeitos danosos, ou pelo menos minimizá-los, é a rapidez com que uma *Fake News* é detectada. Desta maneira, outras ações podem ser tomadas, como a retificação da notícia, seu desmentido ou refutação ou a conscientização da população sobre sua veiculação.

As abordagens para a detecção automatizada de *Fake News*, em um nível abstrato, como descrito por Conroy, Rubin e Chen(7) e também por Bondielli e Marcelloni(1), podem ser divididas em dois ramos complementares. Por um lado temos a chamada abordagem de rede, que se baseia em características como origem, autoria, forma de propagação nas redes, reputação dos sites, volume de divulgação, ou ainda a reação do público. Por outro lado

temos a abordagem linguística, quando considera as informações que podem ser extraídas diretamente do texto, como estilo da escrita, polarização em relação ao assunto, expressão de sentimentos e emoções, são exemplos do objetivo desta abordagem. Quando ambas as abordagens são implementadas em conjunto, é chamada abordagem híbrida.

Dando destaque à abordagem linguística, métodos promissores vêm utilizando técnicas de processamento de linguagem natural para identificação de características e extração de atributos tais como a classificação gramatical de termos e palavras, análises de elementos sintáticos e semânticos, assim como o uso de análise de sentimentos (8), que permitem reconhecer e classificar uma notícia falsa, inclusive alguns trabalhos para textos escritos em língua portuguesa (9).

Apesar da expressão "análise de sentimentos" ser caracterizada pela aplicação de técnicas computacionais que identifiquem a polaridade (opinião) e/ou a emoção presente em um texto (10), até onde foi possível observar, os métodos de detecção de *Fake News* desenvolvidos até o momento concentram-se exclusivamente na determinação da polaridade, ou seja, se o texto da notícia em análise expressa opinião positiva, negativa ou neutra. A utilização da análise de sentimentos para identificar *Fake News* restrita apenas à polaridade se apresenta como uma limitação devido, basicamente, a dois fatores. O primeiro é que os resultados dos trabalhos de detecção de *Fake News* baseados em polaridade apontam que há pouca capacidade de separação das notícias em *fake* e não *fake* na tarefa de classificação dos textos (9). Segundo que ao não incluir atributos de emoção deixamos de considerar estudos que indicam que emoções contidas na escrita dos textos podem revelar características psicológicas ou sociais do contexto onde seu autor está inserido (11) (12), podendo, inclusive, identificar se a narrativa possui indícios de não ser verdadeira.

1.3 Objetivos

O presente trabalho levanta a hipótese de que a ampliação do uso de técnicas de análise de sentimentos, em particular com a inclusão da classificação de emoções, associadas a atributos de classificação gramatical, entre outras métricas que podem ser levantadas das notícias, pode viabilizar a construção de modelos de detecção de *Fake News* mais robustos que os existentes na literatura, em especial focando nas nuances da língua portuguesa.

De forma a obter evidências experimentais que apontem para a validade da hipótese levantada, o presente trabalho propõe e avalia um método estendido que identifica e utiliza a emoção presente nos textos das notícias para detectar *Fake News* no idioma português. Nos experimentos realizados, o método proposto apresentou resultados de acurácia superior a 92% na identificação de *Fake News*, valor em média 1,4% superior quando comparado aos métodos que não utilizam a classificação de emoções.

O objetivo geral deste trabalho é o de apresentar um método capaz de identificar a

ocorrência de uma *Fake News*, baseado unicamente no conteúdo do texto desta notícia e que apresente resultados superiores aos métodos do estado da arte nesta área. Queremos evidenciar que, pode-se melhorar a detecção de *Fake News* pela identificação de caraterísticas linguísticas específicas, utilizando ferramentas de análise de sentimentos, técnicas de classificação das emoções presentes nos textos (e não somente a polaridade das mesmas) e a associação de outras características textuais.

Além disso queremos validar uma metodologia baseada em aprendizado de máquina e na implementação e construção de um modelo funcional, capaz de identificar e classificar uma *Fake News* no idioma português. Tal modelo pode ser parte de um arcabouço mais abrangente, de forma a cobrir várias das facetas da criação, divulgação e proliferação das *Fake News*.

1.4 Justificativa

Embora não seja um fenômeno recente, a proliferação de *Fake News*, vem se tornando um problema cada vez mais complexo e de difícil solução. Seu poder de criar caos ou, ainda pior, seu poder de convencer a população a acreditar em situações falsas, torna esta uma arma poderosa que pode desestabilizar governos e nações, alterar relações diplomáticas e até mesmo induzir conflitos. Na ficção, seu uso como ferramenta de controle sobre as massas pode ser muito bem representado pela obra "1984" de George Orwell, onde o Estado impõe um regime extremamente totalitário para a sociedade e exerce seu poder de controle através do "Ministério da Verdade", que tinha como função alterar dados para que toda a história, comunicado e documento estivesse de acordo com o que o partido pregava.

Em um passado recente, cujos efeitos ainda ecoam nos dias atuais, durante a 2ª Grande Guerra Mundial, embora não haja evidências concretas, é atribuído ao ministro da propaganda de Adolf Hitler, Joseph Goebbels, o pensamento de que "Uma mentira contada mil vezes torna-se verdade". Para ele quanto mais uma pessoa era exposta à mentira, mais fácil essa falsa informação era aceita como verdade. Ele teria dito "Temos que fazer o povo crer que a fome, a sede, a escassez e as enfermidades são culpa dos nossos opositores e fazer que nossos simpatizantes repitam isso a todo momento". Entendemos que a guerra da desinformação é uma arma poderosa na mão de governos ou pessoas inescrupulosas.

Em agosto de 2018, o Exército Brasileiro lançou uma campanha de utilidade pública[3] com foco em esclarecer a população sobre o tema, cujo título era "ENTRAMOS NO COMBATE ÀS 'FAKE NEWS', ENTRAMOS NO COMBATE À DESINFORMAÇÃO". Mostrando que se trata de um assunto de segurança nacional o qual merece a devida atenção

[3] https://tinyurl.com/y5csqyuz

de todos, e que nem mesmo o Exército tem sido poupado dos ataques com *Fake News*[4]. Acreditamos que este projeto, em conjunto com outros que estão sendo desenvolvidos, possam auxiliar esta instituição na sua missão de "Contribuir para a garantia da soberania nacional, dos poderes constitucionais, da lei e da ordem, salvaguardando os interesses nacionais e cooperando com o desenvolvimento nacional e o bem-estar social".

1.5 Metodologia

O início do trabalho se deu com a elaboração de uma revisão bibliográfica sobre o estado-da-arte na detecção de *Fake News*, que tivesse como base a análise linguística. Foram usados como fontes de pesquisa os repositórios: IEEE Xplore [5], ACM Digital Library [6], Scopus [7] e Google Schoolar [8]. Para auxiliar na seleção dos trabalhos, foram elaboradas perguntas norteadoras (descritas na sessão de trabalhos relacionados) e os trabalhos que não atendessem aos critérios foram eliminados como referências. A pergunta "O método pode ser utilizado/adaptado para outros idiomas?" obteve especial atenção, caso a pesquisa não fosse originalmente destinada ao idioma português.

Embora a maioria dos trabalhos fossem desenvolvidos para o idioma inglês, muitos deles foram relevantes para entender como nuances e especificidades da língua eram tratadas. Fez parte desta pesquisa avaliar se as técnicas aplicadas poderiam ser indiferentes ao idioma ou se poderiam ser adaptadas/otimizadas para o português. Como critério adicional somente consideramos os trabalhos que incluíram disciplinas de Análise de Sentimentos, tema de especial interesse desta pesquisa. Dentre os trabalhos que utilizavam análise de sentimentos, nenhum deles se valeu da identificação de emoções como atributo dos textos. As diversas pesquisas existentes, relacionadas ao desenvolvimento de ferramentas que permitissem a realização de experimentos no campo das emoções ultrapassaram os limites da ciência da computação e levou às áreas das ciências humanas, onde foram analisados artigos em psicologia e análise comportamental e a descobertas de ferramentas como o LIWC, que se revelou um dos principais componentes para o resultado final deste trabalho.

Na sequência, foram avaliados os algoritmos mais utilizados em processamento de linguagem natural, sua relevância para a obtenção dos resultados e as possíveis lacunas que pudessem ser exploradas na obtenção de melhores resultados.

De forma a permitir que experimentos fossem realizados utilizando dados reais, foi necessária a busca por conjuntos de dados de notícias que tivessem anotações sobre suas classes entre falsas e verdadeiras, que contivessem um número representativo de textos e

[4] https://www.eb.mil.br/fake-news
[5] https://ieeexplore.ieee.org/Xplore/home.jsp
[6] https://dl.acm.org/
[7] https://www.scopus.com/
[8] https://scholar.google.com.br/

que fossem de domínio público. Referências a *datasets* no idioma inglês foram descartadas para cumprir os requisitos previamente levantados, sendo selecionados apenas conjuntos de notícias em português. Partiu-se então para o desenvolvimento de um protótipo de classificação de *Fake News*, que permitisse flexibilidade na escolha das ferramentas, fosse de rápida implementação e permitisse a análise dos resultados de forma eficiente. Após o desenvolvimento do protótipo, foram executados os experimentos sobre os conjuntos de dados relacionados, com a variação das ferramentas e dos algoritmos de classificação, de forma a fornecer subsídios para uma análise completa dos resultados.

Por fim, um artigo sob o título "A Linguistic Based Method that Combines Polarity Emotion and Gramatical Characteristics to Detect Fake News in Portuguese"[9], foi submetido e aceito no Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia 2020)[10], apresentando o modelo e os resultados obtidos e obtendo elogios e aceitação dos organizadores e do público presente.

1.6 Estrutura da dissertação

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os principais conceitos aplicados neste estudo. A Seção 3 apresenta alguns artigos do estado da arte. O método proposto é detalhado na Seção 4. Em seguida, a Seção 5 descreve os experimentos e discute os resultados obtidos. Finalmente, a Seção 6 faz as ponderações finais e destaca as possibilidades de futuras pesquisas.

[9] https://dl.acm.org/doi/10.1145/3428658.3430975
[10] https://webmedia.org.br/2020/

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Fake News

A expressão *Fake News* tornou-se a forma mais amplamente utilizada para identificar uma informação falsa, tanto na mídia convencional como entre a população em geral. Faz-se necessário, entretanto, ressaltar alguns aspectos para que, do ponto de vista desta pesquisa, uma definição mais precisa possa ser formada.

2.1.1 Definição

Fallis(13) distingue a informação falsa sob duas classificações distintas, descritas em inglês pelos termos: *misinformation* e *disinformation*. Entretanto suas traduções literais para o português são idênticas. Vamos então detalhar um pouco mais estas definições. Quando se refere a *misinformation*, o autor trata de informação falsa que pode originar-se de fontes imprecisas, erros não intencionais, negligência ou viés inconsciente. Já pela outra definição (*disinformation*), refere-se a informação falsa que possui caráter intencional, sendo especificamente fabricada com o objetivo de causar desinformação. Por outro lado, uma informação falsa pode ainda ser definida como rumor, quando a notícia aparentemente verídica, ainda não pôde ser confirmada por fontes oficiais ou é de difícil confirmação.

A desinformação intencional é particularmente perigosa, pois, uma vez que não é causada por acidente, implica que alguém (entidade, pessoa ou corporação) está ativa e intencionalmente atuando para enganar pessoas. Portanto, para este trabalho definiremos *Fake News*, como:

> *Fake News*: notícias intencionalmente falsas, desenvolvidas com propósito de enganar e desinformar, mas que, de alguma maneira, podem ter sua autenticidade verificada. (3)

A partir desta definição as *Fake News* podem ainda ser divididas em três tipo distintos (14):

- a) Profissionalmente fabricadas - têm origens em reportagens jornalísticas fraudulentas e podem representar uma excelente fonte para a criação de um corpus de *Fake News*;
- b) Boatos - trata-se de uma outra forma de desinformação deliberada, erroneamente interpretada como notícia verdadeira ou originada em fonte não confiável;

- c) Notícias humorísticas - sátiras, paródias e programas de humor que utilizam notícias falsas com o objetivo de entretenimento.

A distinção entre os diferentes subtipos não são alvo da abordagem de detecção proposta neste trabalho.

2.1.2 Abordagem

Desenvolver estratégias para lidar com este tipo de ameaça tem sido foco de atenção em diversos trabalhos em computação. No que diz respeito à funcionalidade das abordagens automatizadas para o combate às *Fake News*, podemos dividi-las em dois grupos, Intervenção e Detecção (4):

- Intervenção: Podendo ser reativa ou proativa, atua nas redes sociais visando combater os efeitos das notícias falsas, quer seja mitigando seus efeitos após a sua detecção ou prevenindo sua proliferação.
- Detecção: Visa a implementação de um classificador que permita identificar de forma automatizada se uma notícia é *fake* ou *não fake*.

O foco deste trabalho está no esforço de **detecção**. Ferramentas de detecção podem ser usadas como subsídio para outras iniciativas que visam intervir na divulgação e propagação das *Fake News*. A automatização é um conceito importante nesta tarefa, pois como foi apresentado na introdução, a velocidade na detecção é um aspecto essencial para mitigar os efeitos negativos desta prática.

Podemos dividir em duas linhas as abordagens para a detecção automatizada de *Fake News* (15), conforme sintetizada na Figura 1 :

- Linguística - Na qual informações do texto são extraídas e analisadas para associar padrões da linguagem com a produção de uma notícia falsa.
- de Rede - Onde metadados das mensagens ou conhecimentos estruturados da rede de propagação, como características dos usuários, pontos de origem nas redes e a reação de outros usuários, podem ser incorporadas para fornecer informações adicionais sobre as notícias.

Podemos ainda definir uma abordagem híbrida, com a junção destas duas. Para este trabalho optou-se pela abordagem linguística, baseada no conteúdo textual da notícias, pois, por não estar restrita ao ambiente das redes sociais, pode ser mais abrangente em sua utilização.

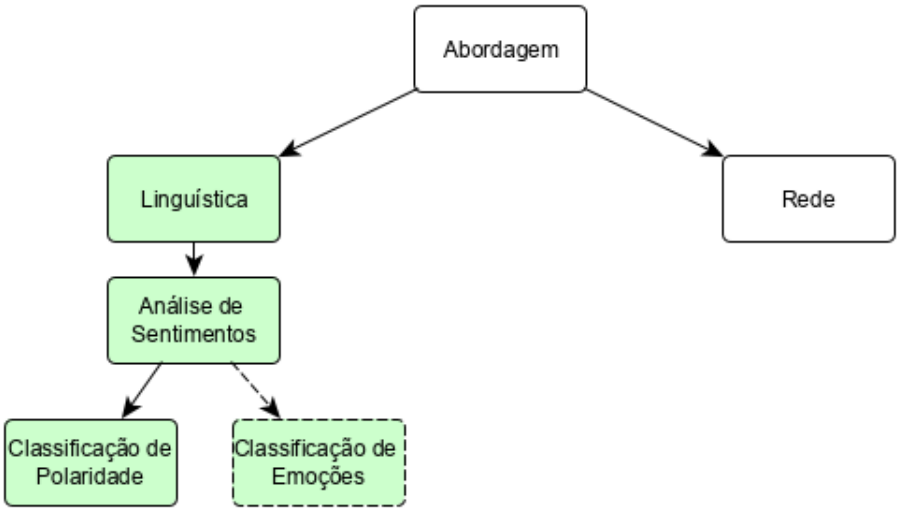


Figura 1 – Abordagens para Detecção Automatizada de *Fake News*. Adaptado de Bondielli e Marcelloni(1)

A criação deliberada de uma notícia falsa constitui uma dificuldade extra para a sua detecção por meio da análise do texto, pois divulgadores intencionais podem ser capazes de manipular o conteúdo para que as notícias $fake$ pareçam não $fake$ (16, 17). Entretanto, um número significativo de métodos computacionais para detecção de *Fake News* utilizam as características linguísticas (1, 7) com resultados significativos. Tal fato pode estar pautado na presença de características linguísticas marcantes, essenciais para que uma notícia *fake* alcance seu objetivo de atingir o maior número possível de pessoas, e tais características podem ser importantes atributos para a tarefa de detecção.

Produzir um texto falso, por princípio requer descrever eventos que não aconteceram. Além de ser convincente, a história deve parecer sincera. Como resultado, tais narrativas são qualitativamente diferentes das baseadas em fatos reais. Uma forma de capturar as diferenças entre histórias verdadeiras e falsas é analisar a linguagem utilizada. Diversas características do estilo linguístico utilizado, como o uso de alguns pronomes, palavras com conotações emotivas, preposições e conjunções que sinalizam esforço cognitivo, têm sido associadas a uma série de respostas emotivas e comportamentais (12).

Por exemplo, algumas das construções linguísticas que podem evidenciar textos que não expressam fatos reais são: poucas auto-referências; maior quantidade de palavras com emoções negativas; e baixa complexidade cognitiva. Primeiramente, o uso da primeira pessoa do singular pode ser considerada uma sutil proclamação de autoria do que está sendo divulgado. Em artigo de 1974 publicado na *American Psychological Association* (18) os autores levantaram a hipótese de que pessoas ao produzirem um texto falso podem evitar o uso de expressões que determinem a propriedade do texto por dois motivos: para se desassociarem de suas palavras ou pela ausência da experiência pessoal sobre o fato. Por este motivo, comunicações enganosas deveriam se caracterizar por poucos

pronomes pessoais em primeira pessoa do singular. Em segundo lugar, estudos sobre pequenas notícias do cotidiano sugerem que pessoas sentem-se desconfortáveis e culpadas no momento do ato de mentir e imediatamente após. Se este estado de espirito se reflete no uso da linguagem, então comunicações enganosas podem ser caracterizadas pelo maior uso de palavras que reflitam emoções negativas (19). Por último, o processo de criar uma história falsa consumiria recursos cognitivos, levando seus criadores a produzirem textos menos complexos (20), com o uso de estruturas mais simples e concretas ao invés de emitir avaliações e julgamentos.

2.2 Processamento de Linguagem Natural

2.2.1 Aspectos Gerais

Para identificação das construções linguísticas (*Stylometry*) (11) que podem auxiliar na detecção de *Fake News*, se faz necessário o processamento automatizado da linguagem humana (21). O campo da ciência que tem por objetivo o desenvolvimento de técnicas que permitem que computadores interpretem a linguagem humana tem o nome de Processamento de Linguagem Natural (PLN) (22), também denominado Linguística Computacional ou, ainda, Processamento de Línguas Naturais [1]. A análise linguística no PLN pode ser dividida nos seguintes segmentos: morfológico, sintático, semântico e pragmático.

Na analise morfológica são identificadas palavras ou expressões isoladas, definidas por delimitadores como pontuação ou espaços em branco. A morfologia lida especificamente com a constituição das palavras e dos grupos de palavras que formam os elementos de expressão de uma língua. Por exemplo, a palavra *'cárcere'* não pode ser dividida em partes menores, já as palavras *'cárceres'* e *'carcererreiro'* podem. As partes constituintes das palavras, denominadas morfemas podem ser independentes, como em cárcere ou dependentes como no caso dos sufixos (*s* em *cárceres*) e prefixos (*im* em *impossível*) (23). Além da estrutura das palavras, a análise morfológica fornece a sua classificação de acordo com seu tipo de uso ou sua categoria gramatical, como substantivos (p.ex.: árvore), adjetivos (p.ex.: curioso), verbos (p.ex: fazer), entre outros (24).

No nível sintático, em que muitas das abordagens existentes para processamento de linguagem são fundamentadas, os métodos que se baseiam principalmente na frequência da ocorrência de palavras (25). O analisador sintático trabalha em nível de agrupamento de palavras, analisando a constituição das frases, verificando se a estrutura é válida e identificando os componentes dessa estrutura. As estruturas das frases compartilham algumas propriedades que permitem que pessoas possam entender e reproduzir sentenças que não foram ouvidas (ou lidas) antes. Também a partir da estrutura da frase é possível

[1] https://www.sbc.org.br/14-comissoes/394-processamento-de-linguagem-natural

identificar relacionamentos entre os elementos do texto, como por exemplo sobre qual entidade se refere o texto (sujeito) ou qual ação está sendo afirmada (predicado). Essa estrutura pode ser caracterizada por uma gramática, consistindo de um conjunto finito de regras e princípios (23).

A representação semântica foca no significado associado às expressões da linguagem natural. Ao invés de simplesmente processar o texto a nível sintático, os métodos baseados em semântica se baseiam em características inerentes ao texto, desta forma indo além do significado isolado de palavras e expressões. O significado, pode ser uma proposição sobre os fatos narrados ou, ainda, pode expressar o propósito ou a intenção do autor. Quando o estudo está centrado no significado das palavras, recebe o nome de semântica lexical, quando o foco está no significado de uma proposição, trata-se da semântica lógica. Um dos grandes desafios da semântica lexical está no tratamento da ambiguidade que algumas palavras apresentam. Como por exemplo o verbo '*ser*', que apresenta dezoito diferentes definições pelo Dicionário Aurélio da Língua Portuguesa. Já a semântica lógica considera o significado a partir da definição de um domínio de conhecimento, similar a teoria dos conjuntos. No âmbito da PLN, o processamento semântico é considerado um dos maiores desafios, pois apresenta questões que são difíceis de tratar de maneira exata e completa. O Significado semântico está associado a um conhecimento da realidade no qual o assunto está inserido e também a questões mais obscuras como estados mentais e consciência. Se de um lado depende da morfologia e da estrutura sintática, de outro lado pode estar vinculado com informações da representação pragmática.

Na representação pragmática, o entendimento da narrativa é a área central para sistemas de cognição automatizada e tomada de decisão. Além de ser parte fundamental da comunicação humana, as narrativas são a forma como a qual as realidades são construídas. Decodificar como as narrativas são geradas e processadas pelo cérebro humano, pode levar ao melhor entendimento da inteligência da consciência humana. É função da análise pragmática fazer uma interpretação do todo e não mais analisar o significado das partes de uma linguagem, considerando outros fatores, como contexto e os interlocutores.

Cabe ressaltar que os princípios explorados neste trabalho estão mais fortemente relacionados as duas primeiras divisões de representação linguística, a morfológica e a sintática.

Na literatura não específica sobre PLN, a linguagem natural é referida como dado não estruturado, enquanto dados estruturados são aqueles presentes em bancos de dados relacionais. Uma abordagem estruturada em PLN significa converter texto não estruturado em dados estruturados de forma a permitir que ferramentas tradicionais de tratamento e preparação de dados sejam aplicadas, objetivando que os resultados possam ser visualizados e para que se possa obter conhecimento através de análises quantitativas e qualitativas desse dado. Esta premissa norteia os esforços detalhados nos próximos tópicos.

2.2.2 Tokenização

Dados uma sequência de caracteres e definida a unidade que representa um documento, *tokenizar* (do inglês *tokenization*), é a tarefa de dividir este texto em pedaços, eventualmente eliminando alguns caracteres, como símbolos de pontuação (26).

Tokens são frequentemente chamados de termos ou palavras, mas algumas vezes é importante fazer uma distinção. Um *token* é uma instância de uma sequência de caracteres em algum documento específico que são agrupados como uma unidade semântica definida para processamento. Podem ser palavras, como podem ser expressões ou caracteres isolados. Um tipo é a classe definida para os *tokens* que contêm a mesma sequência de caracteres. Um termo é um tipo que foi incluído no dicionário de um sistema recuperação de informações. O conjunto de termos pode ser totalmente distinto dos *tokens*, mas eles geralmente são derivados deles por vários processos de normalização e na prática eles estão fortemente relacionados aos *tokens* no documento. A principal questão da fase de *tokenização* é definir quais são os *tokens* corretos a serem usados. O exemplo da Tabela 1 é trivial: corta-se os espaços em branco e descarta-se os caracteres de pontuação.

Entrada:	Isto é um sonho, bem sei, mas quero continuar a sonhar.[2]
Saida:	Isto | é | um | sonho | bem | sei | mas | quero | continuar | a | sonhar

Tabela 1 – Tokenização

2.2.3 Remoção de Termos Comuns

Às vezes são totalmente excluídas do vocabulário algumas palavras extremamente comuns, que são de pouco valor para selecionar documentos que correspondam as necessidades da aplicação. Essas palavras são chamadas palavras de parada (do inglês *stop words*). A estratégia geral para determinar uma lista de *stop words* é classificar os termos por frequência de coleta (o número total de vezes que cada termo aparece na coleção de documentos) e, em seguida, selecionar os termos mais frequentes. As *stop words* podem também ser filtradas por seu conteúdo semântico relativo ao domínio dos documentos sendo analisados (e.g. uma lista de *stop words*). Um fragmento de lista de *stop words* é mostrado como exemplo na Figura 2.

```
de a o que e do da em um para com uma os no
se na por mais as dos como mas ao das tem à
ou ser quando muito há nos já também só pelo
```

Figura 2 – Amostragem parcial uma lista de stop words

Usar uma lista de *stop words* reduz significativamente o número de elementos que um sistema precisa processar, sem afetar a qualidade da tarefa sendo executada e é uma atividade frequente no pré-processamento de textos.

2.2.4 Classificação Gramatical

É atribuído a Dionísio Thrax de Alexandria (c. 100 a.C.) um esboço gramatical do grego que resumia o conhecimento linguístico da época e que é a fonte de grande parte do vocabulário linguístico moderno, incluindo palavras como sintaxe, ditongo e a analogia com partes do discurso. Seu trabalho inclui também uma descrição de oito classes gramaticais: substantivo, verbo, pronome, preposição, advérbio, conjunção, particípio e artigo. Embora estudiosos anteriores (incluindo Aristóteles e os estoicos) tivessem suas próprias listas de partes de discurso, foi o conjunto de oito classes de Thrax que se tornou a base para praticamente todas as descrições de classes gramaticais da maioria das línguas europeias nos 2.000 anos seguintes (27).

A classificação ou etiquetagem de partes do discurso [3] (28) (do inglês Parts of Speech ou POS Tagging) é uma técnica extremamente útil pois revela muito sobre a palavra avaliada e as palavras vizinhas e tem recebido grande atenção como um componente crítico de muitos sistemas de processamento de linguagem natural (29). Saber se uma palavra é um substantivo ou um verbo pode dar indicações sobre prováveis palavras vizinhas, tornando esta técnica um aspecto chave para, por exemplo, sistemas de análise textual, em operações de identificação de entidades como pessoas e organizações em sistemas de extração de informações, ferramentas de busca ou ainda em reconhecimento e síntese de voz. A *POS Tagging*, referida aqui como classificação gramatical, é utilizada com o objetivo de obter medidas de sua utilização (30).

Os etiquetadores automáticos existentes podem ser baseados em duas abordagens principais: baseada em regras, inicialmente, caracterizada pelo uso de regras codificadas manualmente (31, 32) e, posteriormente, pela aplicação de abordagens semi-automatizadas (33); e probabilística, que aplica métodos de etiquetagem estatísticos como Hidden Markov Model (HMM) (34), HMM e Árvore de Decisão (35) , Maximum Entropy (36), Memory-based (37), dentre outros.

Independente da abordagem utilizada, os algoritmos para a classificação das palavras, em sua maioria, utilizam as categorias gramaticais tipicamente presentes na *Universal POS tags* [4], tais como adjetivos (ADJ), verbos (VERB), pronomes (PRON), etc. E, embora não sejam consideradas classes gramaticais, estão incluídas nessa classificação elementos de pontuação (PUNCT), alguns caracteres não alfanuméricos (SYM) e uma classe especial (X) para palavras que, por alguma razão, não possam ser classificadas nas demais categorias.

[3] Também referidas como: classes de palavras, categorias sintáticas ou ainda rotulação morfossintática.
[4] https://universaldependencies.org/u/pos/

Tipicamente, dado um texto, para cada palavra é devolvida uma "etiqueta" com informação a respeito de sua categoria gramatical.

Uma outra forma de classificação pode ser realizada por meio de léxicos, que relacionam as palavras à sua classe por meio de uma busca (38). A classificação por meio de léxicos é realizada de forma fixa, não considerando o contexto onde a palavra está inserida, podendo portanto levar a classificações ligeiramente diferentes das obtidas pelos algoritmos de etiquetagem gramatical. Exemplos de classes presentes em léxicos são: *pronoun, ppron, ipron, article, prep, auxverb, adverb, conj, negate, verb, adj, etc.*. A marcação gramatical pode então ser representada por um classificador $t = \alpha(p)$, onde dentro de um contexto, cada palavra p estará associada a sua classe t correspondente.

2.2.5 Análise de Sentimentos

A Análise de Sentimentos é o campo de estudo onde são usados métodos, técnicas e ferramentas para detecção e extração de informações subjetivas como opiniões, atitudes, sentimentos, avaliações ou emoções a partir da linguagem (39). Tradicionalmente os trabalhos que abordam a análise de sentimentos, tratam de polaridade de opiniões. Por exemplo, se uma pessoa ou grupo de pessoas tem opiniões positivas, negativas ou neutras sobre determinado produto ou serviço. Isto pode explicar o porquê dos termos Análise de Sentimento e Mineração de Opiniões são frequentemente utilizados como sinônimos (40).

Embora os conceitos de classificação de emoções e polaridade de opinião não sejam equivalentes, eles possuem grande intercessão. Esta análise divide classificação de polaridade de classificação de emoções, sendo ambas brevemente descritas a seguir.

2.2.5.1 Classificação de Polaridade

A classificação de polaridade indica se o texto expressa opiniões positivas, negativas ou neutras em relação a um assunto, também chamada de mineração de opiniões. Trata-se de um problema de classificação de texto, que pode ser dividido em dois subtópicos: classificar um documento contendo texto opinativo, determinando se este expressa uma opinião sobre determinado assunto; ou pode-se chegar a nível de sentença e classificá-la como objetiva (contendo fatos) ou subjetiva (se expressa uma opinião positiva, negativa ou neutra). O primeiro é comumente conhecido como classificação de sentimentos a nível de documento, e se propõe a encontrar o sentimento genérico do autor em um texto opinativo, por exemplo na avaliação de um produto ou em uma notícia publicada na mídia. No segundo caso, chamada de classificação de subjetividade ou classificação de sentimentos a nível de sentença. Existe um grande número de expressões que descrevem opiniões positivas ou negativas, é papel da mineração de opiniões inferir essa opinião, baseado na linguagem usada para se expressá-las (41).

A classificação de polaridade pode se utilizar de léxicos para realizar a tarefa de classificação (42), onde cada termo ou expressão está associado a um valor inteiro entre $[-1, 1]$. Podemos então formalizar a função $P(p)$ que determina a polaridade de uma palavra p qualquer, onde teremos:

$$P(p) = \begin{cases} 1 \rightarrow positivo \\ 0 \rightarrow neutro \\ -1 \rightarrow negativo \end{cases} \tag{2.1}$$

Sendo o neutro adotado para descrever os casos em que o sentimento associado a uma determinada expressão não é claramente positivo ou negativo.

2.2.5.2 Classificação de Emoções

As definições e conceitos apresentados nesta sub-seção foram extraídos dos trabalhos dos seguintes autores: Medhat, Hassan e Korashy(10), Buechel e Hahn(43) e Mohammad(44)

A Classificação de Emoções expande o conceito de Analise de Sentimento e tem por objetivo identificar estados afetivos denominados emoções que são projeções de um sentimento. Emoções são nossos sentimentos e pensamentos subjetivos e têm sido estudadas em vários campos de pesquisa, como psicologia, filosofia, sociologia, biologia, entre outros. Não existe um conjunto acordado de emoções básicas entre os pesquisadores.

Entretanto, existem Modelos Discretos de representação das emoções (no inglês Discrete Emotion Models-DEM), que envolvem colocar as emoções em classes ou categorias distintas (45). Entre os mais proeminentes está o modelo de Paul Ekman (1992) (46) que propôs um conjunto de seis tipos de emoções básicas: alegria, tristeza, raiva, medo, nojo e surpresa. Tais emoções se originam, segundo Ekman, em diferentes sistemas neurais como resultado da percepção de experiências vividas, desta forma as emoções seriam independentes entre si. Entretanto, a sinergia entre elas podeira produzir outras emoções mais complexas como culpa, vergonha, orgulho, luxúria, ganância, entre outras. Outro modelo é o de Robert Plutchik (1991) (47), que apresenta um conjunto semelhante ao de Ekman, incluindo contudo, confiança e antecipação. No modelo de Plutchik as emoções podem ainda ser subdivididas em várias emoções secundárias e terciárias, cada uma pode ainda ter diferentes intensidades. A Figura 3 mostra como Plutchik arrumou estas emoções de tal forma que emoções divergentes aparecem diametralmente opostas e as que se encontram mais perto do centro sendo derivadas das oito emoções fundamentais (círculo seguinte) e tendo maior intensidade que aquelas nas extremidades, que são combinações entre elas. Os modelos discretos têm sido empregados em diversos trabalhos para a tarefa de classificação da emoção devido a sua simplicidade, embora não cubram todas as classes de emoções.

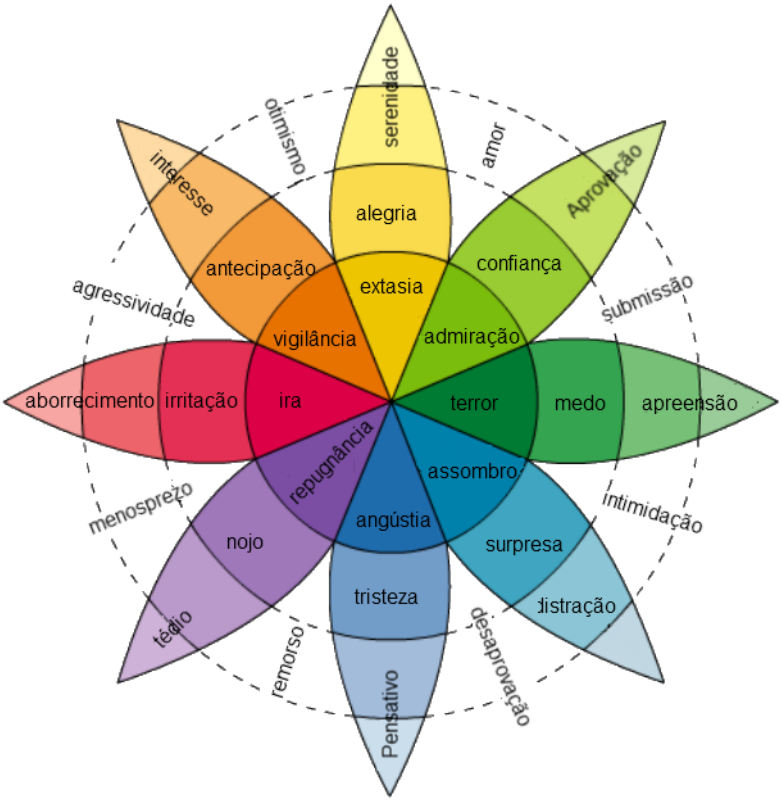


Figura 3 – Roda das Emoções proposta por Plutchik

Uma segunda forma de classificar os modelos de emoção, é levando-se em consideração a sua dimensionalidade. Modelos Dimensionais de Emoções (ou Dimmensional Emotion Models - DiEM), pressupõem que as emoções não são independentes, que existe uma relação entre elas, necessitando desta forma que eles sejam construídos em espaços multidimensionais que demostrem como elas se relacionam e refletindo os dois principais comportamentos fundamentais do bom o do mau.

O modelo bi-dimensional de Russel (2), apresentado na Figura 4 em forma de círculo, distingue as emoções entre valência (*Valence*) e excitação (*Arrousal*), onde a valência distingue entre emoções agradáveis ou desagradáveis e a excitação diferencia as emoções entre alta-ativação (alta resposta neurológica) e baixa ativação ou desativação (baixa ou ausente resposta neurológica). O modelo estabelece que as emoções não são independentes entre si.

O modelo de Plutchik da Figura 3, já comentado anteriormente, também se enquadra no formato dimensional e mostra a valência no eixo vertical e a excitação no eixo horizontal.

Quando se discute os sentimentos subjacentes de opiniões ou emoções, vale distinguir duas diferentes definições: o estado mental da pessoa (ou sentimentos) e a linguagem que ela usa para expressá-lo. Embora haja um número limitado de emoções, existe um grande número de expressões que podem ser utilizadas para expressá-las. Classificação de

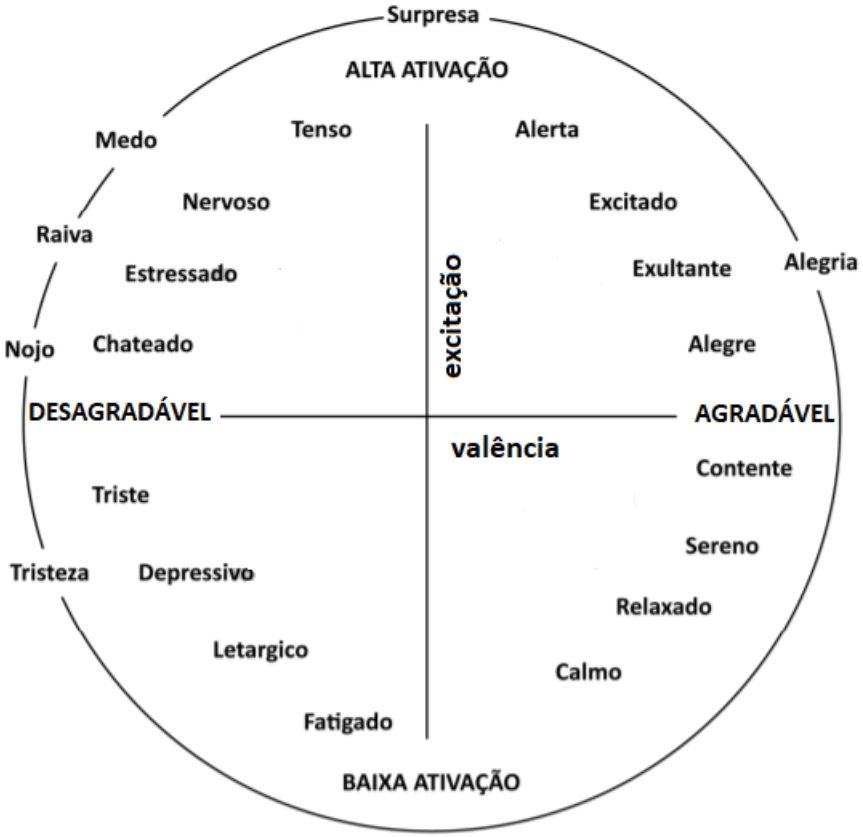


Figura 4 – Modelo circumplexo de Emoções (2)

emoções refere-se à tarefa de identificar as palavras relacionadas as dimensões subjetivas das emoções humanas dentro do texto analisado.

A Figura 5 mostra um exemplo em árvore de classificação de emoções.

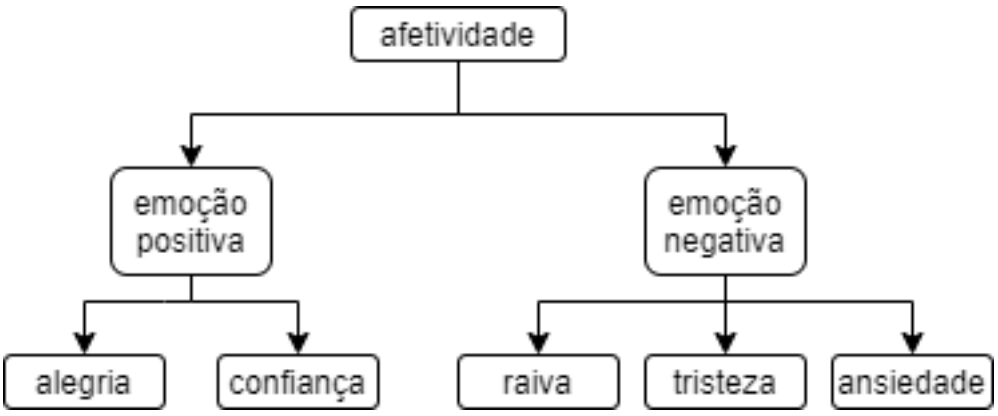


Figura 5 – Exemplo de Estrutura do léxico afetivo LIWC 2007

Neste exemplo, as palavras com um significado afetivo ou relativas a uma emoção são associadas a classe "afetividade", estas são então classificadas seguindo esta hierarquia em forma de árvore. A classificação ocorre com a associação de cada palavra às categorias dos nós da árvore por onde ela se enquadra. Existem sobreposições entre as palavras de "emoção positiva" com palavras de polaridade positiva, e das palavras de "emoção negativa"

com as palavras de polaridade negativa descritas no tópico anterior, mas não são totalmente coincidentes. A título de exemplificação, a palavra "não" possui polaridade negativa em um determinado léxico de polaridade, mas pode não estar presente em uma classe afetiva de um léxico de emoções, ao passo que a palavra "aborrecido", no mesmo léxico de polaridade, tem polaridade negativa e também é classificada como emoção negativa no referido léxico de emoções. De maneira que toda palavra pode ter polaridade "positiva", "negativa" ou "neutra", mas nem todas representam uma emoção. Cabe ressaltar que existem inúmeros modelos de representação de emoções (48), este é apenas um deles.

2.2.6 Léxicos

Um tipo de estrutura utilizada em grande número de sistemas de PLN são os léxicos, também chamados de dicionários. Constituem-se, em sua maioria, de uma lista de itens ou mais formalmente itens lexicais e as informações relacionadas a estes itens. Podem ser palavras isoladas como *'lua', 'mel', 'casa', 'modo'* ou composições de palavras que podem definir um significado específico diferente do inicial, por exemplo, *'lua de mel', 'casa de cultura'* ou *'a grosso modo'*. Os léxicos podem apresentar informações associadas a categoria gramatical, gênero, número, grau, pessoa, tempo, modo, regência verbal ou nominal, representações semânticas, polaridade, emoções e etc. Os léxicos podem ser classificados em estrutura de formas ou estrutura de bases. Se for constituído por todos os itens lexicais, como palavras ou expressões, trata-se de um dicionário de formas. Se, por outro lado, sua estrutura for formada por morfemas [5], trata-se de um léxico de bases (23).

2.2.6.1 SentiLex-PT

A análise automática de sentimentos ou polaridade de opiniões, como descrito na Seção 2.2.5.1, dedica-se ao tratamento computacional de opiniões, sentimentos e atitudes, expressos em textos. As aplicações que realizam este tipo de análise baseiam-se, geralmente, em léxicos de sentimento. Em geral, a informação de sentimento descrita nestes léxicos corresponde à polaridade das palavras ou expressões comumente expressas por valores como: negativo, positivo ou neutro.

O SentiLex-PT (42) é um léxico de sentimento especificamente concebido para a análise de sentimento e opinião sobre entidades humanas em textos redigidos em português, sendo atualmente constituído por 7.014 lemas [6] e 82.347 formas flexionadas. A Tabela 2, apresenta trecho do Sentile-PT.

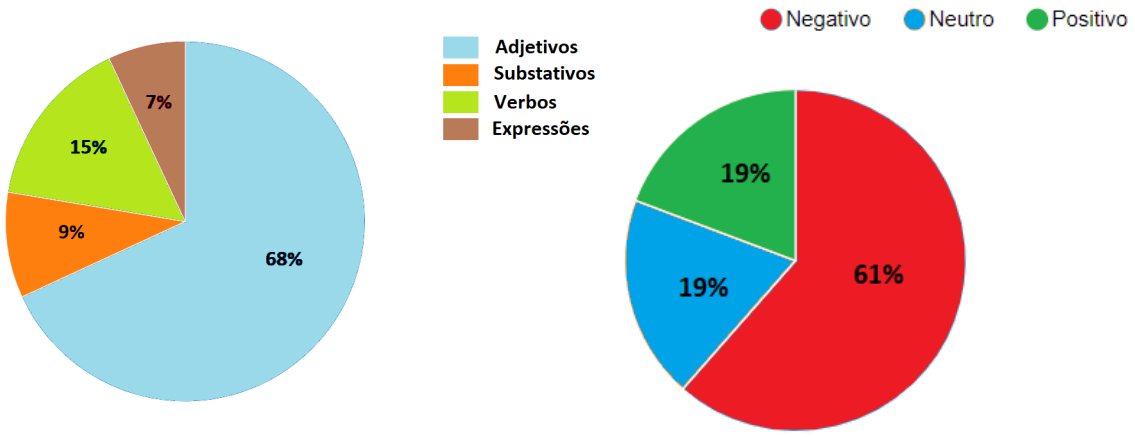
O SentiLex-PT tem dois dicionários associados, um que descreve lemas e o correspondente de formas flexionadas. No dicionário de lemas, cada linha inclui informação

[5] A menor unidade linguística que possui significado, abarcando raízes e afixos, formas livres, formas presas e vocábulos gramaticais - Oxford Languages and Google

[6] Lema: forma masculina do singular para os adjetivos, a forma singular para os nomes que flexionam em número e a forma infinitiva para os verbos e expressões idiomáticas.

inveja	PoS=N;TG=HUM:N0;POL:N0=-1;ANOT=MAN
invejado	PoS=Adj;TG=HUM:N0;POL:N0=0;ANOT=JALC
invejar	PoS=V;TG=HUM:N0:N1;POL:N0=-1;POL:N1=0;ANOT=MAN

Tabela 2 – Trecho de léxico em estrutura de formas



(a) Classes

(b) Polaridade

Figura 6 – Distribuições no Sentilex-PT

sobre:

- Categoria gramatical (POS): Adjetivo **(ADJ)**, Substantivo **(N)**, Verbo **(V)** ou Expressões **(IDIOM)**.
- Polaridade (POL): positiva (1), negativa (-1) ou neutra (0)
- Alvo da polaridade (TG): corresponde a um nome de tipo humano (HUM), com função de sujeito (N0) e/ou complemento (N1)
- Método de anotação da polaridade (ANOT): manual (MAN) ou atribuída por ferramenta automatizada (JALC)

No dicionário de formas flexionadas, as entradas estão associadas ao respetivo lema. Neste formato, além das informações descritas no dicionário de lemas, os adjetivos e nomes contêm informação sobre a flexão (FLEX) em gênero (masculino (m) ou feminino (f)) e número (singular (s) ou plural (p)). Os atributos associados aos verbos e expressões idiomáticas incluem informação de tempo verbal, pessoa e número.

2.2.6.2 LIWC

Palavras usadas no cotidiano refletem personalidades e relações sociais em que estão inseridos os personagens que as utilizam. Palavras são as formas mais comuns e confiáveis das pessoas externalizarem pensamentos e emoções, em um formato que outros possam

entender. O uso das palavras e a linguagem são os principais meios pelos quais psicólogos e analistas comportamentais tentam entender o ser humano. O LIWC (*Linguistic Inquiry Word Counting*)[7] é um software de análise de textos com um conjunto de dicionários em diversos idiomas, que conta as palavras em categorias psicologicamente significativas. Resultados práticos do uso do LIWC mostraram sua habilidade em detectar significado em uma variedade de situações, como identificar foco e atenção, relações sociais, estilos de pensamento, individualidades e, sob o ponto de vista de interesse deste trabalho, a emotividade (11).

O LIWC tem dois componentes principais, o software que processa o texto analisado comparando palavra por palavra e um conjunto de dicionários em diversos idiomas, onde as mesmas são classificadas em categorias. Os dicionários são o coração do LIWC, que quando foi criado tinha o propósito de identificar apenas a percentagem de emoções positivas ou negativas no texto, mas rapidamente evoluiu para mais de 80 categorias. A versão 2015, a mais recente disponível no momento da escrita deste trabalho, continha 73 categorias no total subdivididas em dois grandes grupos: um grupo das palavras que representam conteúdos, normalmente substantivos, verbos, adjetivos e advérbios; um outro grupo de palavras de estilo ou funcionais, compostas de pronomes, preposições, artigos, conjunções, verbos auxiliares e algumas outras categorias. Palavras funcionais estão mais diretamente relacionadas ao ambiente psicológico e social das pessoas e pesquisas sugerem que o LIWC identifica com precisão a emoção usada na linguagem.

A estrutura do dicionário do LIWC está dividida em duas regiões distintas. A primeira contém as classes, representadas por um índice numérico e um texto identificador. A Tabela 3 apresenta o subconjunto das classes do LIWC relevantes para este estudo. Na segunda parte o dicionário apresenta 14458 palavras ou grupo de palavras resumidas por caractere coringa (asterisco), que permite aceitar demais letras, hifens, ou números e suas respectivas classificações segundo o dicionário, exemplificado na Tabela 4.

Para aplicação neste trabalho, as classes do dicionário do LIWC foram divididas em três grupos a saber: de 1 a 23 definidas como classes gramaticais; as identificadas entre os índices 30 até o 35 relativas as emoções; já as classes do 40 ao 125, são relacionadas a comportamentos sociais e não foram utilizadas nos experimentos.

2.2.6.3 Affect-BR

Embora haja uma versão para o português do Brasil do dicionário do LIWC, problemas identificados na sua versão de 2007 (LIWC2007pt) levaram ao desenvolvimento do Affect-BR (49), um léxico emotivo específico para o português brasileiro, baseado na versão em inglês do LIWC 2015. Uma vez que a tradução direta do dicionário poderia trazer

[7] http://liwc.wpengine.com/

1 funct	14 conj
2 pronoun	15 negate
3 ppron	20 verb
4 i	21 adj
5 we	22 compare
6 you	23 interrog
7 shehe	24 number
8 they	25 quant
9 ipron	**30 affect**
10 article	**31 posemo**
11 prep	**32 negemo**
12 auxverb	**33 anx**
13 adverb	**34 anger**
	35 sad

Tabela 3 – Lista parcial das Classes do LIWC

abomina*	70 71
abençoa	20 30 31 91 114
bonita*	21 30 31 60 61
bons	21 30 31 80 84
choro*	30 32 35
limitad*	100 102
mandastes	20 50 52 90
mandava	20 50 52 90
obcecad*	30 32 33
oprimi*	30 32 33
otári*	30 32 34 120 121

Tabela 4 – Amostra do Vocabulário classificado pelo LIWC

problemas de mudança de significado, mais evidentemente para línguas mais complexas que o inglês, o trabalho de desenvolver um dicionário em outra língua é uma tarefa complexa.

O Affect-BR concentra seu vocabulário nas classes afetivas (*affect*) e em suas subcategorias. As palavras foram traduzidas de forma automática usando os tradutores do Google [8] e do Bing [9], posteriormente um dicionário foi utilizado para melhorar o entendimento dos significados das palavras, eliminação de redundâncias e inclusão de sinônimos e outras palavras que poderiam fazer parte do léxico. Como resultado o Affect-BR conta com um total de 1139 palavras associadas a classe afetiva, sendo 479 positivas e 661 negativas. Assim como no LIWC o Affect-BR faz uso de caracteres coringa, para resumir conjugações, variações de gênero ou grau e possui a mesma estrutura de classificação. O Affect-BR foi avaliado nos experimentos e sua performance comparada ao do LIWC de forma a permitir avaliar variações nas construções destes léxicos.

2.3 Aprendizado de Máquina

Em 1959 Arthur Samuel (50) definiu o termo aprendizado de máquina como: "Campo de estudo que permite aos computadores a habilidade de aprender sem serem explicitamente programados". Algoritmos de Aprendizado de Máquina analisam conjuntos de dados em busca (indução) de hipóteses ou funções capazes de descrever as relações entre os dados.

Dentro do PLN, problemas de classificação automatizada de textos utilizando aprendizado de máquina, tem sido amplamente estudado nas últimas décadas. A maioria

[8] https://translate.google.com/
[9] https://www.bing.com/translator

dos sistemas de classificação de texto e categorização de documentos pode ser subdivididos nas quatro fases seguintes: extração de características, reduções de dimensionalidade, seleção de classificador e avaliação. Um dos passos mais significativos na categorização de documentos está em escolher o melhor algoritmo de classificação. Documentos podem ser classificados por três diferentes métodos de aprendizado de máquina: supervisionado, não supervisionado e semi-supervisionado.(45)

2.3.1 Aspectos Gerais

O aprendizado supervisionado é definido quando para cada par de variáveis de entrada (x) e saída (Y), um algoritmo é usado para aprender a função (f) que mapeia o relacionamento da entrada com a saída, calculando o erro para ajustar os parâmetros do algoritmo e minimizar este erro: $Y = f(x)$

O objetivo é aproximar ao máximo a função de mapeamento de forma que quando houver uma nova entrada (x) será possível predizer a variável de saída (Y) para este dado. É chamado de aprendizado supervisionado porque é o processo pelo qual o algoritmo aprende a partir de um conjunto de dados de treinamento. Pode ser pensado como um professor supervisionando o processo. O aprendizado pára quando o algoritmo atinge um nível aceitável de desempenho. As tarefas relacionadas ao aprendizado supervisionado, podem ser agrupadas em tarefas de regressão e de classificação. O processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descrito como um conjunto de características e os rótulos a eles associados, é chamado de análise preditiva. Quando a identificação de uma classe ao qual o evento está associado, está dentro de um conjunto finito de classes existentes (verdadeiro ou falso; ensolarado, nublado ou chuvoso; vírus ou não vírus; etc.), tipifica uma situação de análise preditiva de classificação ou predição categórica. Outra forma possível é quando o evento está associado a um número dentro de um conjunto contínuo de valores possíveis (valores de temperatura; ou número de visualizações de um anúncio), neste caso a tarefa de análise preditiva caracteriza-se como de regressão ou predição numérica.

No aprendizado não-supervisionado não há um *"professor"* para supervisionar o processo, ele busca por redundâncias e regularidades nos padrões de entrada. Caracteriza-se por ter apenas dados de entrada sem variáveis de saída associadas. As tarefas relacionadas ao aprendizado não-supervisionado, podem ser agrupadas em tarefas de agrupamento e de associação. Problemas de agrupamento é quando se pretende descobrir agrupamentos inerentes aos dados, como agrupamentos de clientes pelo comportamento de compra. Já na descoberta de regras de associação, acontece quando se pretende descobrir regras que descrevam o comportamento de conjuntos de dados, como pessoas que compram um determinado produto, tendem a comprar um outro produto. Tarefas de aprendizado não-supervisionado são de natureza descritiva.

O aprendizado semi-supervisionado inclui ambos os problemas discutidos anteriormente: ela usa dados rotulados e não-rotulados. É uma oportunidade para quando não se tem uma quantidade significativa de dados rotulados. Assim, são utilizadas técnicas de aprendizado não supervisionado, para aprender a estrutura dos dados não rotulados. Ao final dessa etapa não supervisionada, o algoritmo conseguirá associar cada observação com uma pontuação proporcional à probabilidade do dado ter vindo de uma das classes. Em seguida, usa-se o conjunto rotulado dos dados para definir um limiar na pontuação, a partir do qual pode-se classificar o conjunto.

2.3.2 Algoritmos de Classificação

A tarefa de classificação consiste em descobrir uma função que relacione um conjunto de registros às suas classes, permitindo que ao se aplicar esta função a um novo conjunto de registros seja possível determinar a classe de seus elementos. Diversas técnicas de aprendizado de máquina vem sendo propostas para classificação automatizada de documentos eletrônicos como os classificadores Bayesianos, árvores de decisão, K-Nearest Neighbor(KNN), Support Vector Machines(SVM), redes neurais, etc. Alguns destes algoritmos são descritos a seguir (51).

2.3.2.1 Classificador Bayesiano Ingênuo

O Classificador Bayesiano Ingênuo, ou *Naive Bayes (NB)* é um classificador probabilístico simples, baseado no teorema de Bayes com a suposição de forte independência entre os atributos. Essa presunção de independência, torna a ordem dos atributos irrelevante e consequentemente a presença de um atributo não afeta o cálculo da probabilidade dos outros. Estas características tornam este classificador muito eficiente computacionalmente, necessitando de volumes relativamente baixos de dados de treinamento, porém sua aplicabilidade pode ser limitada. Em um processo de classificação no qual um exemplar com rótulo desconhecido é apresentado ao classificador, o algoritmo *Naive Bayes* tomará a decisão sobre qual classe o exemplar $\vec{x_i} = \{x_{i1}, x_{i2}, ..., x_{in}\}$ deve estar associado, por meio de probabilidades condicionais, ou seja, as probabilidades dele pertencer a cada uma das classes ck (probabilidade da classe K, dado o exemplar $\vec{x}$) $P(ck|\vec{x})$ encontradas no conjunto de dados para treinamento, segundo a Formula 2.2:

$$P(ck|\vec{x_i}) = P(ck) \prod_{j=1}^{n} P(x_{ij}|ck) \tag{2.2}$$

Onde: $P(ck)$ é a probabilidade da classe ck no conjunto de treinamento e $P(x_{ij}|ck)$ é a probabilidade do atributo x_{ij} dada a distribuição da classe ck.

2.3.2.2 KNN

K-Nearest Neighbor (KNN) é um algoritmo usado para testar o grau de similaridade entre um documento e os “k” vizinhos mais próximos, sendo “k” um número inteiro, obtido de dados de treinamento para os quais mantém algum aspecto de similaridade ou métrica de proximidade. Métricas de distância são usadas para medir a similaridade ou dissimilaridade entre dois exemplares. Existem na literatura especializada várias formas de medir as distâncias entre dois exemplares com atributos numéricos, como: Manhattan, Euclidiana, Minkowski, Cosseno, entre outras. A fase de treinamento consiste em armazenar em vetores os atributos e classes do conjunto de treino. Na fase de classificação, são calculadas as distâncias de um novo vetor, contendo os atributos de um documento de entrada, para todos os vetores armazenados e os ”k” mais próximos são selecionados. A predição da classe do novo documento é representada pela classe mais frequente entre os ”k” registros. No KNN os vetores de atributos podem estar em um espaço n-dimensional, no qual cada atributo corresponde a uma dimensão, a Figura 7 mostra uma representação onde $n = 2$, ”A” e ”B” são classes do conjunto de treino e ”U1” e ”U2”, são documentos de entrada para serem classificados.

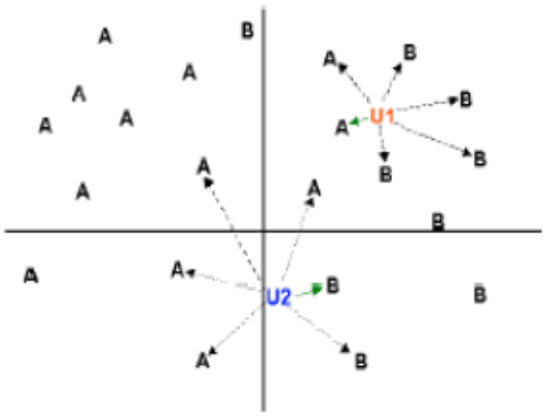


Figura 7 – KNN

2.3.2.3 SVM

Support vector machines (SVMs) é considerado um dos métodos de classificação discriminativo. O SVM define classificadores lineares, que utilizam um hiperplano[10] para separar o conjunto de dados em suas classes. Os dados que estiverem mais próximos da superfície de decisão, são chamados de Vetores de Suporte (Support Vectors). A performance do SVM não se altera se os documentos que não pertencem aos vetores de suporte não estiverem presentes no dados de treinamento.

[10] Hiperplano - figura geométrica de curvatura nula em um espaço euclidiano n-dimensional e cuja equação em coordenadas cartesianas é linear.

O algoritmo pode ser descrito da seguinte forma: dadas "D" amostras de treinamento x_i, y_i, com $i = 1, 2..., D$, onde $x_i \in \Re$ é uma representação vetorial de um conjunto e $yi \in \{-1, 1\}$ é sua classe associada. Neste processo existe uma distribuição de probabilidade $P(x, y)$ desconhecida da qual os dados de treinamento serão retirados. Ou seja, o processo de treinamento consiste em treinar um classificador de forma que este aprenda um mapeamento $x \mapsto y$, por meio de exemplos (classes) de treinamento $\{xi, yi\}$ de forma que a máquina seja capaz de classificar um exemplo (x, y) ainda não visto que siga a mesma distribuição de probabilidade (P) dos exemplos de treinamento.

A separação ótima entre classes ocorre por meio de um hiperplano condicional (C), tal que este plano é orientado para maximizar a margem (distância entre as bordas A e B) e pelo ponto mais próximo de cada classe. A Figura 8, mostra um espaço bi-dimensional, duas diferentes classes e o hiperplano de separação, que neste caso é uma linha.

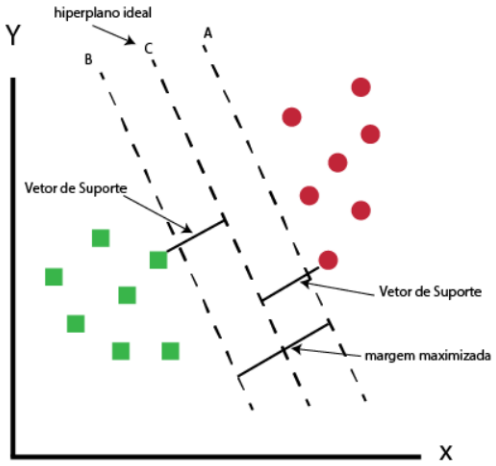


Figura 8 – Máquina de Vetor de Suporte

2.3.2.4 Árvore de Decisão

As árvores de decisão recriam o processo de categorização manual ao construir pontos de decisão bem definidos, em forma de árvore, onde as folhas representam as classes de categorias e os nós representam uma decisão sobre um atributo que determina como os dados são divididos. Modelos baseados em árvores de decisão podem ser construídos por especialistas ou podem ser parte de um algoritmo de aprendizado supervisionado que possibilitam inferir as regras de decisão da árvore.

2.3.2.5 Boosting

O *Boosting* é uma estratégia genérica para aprimorar o desempenho de qualquer algoritmo de aprendizado, que utiliza o princípio de que a combinação de classificadores fracos podem gerar classificadores fortes. Originalmente o método foi proposto para tratar problemas de classificação de padrões, com a introdução do AdaBoost, posteriormente

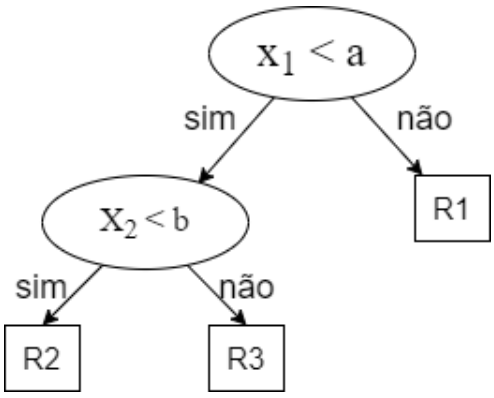


Figura 9 – Árvore de Decisão simplificada

diversas generalizações surgiram a partir da estratégia original, entre elas o algoritmo Gradient Boosting. Os métodos de *boosting* seguem um paradigma sequencial, utilizando uma estratégia que busca atuar sobre os erros da etapa anterior, com o objetivo de reduzir gradativamente os resíduos de previsão.

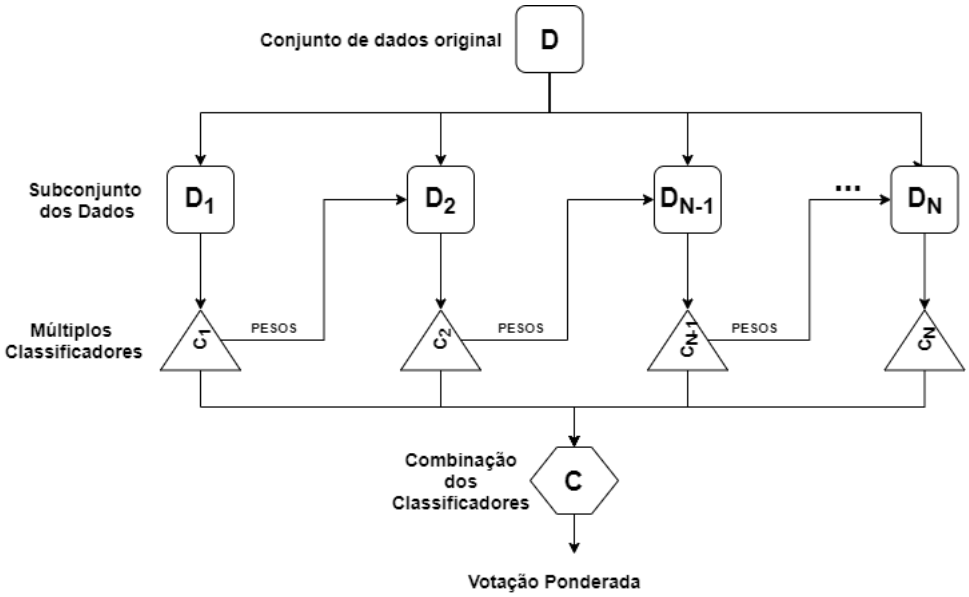


Figura 10 – Formação genérica de um algoritmo de *boosting*

2.3.2.5.1 Ada Boost

O AdaBoost (ou Adaptive Boosting) proposto pela primeira vez por Freund e Schapire(52) utiliza múltiplos classificadores para aumentar a taxa de acerto no processo de classificação e um algoritmo de aprendizado constrói um conjunto de classificadores base com o intuito de produzir um classificador melhor, no qual seus resultados são combinados através do voto ponderado. O algoritmo base é executado de forma repetitiva em várias iterações. Após a primeira iteração o conjunto de dados original é classificado, então é calculada a taxa de erro (ε), importância (α) e pesos (ω), de maneira que as instâncias

classificadas incorretamente têm seus pesos aumentados, e as classificadas corretamente têm seus diminuídos. O método para obtenção da hipótese final é a combinação ponderada das diversas saídas em cada iteração, quanto menor a taxa de erro, maior é a importância atribuída ao classificador.

2.3.2.5.2 Gradient Boosting

Embora possa ser utilizado para qualquer algoritmo de aprendizado, o *Gradient Boosting* (GB)(53) é mais utilizado com árvores de decisão. O poder preditivo individual de uma árvore de decisão (Seção 2.3.2.4) é normalmente inferior aos resultados que podem ser obtidos com outras técnicas de aprendizado computacional. Entretanto é possível aprimorar o poder preditivo de um modelo através da formação de um comitê de previsão.

Usando como exemplo algoritmos de regressão por mínimos quadrados $\frac{1}{n}\sum_i(\hat{y}_i - y_i)^2$, em que o índice i percorre algum conjunto de treinamento de tamanho n dos valores reais da variável de saída y, onde: $\hat{y}_i$ é o valor previsto de $F(x)$, y_i é o valor real e n é o número de amostras em y. Considerando um algoritmo de GB com uma quantidade M de etapas e em cada etapa $1 \leq m \leq M$ temos a função de predição fraca F_m. De maneira a melhorar F_m, uma função complementar de estimação é adicionada $h_m(x)$, de modo que $F_{m+1}(x) = F_m(x) + h_m(x) = y$. O GB ajusta h ao resíduo $y - F_m(x)$ tentando corrigir os erros de seu antecessor. Cada etapa ajustaria uma árvore de decisão $h_m(x)$ a pseudo-resíduos. Seja J_m a sua quantidade de folhas. A árvore divide o espaço de entrada em J_m regiões disjuntas $R_{1m}, \ldots, R_{J_m m}$ e prevê um valor constante em cada região. Usando a notação do indicador, a saída de $h_m(x)$ para a entrada x pode ser escrita como a soma $h_m(x) = \sum_{j=1}^{J_m} b_{jm}\mathbf{1}_{R_{jm}}(x)$. Onde J_m a sua quantidade de folhas, $R_{1m}, \ldots, R_{J_m m}$ regiões do espaço de entrada e b_{jm} é o valor previsto na região R_{jm}.

2.3.3 Medidas de desempenho

Para avaliar o desempenho dos algoritmos de classificação, algumas métricas de desempenho são revistas nesta seção. Considerando as abordagens para detecção de *Fake News* como um problema de classificação, onde se tenta descobrir se uma notícia pode ser falsa, podemos ter as seguintes situações:

- Verdadeiro Positivo (VP): quando a notícia foi prevista como falsa e foi anotada como falsa.
- Verdadeiro Negativo (VN): quando a notícia foi prevista como verdadeira e foi anotada como verdadeira.
- Falso Positivo (FP): quando a notícia foi prevista como falsa mas foi anotada como verdadeira.

- Falso Negativo (FN): quando a notícia foi prevista como verdadeira mas foi anotada como falsa.

As métricas a seguir são comumente utilizadas em aprendizado de máquina para avaliar o desempenho de classificadores por diferentes perspectivas:

$$Precisão = \frac{|VP|}{|VP|+|FP|} \tag{2.3}$$

$$Revocação = \frac{|VP|}{|VP|+|FN|} \tag{2.4}$$

$$F1 = 2.\frac{Precisão.Revocação}{Precisão + Revocação} \tag{2.5}$$

$$Acurácia = \frac{|VP|+|VN|}{|VP|+|VN|+|FP|+|FN|} \tag{2.6}$$

A Acurácia, por exemplo, avalia o nível geral de acertos entre as classes previstas com o conjunto total de notícias do conjunto. A Precisão mede a fração de acertos na identificação de notícias como *Fake News* entre todas aquelas que foram efetivamente identificadas como *Fake News*. A Revocação é usada para medir a sensibilidade ou a fração de todas as notícias efetivamente anotadas e que foram preditas como *Fake News*. O cálculo do F1 é usado para combinar a precisão e a revocação, que pode fornecer uma medida de desempenho geral na tarefa de predição de *Fake News*. Para todos estas medidas, quanto maior o valor obtido melhor é o desempenho.

3 TRABALHOS RELACIONADOS

Segundo Bondielli e Marcelloni(1), nos últimos anos, em especial na última década, notou-se um aumento substancial na quantidade de artigos de pesquisa relacionados aos estudos sobre rumores e *Fake News*. Em especial depois das eleições presidenciais americanas de 2016, a questão das *Fake News* recebeu especial atenção, promovendo grande interesse na comunidade científica com um aumento na quantidade de artigos sobre o tema. A Figura 11, mostra a análise realizada sobre a base de publicações Scopus [1], que comprova essa tendência.

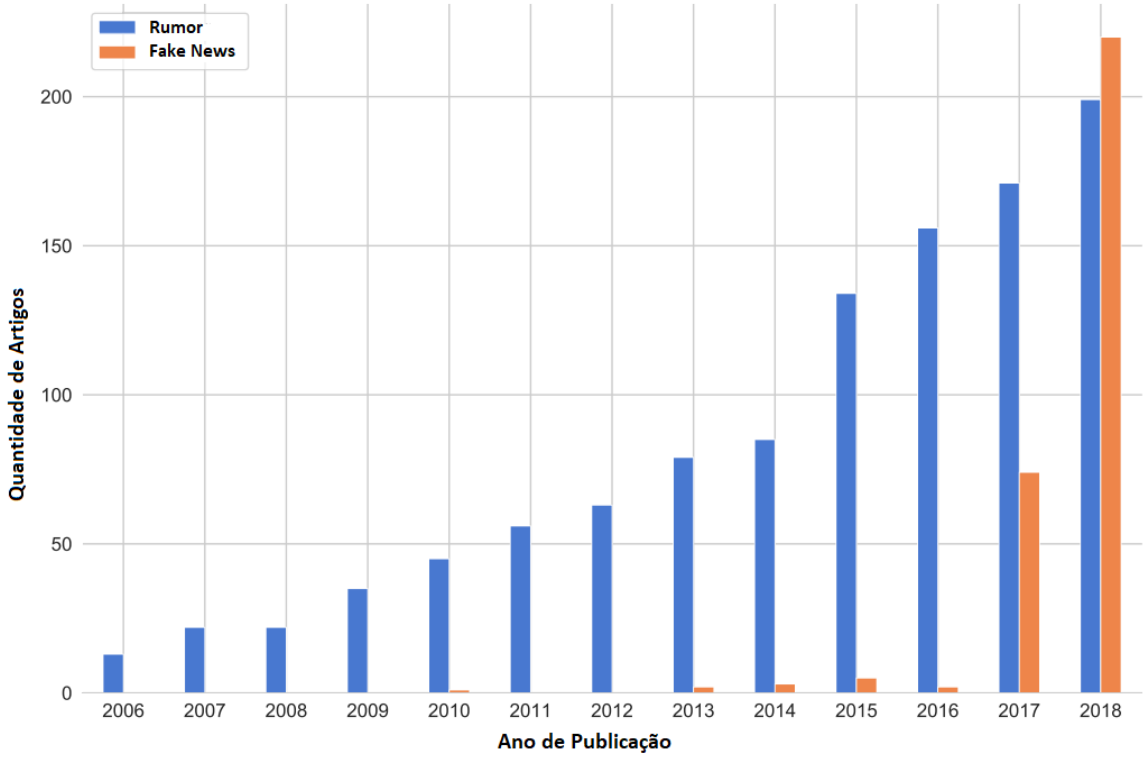


Figura 11 – Publicação de artigos sobre Fake News nos útimos anos (adaptado de Bondielli e Marcelloni(1))

Conforme descrito na seção de fundamentação teórica, este trabalho concentra seus esforços na detecção de *Fake News*, não sendo parte do seu contexto a as notícias caracterizadas como rumores.

Ainda em relação aos trabalhos desenvolvidos sobre as *Fake News*, no que diz respeito a língua, sendo o português um dos dez idiomas mais falados no mundo [2], ele representa um pequeno percentual dos trabalhos sobre *Fake News*. Uma busca pelo assunto

[1] https://www.scopus.com/
[2] https://pt.wikipedia.org/wiki/Lista_de_línguas_por_número_de_falantes_nativos

no *Google Scholar*, indicou que apenas cerca de 7% dos trabalhos (Figura 12) dedicam esforços sobre a língua lusófona.

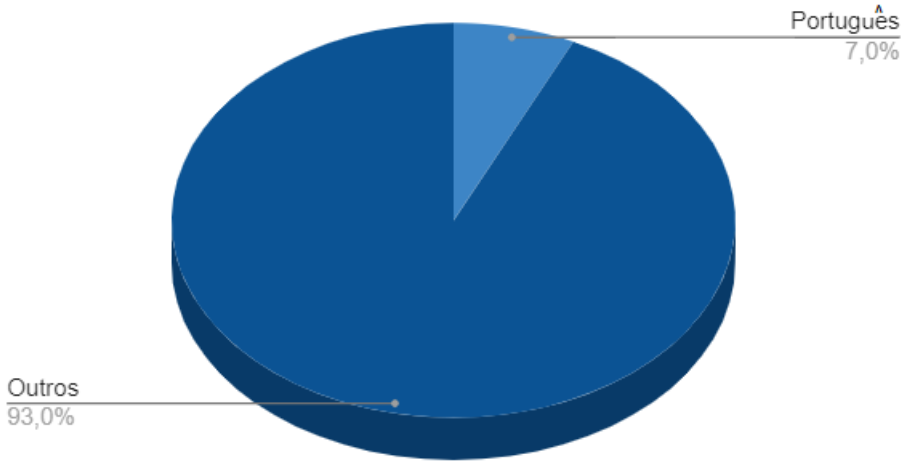


Figura 12 – Distribuição de artigos sobre Fake News de acordo com o idioma (fonte Google Scholar)

Diferentes aspectos tem sido analisados com a motivação de se obter informações de como a *Fake News* se prolifera nos MDDNs e como detectá-la de forma rápida e eficiente de forma a reduzir seus impactos nos usuários dessas mídias e na sociedade como um todo. Alguns dos aspectos mais explorados são, por exemplo, a possibilidade de inferir a veracidade de uma notícia baseado em informações inerentes ao seu conteúdo (e.g. texto, autor, etc.).

Alguns dos mais recentes trabalhos que se relacionam a este estudo utilizam métodos de detecção de *Fake News* com abordagens linguísticas (i.e., informações extraídas diretamente dos textos das notícias (7)), mais especificamente utilizam técnicas de classificação gramatical e uso da polaridade. Assim sendo, o restante desta seção apresenta resumidamente esses trabalhos, priorizando aqueles com foco no idioma português:

A revisão elaborada por Shu et al.(3), ressalta que a tarefa de detecção de *Fake News* no contexto das mídias tradicionais, é baseada no conteúdo das notícias, desta forma mostra-se de grande relevância a utilização de recursos que permitam extrair e representar características destes textos. Visto que *Fake News* são criadas intencionalmente para obter ganhos financeiros ou políticos ao invés de apresentar fatos objetivos, elas frequentemente contém linguagem opinativa e inflamada, de forma a incitar confusão. Portanto, seria razoável explorar as características linguísticas para capturar diferentes estilos de escrita e linguagem sensacionalista para sua detecção. Tais características podem ser extraídas, em termos de organização dos documentos, em diferentes níveis. A nível de caracteres e de palavras, algumas das características tipicamente levantadas podem ser: total de palavras, caracteres por palavra e frequência de palavras. A nível de sentenças e de documentos,

são destacados como relevantes, aspectos como utilização de elementos de pontuação, da marcação de partes do texto, bem como de outras características específicas (e.g.: citações).

O estudo propõe ainda métodos que buscam identificar estilos de escrita e construções que tipificam a intenção de manipulação da atenção do leitor. Segundo os autores, fabricantes de *Fake News* utilizam-se de particularidades na escrita de maneira a persuadir e convencer uma grande variedade de leitores. Estilos de escrita que normalmente não são encontrados em notícias verdadeiras. Estes métodos são divididos em duas categorias principais: orientado a falsidade no texto, quando métodos de estilometria são usados para capturar declarações falsas no texto, técnicas também encontradas em psicologia forense; ou orientado a (falta de) objetividade, abordagens que capturam sinais que indicam a falta ou uma diminuição de objetividade que podem enganar o consumidor da notícia.

Em um trabalho subsequente, Ajao, Bhowmik e Zargari(8) incorporam análise de sentimentos às características linguísticas na classificação de *Fake News*. Embora aponte que a análise de sentimentos não se limite à polaridade, introduzindo o conceito de identificação de emoções (como raiva, ansiedade, depressão e excitação), foca seu estudo na polaridade dessas emoções. Analisando textos do Twitter [3], utiliza recursos de modelagem de tópicos para selecionar os assuntos e palavras mais relevantes dentro do corpus analisado. Desta forma as palavras relevantes extraídas dos dez tópicos mais relacionados, são utilizadas no classificador de sentimentos. Uma vez que a classificação de sentimentos envolve a determinação da polaridade (positiva, negativa ou neutra) do texto, ao procurar pelas palavras chave encontradas nas publicações, eles foram capazes de identificar as melhores palavras para descrever emoções positivas ou negativas. A função de pontuação do LIWC foi utilizada como aplicação para determinar o sentimento dos textos analisados. Uma métrica foi então elaborada na razão entre a contagem das palavras com emoção negativa sobre a contagem das palavras com emoção positiva, definida pela equação 3.1:

$$razão\ de\ emoção = \frac{contagem\ de\ palavras\ de\ emoçoes\ negativas}{contagem\ de\ palavras\ de\ emoçoes\ positivas} \tag{3.1}$$

Essa razão, associada a extração de características dos textos, forma um vetor de atributos que é submetido aos algoritmos de classificação.

Bhutani et al.(54) levantam a hipótese de que a categorização de uma *Fake News* depende da atitude do autor para com o assunto, se é positiva, negativa ou neutra. Por exemplo, se um partido político faz uma declaração sobre um outro partido de oposição, e essa declaração é negativa, há uma indicação de que essa declaração possa ser falsa. Embora isso não seja determinante em todas as situações, certamente, afirmam os autores, é um indicativo importante para determinar se uma notícia é falsa. Desta forma, propõem

[3] Twitter é um serviço de rede social livre que permite que usuários registrados postem textos curtos chamados tweets.

um método que incorpora sentimentos, mais especificamente a polaridade do texto, como um atributo do *dataset* para detecção de *Fake News*. Diferentes técnicas de extração de informação do texto são aplicadas, como a vetorização do texto pela contagem de palavras de forma a determinar qual delas obtém melhor resultado utilizando um classificador Naive Bayes.

Morais et al.(55) levantam a questão de que certas mídias sociais utilizam linguagem irônica para produzir conteúdo humorístico. Usando uma linguagem jornalística em paródias que envolvem personagens reais em histórias cômicas, estes sites tem o objetivo de inspirar a crítica social. Entretanto, um relevante número de pessoas tende a confundir estes textos humorísticos com histórias reais. Para endereçar estes casos os autores propõem uma abordagem diferente ao sugerir uma classificação múltipla para as notícias, podendo variar entre falsa ou legítima e entre satírica ou objetiva. Para a realização desta tarefa o trabalho apresenta uma metodologia para extração de um total de 20 atributos textuais, entre cálculo das frequências de *POS tags*, média de sinônimos por termo e medidas de complexidade e estilo de texto. Vale ressaltar a solução encontrada para identificar termos de uso informal (e.g. gírias) e neologismos, verificando palavras marcadas como adjetivos, advérbios, verbos ou substantivos contra um corpus com mais de um milhão de palavras do idioma português. Os termos ou palavras não encontrados no dicionário são então computados como ”fora do vocabulário” e se somam a lista de atributos extraídos dos textos. Utilizando um *dataset* próprio, construído a partir de sites de notícias legítimas, sites de sátira e agências de checagem de fatos e após a extração dos atributos elencados acima, as notícias são classificadas utilizando algoritmos de aprendizado de máquina, que combinadas perfazem um total de quatro classes possíveis, distinguindo entre notícias com intensão ou não de enganar e se são ou não uma sátira.

Durier e Garcia(56), por sua vez, apresentam um modelo que sugere ser capaz de identificar falsos positivos ao distinguir sarcasmo de *Fake News*. O processo é realizado com o levantamento de características textuais, tais como contagem de letras maiúsculas, contagem de caracteres especiais, entre outras não explicitamente descritas. Com base nestes atributos os autores afirmam poder identificar subjetividade no texto e realizar análise de sentimentos. Em complemento a extração de atributos, os textos são codificados usando técnicas de incorporação de palavras (*word embeddings*) onde são utilizados dicionários vetorizados previamente treinados. Vale ressaltar que o *dataset* utilizado foi manualmente desenvolvido, tendo como assunto as eleições presidenciais de 2018 no Brasil.

Moraes, de Oliveira Sampaio e Charles(9) procuram identificar indícios de *Fake News* através de padrões na escrita. Seus autores propõem duas abordagens, que são comparadas ao final do trabalho para verificação de eficácia e eficiência de cada método. São elas a vetorização do texto a partir de um dicionário pré-treinado (não detalhado no artigo) e outra que realiza a engenharia de atributos para levantar características que possam

ser processadas por algoritmos de classificação. Sendo este segundo método detalhado a seguir. A partir das marcações de partes do texto (*POS Tagging*), são contabilizadas as quantidades de cada classe, bem como sua razão em relação ao total de palavras do texto. Adicionalmente os autores calcularam as quantidades e o percentual de caracteres maiúsculos e pontos de exclamação, além da polaridade do texto. A determinação da polaridade foi realizada utilizando-se um léxico de sentimentos (Sentilex), em que o método proposto calcula a polaridade de cada palavra ou *token*, comparando-as com as entradas do dicionário e verificando sua polaridade, para posteriormente calcular a polaridade total do texto a partir do somatório das da polaridades individuais. Este valor calculado é adicionado ao conjunto de atributos utilizados no processo de classificação.

De maneira a apresentar uma alternativa aos algoritmos de aprendizado supervisionado tradicionais, nos quais os modelos são treinados utilizando-se as classes *fake* e não *fake*, que se por um lado permitem que uma grande quantidade de notícias sejam classificadas em um curto período de tempo, por outro demandam de dados anotados nas classes positivas e negativas de forma balanceada, Faustini e Covões(57) apresentam uma metodologia que utiliza apenas a classe *fake*, ou quando se tem um grande desbalanceamento entre as classes, para o treinamento do modelo, chamada *One Class Classification (OCC)*, comparando o algoritmo nomeado de *Document-Class Distance OCC (DCDistanceOCC)* a outros publicados na literatura. Utilizando o cenário político brasileiro das eleições presidenciais de 2018, coletaram mensagens transmitidas pelas redes sociais do Twitter e WhatsApp, criando seu próprio *dateset* (BRACIS2019_FAKENEWS). O método utiliza um total de quatorze atributos, alguns deles obtidos diretamente do texto de forma trivial, como a contagem de letras maiúsculas ou total de caracteres, outros obtidos com a utilização de marcação de partes do texto *(POS tagging)*. Embora o método utilize o léxico de emoções LIWC, não avalia a contribuição específica das emoções, pois trabalha apenas com a média das polaridades obtidas em cada sentença.

A fim de resumir a análise dos trabalhos relacionados, a Tabela 5 apresenta um modelo comparativo que classifica os trabalhos de acordo com os seguintes critérios:

- (C1) Idioma Português - Indica se o trabalho considera textos de notícias escritas no idioma Português;
- (C2) Características Gramaticais - Informa se o método utiliza características gramaticais como atributos dos textos analisados;
- (C3) Análise de sentimentos com Polaridade - Aponta a utilização de técnicas de análise de sentimento que classificam a polaridade do texto;
- (C4) Classificação de Emoções - Manifesta a utilização de técnica explícita para identificação de emoções dos textos como atributos de classificação.

Referência	**C1**	**C2**	**C3**	**C4**
Shu et al.(58)	-	X	-	-
Ajao, Bhowmik e Zargari(8)	-	X	X	-
Bhutani et al.(54)	-	-	X	-
Morais et al.(55)	X	X	-	-
Durier e Garcia(56)	X	X	-	-
Moraes, de Oliveira Sampaio e Charles(9)	X	X	X	-
Faustini e Covões(57)	X	-	X	-

Tabela 5 – Resumo dos Trabalhos Relacionados

Diante do exposto, identificou-se que, dentre os trabalhos analisados, nenhum considera explicitamente a emoção contida no texto como um atributo para classificação de *Fake News*. Constituindo esta a lacuna de pesquisa a ser explorada neste trabalho.

4 MÉTODO PROPOSTO

Neste capítulo é apresentada a abordagem para análise da contribuição da classificação de emoções na tarefa de detecção de *Fake News*. A descrição conceitual do modelo é apresentada na Seção 4.1. Na seção 4.2 um protótipo do modelo é detalhado para implementação dos experimentos.

4.1 Descrição Conceitual

Denominado FNE, o método proposto pelo presente trabalho constrói modelos de aprendizado de máquina voltados à detecção de *Fake News*. Para tanto, baseia-se na combinação de informações linguísticas, entre elas classificações gramaticais, polaridade e emoções extraídas de textos escritos de notícias previamente rotuladas como *fake* e *não fake*, utilizando *datasets* distintos para comparação dos resultados. A Figura 13 apresenta uma visão macro-funcional do método proposto.

O FNE recebe como entrada um conjunto de notícias N, onde cada notícia $n \in N$ possui dois atributos: $n.t$ que contém o texto divulgado em n e $n.r$ o rótulo indicando se n é *fake* ou *não fake*. As próximas seções detalham as etapas do método proposto. Vamos utilizar a seguinte sentença ($n.t_1$): "**Dilma poderá renunciar para evitar vexame na votação do impeachment no senado.**", para exemplificar o processo. Neste caso, uma sentença FALSA. Após o processo de remoção de termos comuns (do inglês *Stop Words*), a única tarefa de pré-processamento realizada no texto, a mesma sentença ($n.t_1$) fica: "**Dilma poderá renunciar evitar vexame votação impeachment senado.**"

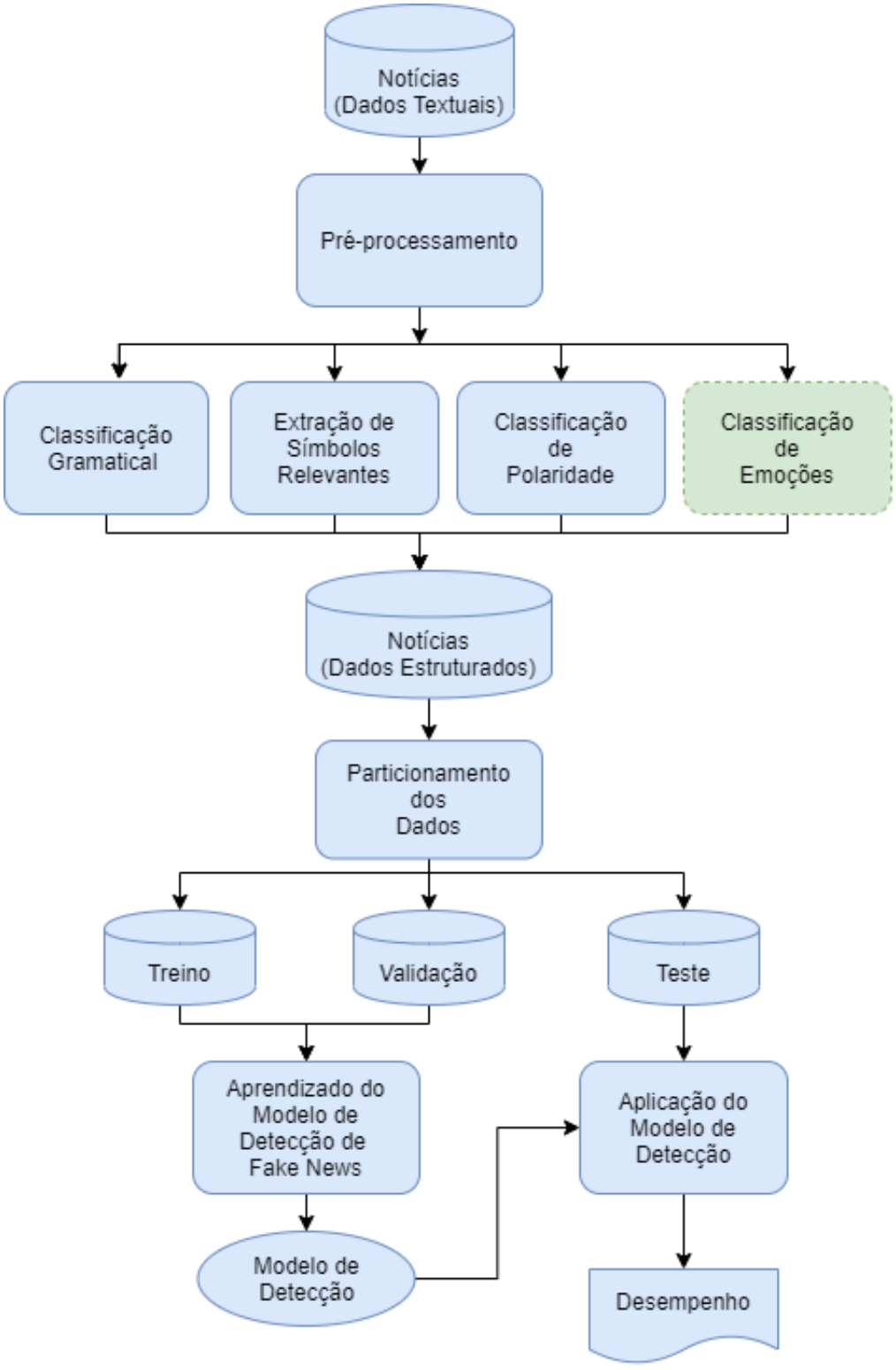


Figura 13 – Modelo Conceitual

4.1.1 Classificação Gramatical

Para cada notícia $n \in N$, esta etapa tem como objetivo identificar todas as palavras ou símbolos de pontuação, também chamados de *tokens*, existentes em n.t, gerando um conjunto ordenado $TK_{n.t} = \{p_1, p_2, ..., p_k\}$, onde cada p_i é um *token*.

Para o nosso exemplo, o conjunto será:

$TK_{n.t_1} = \{Dilma, poderá, renunciar, evitar, vexame, votação, impeachment, senado, .\}$

Assim, para cada token $p_i \in TK_{n.t}$, esta etapa aplica um processo de etiquetagem $\alpha(p_i)$ cujo objetivo é identificar a classe gramatical $cg \in CG$, a qual p_i pertence , onde $CG = \{cg_1, cg_2, ..., cg_{|CG|}\}$ é o conjunto de classes gramaticais consideradas. É importante observar, neste ponto, que o FNE é configurável, cabendo ao analista de dados escolher a implementação de α a ser adotada e, consequentemente, o conjunto CG de classes gramaticais a ser utilizado.

Para otimizar a análise pelos algoritmos de aprendizado de máquina, a frequência de *tokens* em cada classe é contabilizada, sendo os valores resultantes organizados em uma matriz linha $T_{n.t}$ de forma que cada coluna indica a quantidade de *tokens* etiquetados em uma das $|CG|$ classes de CG, conforme mostra a Equação (4.1). Nela $|n.t_{cg_j}|$ corresponde à quantidade de *tokens* de $n.t$ cuja classe gramatical é cg_j. Cabe destacar que as referidas quantidades são normalizadas pelo total de *tokens* em $TK_{n.t}$ (ie., k). Ver Algoritmo 1.

$$T_{n.t} = \frac{1}{k}\Big[\mid n.t_{cg_1} \mid, \mid n.t_{cg_2} \mid, ..., \mid n.t_{cg_{|CG|}} \mid\Big] \tag{4.1}$$

Algoritmo 1: Classificador Gramatical

Entrada: O texto *tokenizado*($TK_{n.t}$)
Entrada: CG
Saída: Matriz $T_{n.t}$
início
  Inicializa Matriz $T_{n.t}(CG)$;
  repita
    Ler token (p_i);
    $T_{n.t}[\alpha(p_i)] += 1$; $k += 1$;
  até *fim do texto*;
  $T_{n.t} = T_{n.t}/k$;
  retorna ($T_{n.t}$)
fim

Aplicando a Equação 4.1 a $TK_{n.t_1}$, onde $k = 9$, teremos (a depender do método específico usado para classificação) algo como:

$T_{n.t_1} = \frac{1}{9}[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 1, 0, 0, 3, 0, 0, 0]$

Neste exemplo, usando a *Universal POS Taging*, temos um termo classificado como "AUX", 4 como "PROPN", 1 como "PUNCT" e 3 como "VERB".

4.1.2 Identificação de Símbolos Relevantes

Esta etapa analisa cada símbolo s em $n.t$ de forma a contabilizar símbolos considerados relevantes para o processo de detecção de *Fake News*, como por exemplo, pontos

de exclamação, aspas, caracteres em caixa alta, entre outros. Quanto mais frequentes forem esses símbolos em um texto, maior a possibilidade de que esse texto seja *fake* (9). A fim de flexibilizar a escolha de quais símbolos especiais devem ser considerados, o FNE permite que o analista de dados especifique o conjunto de símbolos relevantes $CSR = \{s_1, s_2, ..., s_{|CSR|}\}$, onde, conforme o nome sugere, cada s_j é um símbolo relevante.

Assim, esta etapa percorre a cadeia $n.t$ de forma a contabilizar, para cada $s_j \in CSR$, $|n.t_{s_j}|$, i.e., o número de vezes que o símbolo s_j ocorreu em $n.t$. Após contabilizar a frequência de todos os termos de CSR, uma matriz linha $C_{n.t}$ é construída, como mostra a Equação 4.2, onde z é a quantidade total de caracteres em $n.t$ (Algoritmo 2).

$$C_{n.t} = \frac{1}{z}\Big[\mid n.t_{s_1} \mid, \mid n.t_{s_2} \mid, ..., \mid n.t_{s_{|CSR|}} \mid\Big] \tag{4.2}$$

Algoritmo 2: Identificação de Símbolos Relevantes

Entrada: O texto *tokenizado*($TK_{n.t}$)
Entrada: CSR
Saída: Matriz $C_{n.t}$
início
  Inicializa Matriz $C_{n.t}(CSR)$;
  repita
    se *caracter* $\in CSR$ **então**
      $C_{n.t}[CSR]$ += 1 ;
    fim
    $z+ = 1$;
  até *fim do texto*;
  $C_{n.t} = C_{n.t}/z$;
  retorna ($C_{n.t}$)
fim

Sendo $z = 57$ para $n.t_1$, e o conjunto CSR = {MAIÚSCULAS, ASPAS, EXCLAMAÇÃO} (i.e.) teremos :

$C_{n.t_1} = \frac{1}{57}[1, 0, 0]$

4.1.3 Classificação de Polaridade

Para cada *token* $p_i \in TK_{n.t}$, esta etapa aplica a função parcial $P : L \to \{-1, 0, 1\}$ que procura se p_i pertence ao léxico L a fim de recuperar o valor de polaridade associado a p_i. Os valores de polaridade recuperados são somados à polaridade total de $n.t$, conforme indicado na Equação (4.3) e o Algoritmo 3. Para estabelecer uma independência em relação aos diferentes tamanhos de texto, os valores de polaridade $P_{n.t}$ são normalizados (i.e., processados de forma a assumir valores no intervalo $[-1, 1]$), considerando os resultados de polaridade obtidos para todas as notícias do *dataset*. É importante destacar, neste ponto,

que cabe ao analista de dados configurar o FNE com o léxico de polaridade desejado, incluindo a função P a ser utilizada.

$$P_{n.t} = \sum_{i=1}^{k} P(p_i) \tag{4.3}$$

Algoritmo 3: Classificador de Polaridade

Entrada: O texto *tokenizado*($TK_{n.t}$)
Entrada: Léxico de Polaridade
Saída: $P_{n.t}$
início
  Inicializa $P_{n.t} = 0$;
  repita
    Ler token (p_i);
    $P_{n.t}$ += P(p_i);
  até *fim do texto*;
  retorna ($P_{n.t}$)
fim

A sentença $n.t_1$, apresenta um dos termos como polaridade negativa e os demais termos com polaridade neutra , terá o valor de $P_{n.t_1} = -1$, depois do processo de normalização teremos o valor de $P_{n.t_1} = -0,01$.

4.1.4 Classificação de Emoção

Usando léxicos afetivos (i.e., dicionários que relacionam emoções às palavras de um texto), é possível extrair atributos que permitem identificar a presença de emoções em textos (11).

Cada palavra $p_i \in n.t$, é classificada inicialmente segundo uma função binária $affect(p_i)$, cujo resultado é 1, caso p_i pertença ao léxico afetivo escolhido pelo analista, ou 0, caso contrário. Então, cada p_i em que $affect(p_i) = 1$ é submetida às funções binárias mutuamente exclusivas $posemo(p_i)$ ou $negemo(p_i)$. Caso a emoção relacionada à p_i seja positiva no léxico afetivo, então $posemo(p_i) = 1$ e $negemo(p_i) = 0$. Caso seja negativa, então $posemo(p_i) = 0$ e $negemo(p_i) = 1$. Cada p_i com $negemo(p_i) = 1$ pode ainda ser subclassificada em uma de três emoções negativas: raiva, ansiedade ou tristeza. Para tanto, são utilizadas, respectivamente, as funções binárias $anger(p_i), anx(p_i)$ e $sad(p_i)$, que retornam valor 1, caso p_i se enquadre na referida emoção, e 0, caso contrário como descrito pelo Algoritmo 4. Como resultado deste processo, tem-se uma matriz linha $E_{n.t}$ representada na Equação (4.4), onde cada coluna indica o somatório de ocorrências de *tokens* em cada uma das seis categorias afetivas apresentadas acima.

$$E_{n.t} = \left[\sum_{i=1}^{k} affect(p_i), \sum_{i=1}^{k} posemo(p_i), \sum_{i=1}^{k} negemo(p_i), \sum_{i=1}^{k} anx(p_i), \sum_{i=1}^{k} anger(p_i), \sum_{i=1}^{k} sad(p_i)\right] \tag{4.4}$$

Algoritmo 4: Classificador de Emoções

Entrada: O texto $tokenizado(TK_{n.t})$
Entrada: Léxico de Emoções
Saída: Matriz $E_{n.t}$
início
  Inicializa Matriz $E_{n.t}$;
  repita
    Ler token (p_i);
    $E_{n.t}[Emocao(p_i)] += 1$;
  até *fim do texto*;
  retorna $(E_{n.t})$
fim

Novamente aqui, usando a sentença de exemplo $n.t_1$, teremos:

$E_{n.t_1} = [2, 0, 2, 1, 0, 1]$

Sendo que dois termos (renunciar e evitar) foram enquadrados no léxico afetivo (*affect* = 2), ambos os termos estão também classificados como emoções negativas (*negemo* = 2). Um se enquadra como tristeza ($sad = 1$) e outro como ansiedade ($anx = 1$).

Depois de normalizados por todo o conjunto de dados, teremos i.e.:

$E_{n.t_1} = [0.222, 0.0, 0.222, 0.111, 0.0, 0.111]$

4.1.5 Formação do Conjunto de Dados Estruturados

Ao final do processo de categorização descrito nas etapas anteriores, para cada notícia $n \in N$ um conjunto de atributos estruturados Ne_n é formado e definido pela tupla indicada na Equação 4.5.

$$Ne_n = (T_{n.t}, C_{n.t}, P_{n.t}, E_{n.t}) \tag{4.5}$$

Para o exemplo citado teremos então algo próximo do apresentado abaixo:

$Ne_{n_1} = (0.0, 0.0, 0.0, 0.1111111111111111, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.4444444444444444, 0.1111111111111111, 0.0, 0.0, 0.3333333333333333, 0.0, 0.0, 0.017543859649122806, 0.0, 0.0, 0.3739837398373984, 0.2222222222222222, 0.0, 0.2222222222222222, 0.1111111111111111, 0.0, 0.1111111111111111)$

O conjunto Ne formado pelas tuplas geradas pelas etapas descritas anteriormente, a partir de todas as notícias de n, é então armazenado para posterior processamento pelas etapas seguintes. A Equação 4.6 descreve formalmente tal conjunto.

$$Ne = \{Ne_n | n \in N\} \tag{4.6}$$

4.1.6 Particionamento de Dados, Aprendizado e Aplicação do Modelo

A etapa de particionamento de dados é responsável por separar Ne em três conjuntos: treino, validação e teste. Tal separação ocorre de forma aleatória, porém estratificada, assegurando a mesma distribuição de classes em cada conjunto.

Em seguida, na etapa de aprendizado do modelo de detecção de Fake News, o FNE treina cada um dos algoritmos de classificação indicados pelo analista de dados com as notícias do conjunto de treino. O conjunto de validação é utilizado de forma a selecionar o melhor modelo gerado por cada algoritmo, a partir de diferentes configurações de parâmetros especificados pelo analista. Note que cabe ao analista escolher a métrica de avaliação dos modelos de classificação a ser utilizada (e.g. acurácia, precisão, F1-score, dentre outras).

Por fim, a etapa de aplicação do modelo de detecção é responsável por avaliar no conjunto de teste o desempenho do melhor modelo de classificação gerado por cada algoritmo na etapa anterior. A mesma métrica utilizada no treinamento dos modelos deve ser utilizada nesta etapa.

4.2 Protótipo

O protótipo do FNE utilizado nos experimentos deste trabalho foi implementado na plataforma KNIME [1] (59), integrada a um banco de dados MySQL. Tanto os conteúdos dos *datasets*, como os atributos das notícias calculados pelos algoritmos do modelo são persistidos no banco de dados. Algumas das funções do modelo, foram implementadas em linguagem Python, como por exemplo a execução das funções das bibliotecas de *POS Tagging* e alguns dos algoritmos de aprendizado de máquina. As métricas de performance dos algoritmos de aprendizado foram armazenados em planilhas na nuvem, como mostrado no modelo simplificado da Figura 14.

Este protótipo permite ao analista de dados configurar os métodos de cálculo dos atributos de alguns dos componentes da Equação 4.5, conforme as opções indicadas na Tabela 6. A coluna "Opções de implementação" indica as bibliotecas que calculam os atributos apontados na coluna "Atributos".

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

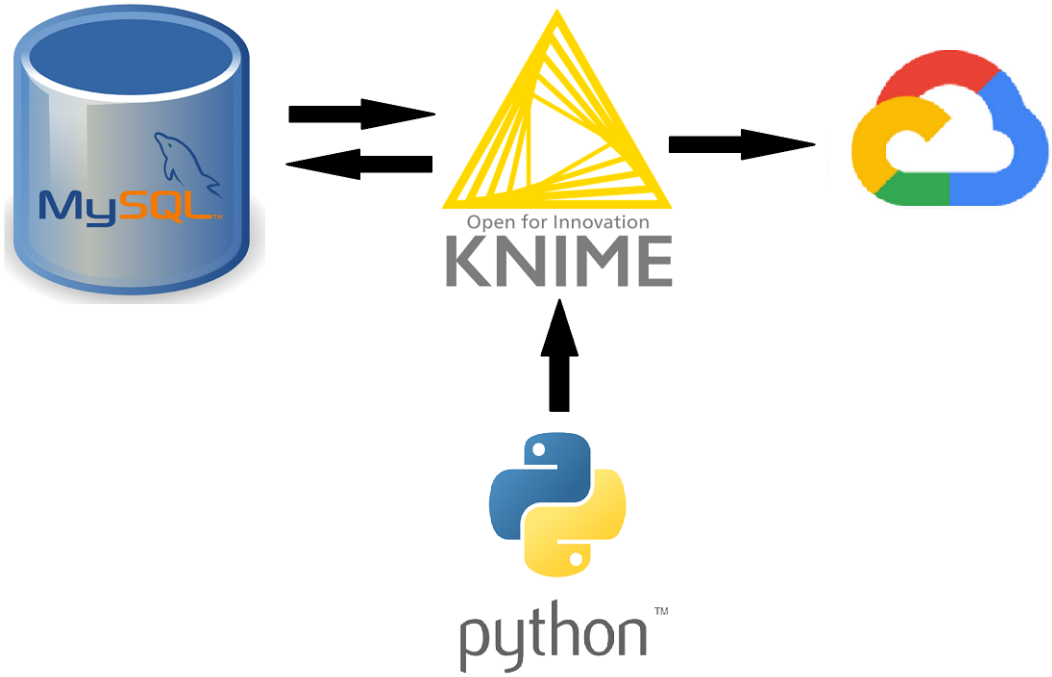


Figura 14 – Modelo esquemático do protótipo

Para o componente $T_{n.t}$ (ie. classificação gramatical), duas opções foram implementadas. Primeiramente a função *token.pos_* da biblioteca SpaCy [2] desenvolvida em *Python*. Esta biblioteca foi escolhida por ter suporte consistente ao idioma português na tarefa de marcação de partes do texto (*Parts-of-Speech Tagging*). Como alternativa, foi utilizada uma parte, da versão mais recente na época da execução dos experimentos, do dicionário para o português do LIWC 2015 (Seção 2.2.6.2), com os atributos que representam classes gramaticais. Foi implementado um *script* que conta a ocorrência dos atributos do subconjunto do LIWC e associa os resultados à matriz $T_{n.t}$.

O Conjunto de Símbolos Relevantes CSR foi configurado com as letras maiúsculas do alfabeto romano (MAIÚSCULAS), o ponto de exclamação (!) e as aspas (”). Este conjunto mostrou-se importante para identificação de *Fake News* em (9) e (56). Denominado Extrator FNE, um módulo foi implementado de forma a contabilizar a quantidade de ocorrências de cada símbolo do conjunto CSR (componente $C_{n.t}$).

A classificação de polaridade (componente $P_{n.t}$) foi desenvolvida com base no dicionário Sentilex-PT (42), que consiste de um léxico especificamente concebido para esta tarefa em textos redigidos em português que possui 7.014 palavras ou expressões relacionadas individualmente a um valor de polaridade. Foi desenvolvido um *script* que percorre o texto de entrada, que separa cada palavra, recupera a polaridade no léxico e acumula a polaridade total do texto.

[2] https://spacy.io/

Componente	Opções de Implementação	Atributos
$T_{n.t}$	SpaCy	$\mid n.t_{ADJ} \mid, \mid n.t_{X} \mid, \mid n.t_{ADP} \mid, \mid n.t_{ADV} \mid, \mid n.t_{AUX} \mid,$ $\mid n.t_{CONJ} \mid, \mid n.t_{CCONJ} \mid, \mid n.t_{DET} \mid, \mid n.t_{INTJ} \mid,$ $\mid n.t_{NOUN} \mid, \mid n.t_{NUM} \mid, \mid n.t_{PART} \mid, \mid n.t_{PRON} \mid,$ $\mid n.t_{PROPN} \mid, \mid n.t_{PUNCT} \mid, \mid n.t_{SCONJ} \mid, \mid n.t_{SYM} \mid,$ $\mid n.t_{VERB} \mid$
	LIWC	$\mid n.t_{funct} \mid, \mid n.t_{pronoun} \mid, \mid n.t_{ppron} \mid, \mid n.t_{i} \mid,$ $\mid n.t_{we} \mid, \mid n.t_{you} \mid, \mid n.t_{shehe} \mid, \mid n.t_{they} \mid,$ $\mid n.t_{ipron} \mid, \mid n.t_{article} \mid, \mid n.t_{prep} \mid, \mid n.t_{auxverb} \mid,$ $\mid n.t_{adverb} \mid, \mid n.t_{conj} \mid, \mid n.t_{negate} \mid, \mid n.t_{verb} \mid,$ $\mid n.t_{adj} \mid, \mid n.t_{interrog} \mid$
$C_{n.t}$	Extrator FNE	$\mid n.t_{,,} \mid, \mid n.t_{!} \mid, \mid n.t_{MAIUSCULAS} \mid$
$P_{n.t}$	Sentilex-PT	P (Polaridade de n.t)
$E_{n.t}$	LIWC	affect, posemo, negemo, anx, anger, sad
	Affect-BR	

Tabela 6 – Conjunto de atributos obtidos dos textos

Quanto à classificação de emoções, dois léxicos com recursos semelhantes foram utilizados como alternativas de implementação: o LIWC (60) e o Affect-Br (49) . Mencionado anteriormente, o LIWC foi também usado para a classificação de palavras ligadas às emoções, sendo de 2015 a versão utilizada e a mais recente do seu dicionário para o idioma português. Já o Affect-Br foi construído a partir da versão em Inglês do LIWC2015. Para ambas as opções, foi implementado um *script* que identifica os atributos relacionados à emoção de cada palavra do texto, utilizando o dicionário selecionado entre as duas opções e contabiliza o componente $E_{n.t}$. Os valores são normalizados entre 0 e 1 para todo o conjunto de notícias analisado.

Como pode ser percebido, diferentes versões do FNE podem ser configuradas e avaliadas. De maneira a padronizar a denominação dessas diferentes versões, optou-se pela seguinte notação: $FNE(OpT, OpC, OpP, OpE)$, onde $OpT \in \{SpaCy, LIWC\}$, $OpC \in \{FNE\text{-}CSR\}$, $OpP \in \{Sentlex\text{-}Pt\}$ e $OpE \in \{Affect\text{-}br, LIWC\}$.

Por fim, os algoritmos de classificação implementados nesta versão do FNE foram *Naive Bayes*, *AdaBoost* e *SVM* conforme evidenciados em (9) e adicionados *Gradient Boost* e *KNN*. Sua escolha deveu-se, basicamente, à popularidade e ao histórico bem sucedido desses algoritmos em diversos problemas envolvendo classificação (44).

5 EXPERIMENTOS E RESULTADOS

Nesta seção, apresentaremos aspectos específicos desta implementação, descrevendo primeiramente as características dos *datasets* selecionados. Em seguida detalharemos a execução dos experimentos, a metodologia de comparação e avaliação e os resultados obtidos.

5.1 Datasets

Algoritmos de aprendizado de máquina supervisionados aprendem a identificar padrões recorrentes em dados anotados (*datasets*). Anotação ou rotulação de dados é o processo de adicionar informações que evidenciem a utilidade da informação presente. A anotação de dados é portanto um estágio indispensável do pré-processamento da informação. Depois que um algoritmo processa dados anotados suficientes, ele pode começar a reconhecer os mesmos padrões quando apresentado a dados novos não anotados. Com esta finalidade, os cientistas de dados precisam usar dados limpos e anotados para treinar modelos de aprendizado de máquina, na tarefa de identificação de *Fake News*.

De maneira a avaliar a efetividade do método proposto, foram consideradas alternativas de *datasets* para representar o conjunto de notícias N. Dentre as opções avaliadas para o idioma português, temos:

- Fake.Br (61)
- Factck.BR (62)
- FakeNewSet (63)

5.1.1 FakeBR

O corpus Fake.Br (61), é uma base composta por um conjunto balanceado de notícias *fake* e *não fake* em Português e possui os textos completos das notícias. A coleta, realizada de forma manual, selecionou textos de notícias falsas disponíveis na Internet no período de dois anos (janeiro/2016 a janeiro/2018). Notícias que não podiam ser corretamente classificadas foram eliminadas, mantendo-se apenas as que eram consideradas totalmente falsas. Em seguida de forma semi-automática buscaram pela notícia verdeira correspondente em relação ao fato. Como uma forma de promover a normalização dos dados entre as notícias *fake* e *não fake*, uma vez constatado que notícias verdadeiras apresentam em média textos mais longos que os das notícias falsas, e de maneira a evitar um viés nas análises, os autores truncaram os textos pelo número de palavras de suas contrapartes

falsas. Sendo que o *dataset* apresenta as duas versões (uma com textos normalizados pelo tamanho e outra com os textos completos) em tabelas distintas.

Neste *dataset*, seis categorias diferentes proporcionam uma diversidade de temas, embora política concentre a maior parte das notícias. A Tabela (7) apresenta um resumo estatístico da composição das notícias do corpus Fake.Br.

Atributos		Fake	Não Fake
quantidade de notícias		3600	3600
quantidade média de tokens		216,1	1268,5
quantidade média de sentenças		12,7	54,8
tamanho médio das sentenças em palavras		15,3	21,1
categorias	ciência e tecnologia	1,5%	
	economia	0,7%	
	política	58%	
	religião	0,7%	
	sociedade	17,7%	
	tv_celebridades	21,4%	

Tabela 7 – Características do Fake.BR

O *dataset* ainda fornece de forma adicional algumas métricas pré-calculadas de características linguísticas, que não foram consideradas neste trabalho, bem como não foi considerada a versão truncada dos textos.

5.1.2 Factck.BR

Em 2019 estimava-se que existiam 160 agências de *fact-checking* em todos o mundo e cerca de 9 iniciativas no Brasil. Em 2020 este número subiu para 298 no mundo e 10 no Brasil segundo o site Reporters' Lab da Universisadade de Duke nos EUA [1]. O site Observatório da Imprensa [2], define *fact-checking* como:

> "... uma checagem de fatos, isto é, um confrontamento de histórias com dados, pesquisas e registros. Se, por exemplo, um político jura que nunca foi acusado de corrupção, há registros judiciais que irão atestar se é verdade. Se o governo diz que a inflação diminuiu, é preciso checar nos índices se isso realmente ocorreu. E se uma corrente diz que há um projeto de lei para cancelar as eleições, é preciso conferir nas propostas em tramitação se essa informação é real. O *fact-checking* é uma forma de qualificar o debate público por meio da apuração jornalística. De checar qual é o grau de verdade das informações."

[1] https://reporterslab.org/fact-checking/
[2] http://www.observatoriodaimprensa.com.br/checagem-de-informacoes/o-que-e-fact-checking/

O Factck.br é um *dataset* desenvolvido para utilização em algoritmos de aprendizado de máquina para classificação de *Fake News* no idioma português do Brasil, que apresenta texto de notícias supostamente falsas e sua respectiva classificação realizada por uma variedade de agências de *Fact Checking* (62). Os dados são coletados do *ClaimReview*[3], uma plataforma de dados estruturados usada por agências de checagem de fatos para compartilhar seus resultados em mecanismos de busca que permite a coleta de dados em tempo real. Para composição do Factck.br seus autores consideraram três das mais ativas agências do Brasil. Sendo elas: Aos Fatos[4], Lupa[5] e Truco[6].

Atributo	Descrição
URL	Site que contém a revisão
Author	Identificação do autor da Revisão
datePublished	Data de publicação da checagem
claimReviewed	Notícia analisada
reviewBody	Texto analisado
title	Título do artigo
ratingValue	Classificação numérica
bestRating	Valor máximo da escala de classificação
alternativeName	Rótulo da classificação

Tabela 8 – Estrutura do dataset Factck.BR

Sua estrutura armazenada em arquivo texto e apresentada na Tabela 8, é composta por nove atributos separados por tabulação (TSV), sendo uma linha para cada revisão de notícia. Cada agência de checagem de fatos utiliza sua própria escala de classificação, com rótulos que podem ser, por exemplo: falso, verdadeiro, impossível de provar, distorcido, impreciso, entre outros. Sendo que no ClaimReview a classificação pode ser obtida de forma linear, onde 1 é falso e o valor máximo da escala (bestRating), corresponde ao verdadeiro, com os valores intermediários representando outras classificações. A Tabela 9 apresenta a distribuição das notícias no *dataset* por cada classificação presente.

[3] https://schema.org/ClaimReview
[4] https://www.aosfatos.org/
[5] https://piaui.folha.uol.com.br/lupa/
[6] https://apublica.org/checagem/

Rótulo da Classificação	Quantidade de Notícias
Ainda é cedo para dizer	6
De olho	3
Discutível	12
Distorcido	25
Exagerado	87
Falso	615
Impossível provar	20
Sem contexto	42
Subestimado	6
Verdadeiro	119
Verdadeiro, mas	4
distorcido	29
exagerado	4
falso	328
impreciso	2
insustentável	5
outros	1
verdadeiro	1

Tabela 9 – Estatísicas do Factck.BR

5.1.3 FakeNewsSet

Como uma estrutura mais completa, o FakeNewsSet (63) apresenta atributos sobre o conteúdo das notícias, assim como outros atributos. Utiliza as agências de checagem de fatos (e.g. Lupa, Aos Fatos e AFP) para definir um rótulo para as notícias e classificar como *fake* ou não *fake*. E a rede social Twitter para obter dados da propagação das mesmas. Para a proposta desta dissertação, apenas os dados do componente texto das 600 notícias serão considerados, bem como o campo rótulo do resultado da checagem. Um resumo estatístico do dataset é apresentado na Tabela 10.

Recursos	Fake	Not Fake
Notícias	300	300
Divulgações (tweets)	20.498	6.506
Divulgações (retweets)	17.927	4.816
Usuários divulgadores	15.983	
Média de seguidores por usuário	96.239	
Média de seguidos por usuário	6.120	
Média de notícias por usuário	1,53	
Média de usuários por notícia	40,7	

Tabela 10 – Estatísticas do FakeNewsSet

5.2 Experimentos

Os experimentos foram realizados utilizando-se inicialmente o *dataset* Fake.BR (5.1.1) e os resultados detalhados destes experimentos são mostrados nesta seção. Posteriormente, o método é avaliado nos demais *datasets* apresentados na Seção 5.1 para validação.

5.2.1 Baselines

Para realização dos experimentos, foram criados dois *baselines*, indicados abaixo, para aferir a capacidade de detecção de *Fake News* sem o uso da classificação de emoções, mas variando o método para classificação gramatical (SpaCy e LIWC).

Vale reforçar que tanto a biblioteca SpaCy, como os atributos gramaticais do LIWC proporcionam mecanismos diferentes para a classificação gramatical e cada um destes mecanismos foi avaliado separadamente. Note que o símbolo '_' foi utilizado para denotar versão de FNE em que a etapa de classificação de emoção não é realizada e, portando, os atributos extraídos nesta etapa não são considerados para fins de construção dos modelos de classificação.

baseline 1 → FNE(SpaCy, FNE-CSR, *Sentlex-PT*, _)

baseline 2 → FNE(LIWC, FNE-CSR, *Sentlex-PT*, _)

5.2.2 Inclusão dos Léxicos de Emoção

A Figura 15 apresenta, a título ilustrativo, uma representação da amplitude dos atributos de emoções obtidos com o uso do *LIWC*. Este exemplo foi construído a partir de uma amostra aleatória de 250 notícias *fake* e 250 não *fake* do Fake.Br. Nelas, podem-se observar uma diferença de distribuição das emoções, em especial, no que se refere à presença de emoções negativas. Que foi interpretado como um indicativo importante de evidências de que estas diferenças podem ser relevantes na diferenciação das notícias.

Para avaliar o efeito da utilização da emoção presente nos textos das notícias, na classificação de *Fake News*, os dois léxicos afetivos Affect-br e LIWC foram empregados isoladamente entre eles. Desta forma, quatro versões do FNE foram configuradas para serem comparadas com os baselines:

FNE(SpaCy, FNE-CSR, *Sentlex-PT*, LIWC)

FNE(SpaCy, FNE-CSR, *Sentlex-PT*, Affect-BR)

FNE(LIWC, FNE-CSR, *Sentlex-PT*, LIWC)

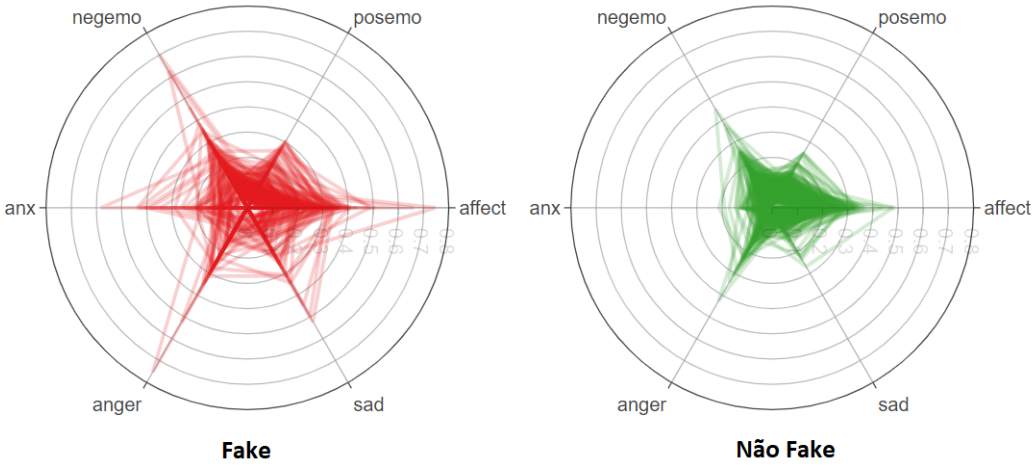


Figura 15 – Distribuição de emoções em textos FAKE e NÃO FAKE

FNE(LIWC, FNE-CSR, *Sentlex-PT*, Affect-BR)

5.2.3 Classificação

A acurácia (Seção 2.6) foi a métrica de avaliação escolhida para os modelos de classificação. Os algoritmos classificadores utilizados nos experimentos foram os já mencionados na Seção 4.2. A Tabela 11 resume a configuração final de hiper-parâmetros utilizada em cada um deles.

algoritmo	parâmetros
Naive Bayes (NB)	padrão
AdaBoost (AB)	Random Forest I = 10 (iterações) depth = 0 (sem limite)
SVM	C = 1000 gama = 0.001 kernel linear
Gradient Boost (GB)	níveis =5 modelos = 100 learning rate = 0.1
KNN	k = 3

Tabela 11 – Parametrização dos algoritmos de classificação

Tendo levantado todos os atributos das notícias e gerado o conjunto estruturado Ne, um sistema de seleção garante que os atributos corretos são passados para os algoritmos de classificação em cada uma das fases do processo (*baselines* e experimentos com emoções). Paralelamente estes atributos são processados pelos os algoritmos implementados. O

conjunto Ne é dividido em blocos distintos nomeados como "Treinamento", "Validação" e "Testes", na proporção de 70%, 15% e 15% respectivamente. Com os dados do conjunto de treinamento os modelos inferem a função que representa o comportamento das notícias. A verificação deste modelo é realizada usando o conjunto de validação. Durante este processo os hiper-parâmetros dos algoritmos são ajustados utilizando-se a função GridSearch[7] da biblioteca SciKit-learn do Python, de forma a produzir os melhores resultados com base no indicador de desempenho escolhido. Ao final desta etapa, para garantir que os algoritmos não estão super-ajustados aos dados de treinamento, os modelos são submetidos ao conjunto de dados de testes.

5.2.4 Resultados

Para confirmar que os resultados obtidos não seriam associados a um viés relacionado às características de um conjunto de dados específico, replicamos os experimentos com os *datasets* listados na Seção 5.1. Como as características das notícias podem diferir entre os *datasets*, quer seja em relação aos assuntos, em relação aos sites de onde foram coletadas ou aos estilos linguísticos, estes experimentos permitem verificar se o método é robusto e independe do conjunto de dados.

Cabe ressaltar que toda a metodologia foi repetida de forma rigorosamente idêntica, quer seja nos hiper-parâmetros dos algoritmos de classificação ou na composição dos *baselines*.

5.2.4.1 Fake.BR

A Tabela 12 resume os resultados dos experimentos realizados sobre o *dataset* Fake.BR. De uma forma geral, existem evidências que apontam para a validade da hipótese levantada neste trabalho de que a combinação de informações gramaticais, com a polaridade e as emoções presentes nos textos das notícias pode levar a melhores modelos de detecção de *Fake News* do que aqueles modelos que consideram apenas informações gramaticais e a polaridade desses textos. Abaixo segue uma análise mais detalhada sobre os resultados obtidos.

Em 70% (7 em 10) das comparações entre as versões do FNE com o uso da classificação de emoções e seus respectivos *baselines*, percebe-se que alguma versão baseada nas emoções superou o *baseline* correspondente.

Uma comparação direta entre os *baselines 1* e *2* revela uma superioridade do uso do LIWC em relação ao SpaCy como ferramentas para a realização da classificação gramatical. De fato, com exceção dos experimentos realizados com o SVM, em todos os demais o *baseline 2* superou o *baseline 1*.

[7] https://scikit-learn.org/stable/modules/grid_search.html

Numa comparação entre a influência dos léxicos afetivos em relação ao *baseline 1*, percebe-se uma pequena superioridade do LIWC em relação ao Affect-BR. O mesmo já não ocorre com relação ao *baseline 2*. Nesta visão comparativa, o Affect-BR se mostrou superior em três dos cinco algoritmos de classificação utilizados.

	baseline1	baseline 1 + emoções		baseline2	baseline 2 + emoções	
classificadores	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, \_)$	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, LIWC)$	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, \_)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, LIWC)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$
NB	79,42%	81,26%	**81,57%**	83,10%	**86,35%**	84,46%
GB	90,89%	**92,13%**	91,90%	91,40%	92,40%	**92,53%**
AB	88,08%	**89,25%**	89,22%	**90,90%**	90,47%	89,88%
SVM	88,10%	**88,15%**	88,06%	**84,10%**	78,76%	83,33%
KNN	**77,81%**	77,29%	76,93%	80,10%	78,99%	**80,79%**

* Em negrito os valores máximos de cada experimento
** Em vermelho o melhor resultado geral

Tabela 12 – Valores de Acurácia dos Experimentos com o Fake.BR

Pode-se ainda observar que a escolha do algoritmo de classificação pode influenciar negativamente os resultados, apesar da inclusão dos atributos de emoção. Esse comportamento pode ser visto nos resultados dos testes com o *baseline 1*, quando utilizados o KNN, e nos testes com o *baseline 2*, nas linhas referentes ao AdaBoost e SVM.

Não obstante, pode-se notar que a utilização tanto do Naive Bayes, quanto do Gradient Boost, apresentam melhorias na acurácia dos experimentos. Com o destaque para o NB, que apresentou a maior diferença (3,25%) em relação ao *baseline 2* e para o GB, que alcançou o melhor resultado geral de acurácia (92,53%).

A métrica de acurácia nestes experimentos mostrou-se interessante, por mostrar de forma mais efetiva a qualidade dos resultados. Tomemos como referência os resultados obtidos para a classe FALSO (textos anotados como falsos) na configuração $FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$, como podemos ver na Tabela 13, embora alguns classificadores apresentem um valor elevado de revocação, a quantidade de falsos positivos é alta, reduzindo o valor de acurácia total do processo.

	Métricas		
Algoritmos	Revocação	Precisão	F-1
NAIVE BAYES	92,56%	79,66%	85,62%
GRADIENT BOOST	92,94%	92,18%	92,56%
ADABOOST	90,14%	89,67%	89,90%
SMV	88,53%	80,20%	84,16%
KNN	66,67%	92,92%	77,63%

Tabela 13 – Comparação entre Métricas

Dentre os algoritmos escolhidos para os experimentos, os modelos de *Boosting* (GRADIENT BOOST e ADABOOST), independente do resultado absoluto de acurácia, foram o que mostraram melhor balanceamento entre as métricas de precisão e revocação, apresentando resultados mais consistentes (Figura 16), com menor taxa de falsos positivos, conclusão corroborada pelos valores de F-1 para estes algoritmos. Já o KNN, embora apresente um valor elevado para a métrica de precisão (indicando que 92,92% dos registros classificados como falsos eram realmente falsos), sua capacidade de cobertura é baixa identificando apenas 66,67% dos casos (revocação).

Os resultados obtidos mostraram que a utilização dos léxicos de emoção pode, de fato, trazer melhorias na tarefa de classificação de *Fake News*. Tanto o LIWC como o Affect-BR apresentaram resultados semelhantes, com uma pequena vantagem para o Affect-BR, quando utilizado em conjunto com o GB no *baseline 2*. Entretanto, fica claro que a escolha dos algoritmos de classificação é crucial para a obtenção dos resultados favoráveis. A íntegra dos valores obtidos para as diversas métricas apuradas estão relacionados no Apêndice B.

5.2.4.2 FakeNewsSet

Como pode ser observado na Tabela 14, repetimos os experimentos com o conjunto de dados FakeNewsSet e apenas em duas situações a comparação dos experimentos mostrou-se favorável aos conjuntos de *baseline*, ambas ocorreram quando se utilizou o KNN como algoritmo de classificação. Pode-se notar que os melhores resultados absolutos na tarefa de classificação ficaram por conta do AdaBoost, tanto para o *baseline* 1 quanto para o 2. Novamente aqui a maior diferença percentual (9,83pp) foi apresentada pelo Naive Bayes.

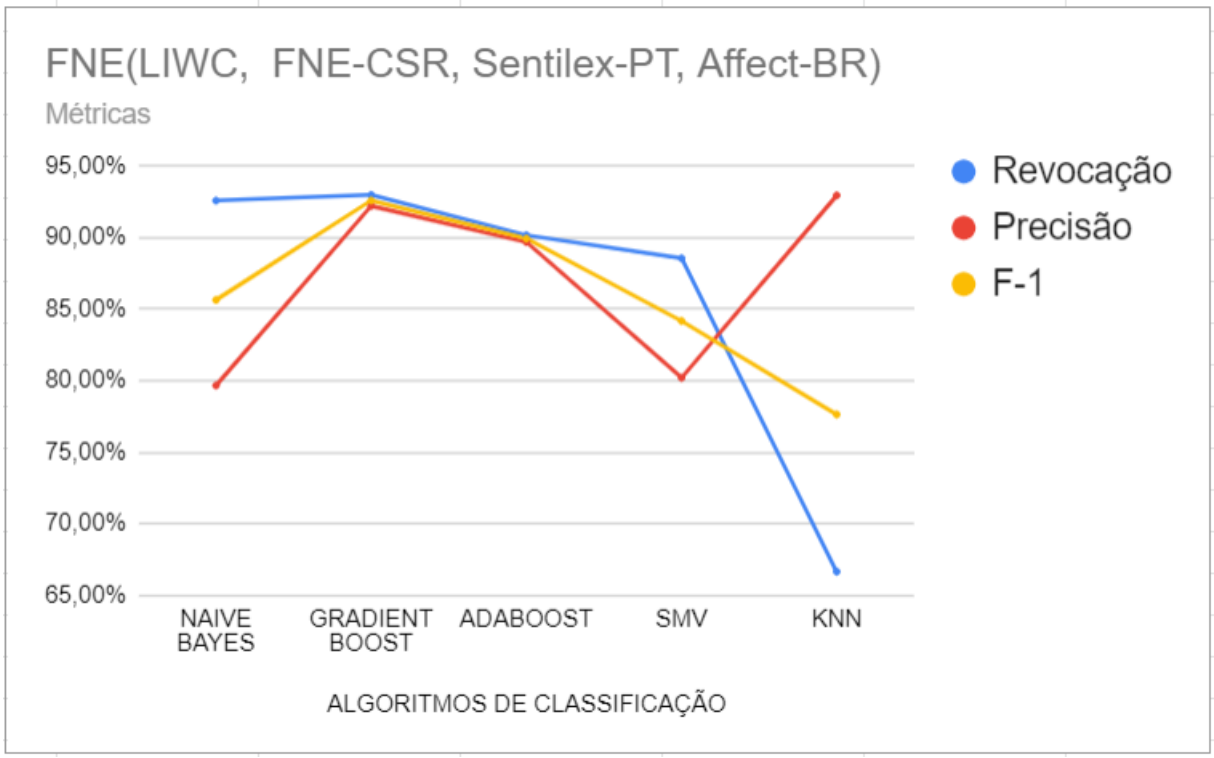


Figura 16 – Comparação entre Métricas

5.2.4.3 Factck-BR

O Factck.BR, apresenta um conjunto complexo de rotulação das notícias (Tabela 9), com classificações intermediárias entre o verdadeiro e o falso. Pode-se notar que algumas classificações apresentam definições imprecisas, não sendo possível determinar de forma inequívoca se o texto se trata de uma notícia falsa ou verdadeira. Além disso, as classificações não estão com seus títulos padronizados, permitindo duplicidades. Por estes motivos, para efeito deste trabalho, apenas as classes verdadeiro e falso foram utilizadas, unindo os conteúdos de Falso com falso e Verdadeiro com verdadeiro.

Ao realizar esta separação, pôde-se notar um desbalanceamento entre as classes com 943 notícias falsas e apenas 120 verdadeiras. Alguns algoritmos de aprendizado de máquina supervisionado como árvores de decisão, requerem uma distribuição igualitária entre as classes para obter uma boa performance de classificação. De forma a minimizar os efeitos de viés nos algoritmos de classificação por conta deste desbalanceamento, foi utilizada a técnica SMOTE (64) (*Synthetic Minority Over-sampling Technique*), que ajusta a quantidade de amostras adicionando de forma artificial registros na classe minoritária. O algoritmo cria registros sintéticos entre um registro real e o seu vizinho mais próximo da mesma classe, selecionando pontos aleatórios entre estes registros e determinando novos atributos baseados nestes pontos.

A Tabela 15 apresenta a consolidação dos resultados da aplicação do modelo a este *dataset*. Reforça-se aqui a confirmação da hipótese, quando se verifica que em apenas uma

	baseline1	baseline 1 + emoções		baseline2	baseline 2 + emoções	
classificadores	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, \_)$	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, LIWC)$	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, \_)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, LIWC)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$
NB	54,33%	62,50%	**62,67%**	70,17%	79,00%	**80,00%**
GB	89,50%	89,17%	89,50%	86,33%	**87,00%**	86,50%
AB	89,33%	**90,17%**	89,33%	87,17%	89,67%	**90,67%**
SVM	83,33%	**85,83%**	85,17%	81,50%	**83,50%**	80,83%
KNN	**69,17%**	69,00%	69,00%	**71,17%**	69,67%	67,33%

* Em negrito os valores máximos de cada experimento
** Em vermelho o melhor resultado geral

Tabela 14 – Valores de Acurácia dos Experimentos com o FakeNewsSet

das comparações o *baseline* apresentou resultado melhor que o seu respectivo experimento com a associação dos atributos de emoções.

	baseline1	baseline 1 + emoções		baseline2	baseline 2 + emoções	
classificadores	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, \_)$	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, LIWC)$	$FNE(SpaCy, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, \_)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, LIWC)$	$FNE(LIWC, FNE\text{-}CSR, Sentilex\text{-}PT, Affect\text{-}BR)$
NB	50,36%	50,00%	**50,54%**	**57,14%**	50,71%	50,36%
GB	85,18%	87,32%	**88,75%**	79,11%	81,43%	**82,68%**
AB	87,50%	89,82%	91,96%	76,25%	**81,25%**	80,71%
SVM	71,79%	71,79%	**76,07%**	70,00%	70,00%	70,00%
KNN	77,50%	81,79%	**82,14%**	72,32%	69,82%	**74,11%**

* Em negrito os valores máximos de cada experimento
** Em vermelho o melhor resultado geral

Tabela 15 – Valores de Acurácia dos Experimentos com o Factck.BR

5.2.4.4 Análise Global dos Resultados

Convém aqui ressaltar que nosso objetivo era de provar que o uso de atributos linguísticos poderiam ser utilizados para classificação de *Fake News*, e especial utilizando-se a classificação de emoções contidas nestes textos. Pode-se notar, em uma avaliação extensiva dos resultados obtidos entre os três *datasets* utilizados nos experimentos, que a inclusão dos atributos de emoção em conjunto com os demais atributos linguísticos, resultaram em valores expressivos de acurácia e uma melhoria em relação aos *baselines*. Ao utilizar-se diferentes *datasets*, com conteúdos distintos e diferentes critérios de construção, nossa intenção era mostrar que o processo é robusto e que pode ser aplicável a diferentes situações.

Nos experimentos com o Fake.BR, sete das dez combinações entre *datasets* e algoritmos de classificação, resultaram em melhoria na acurácia da classificação. Sendo

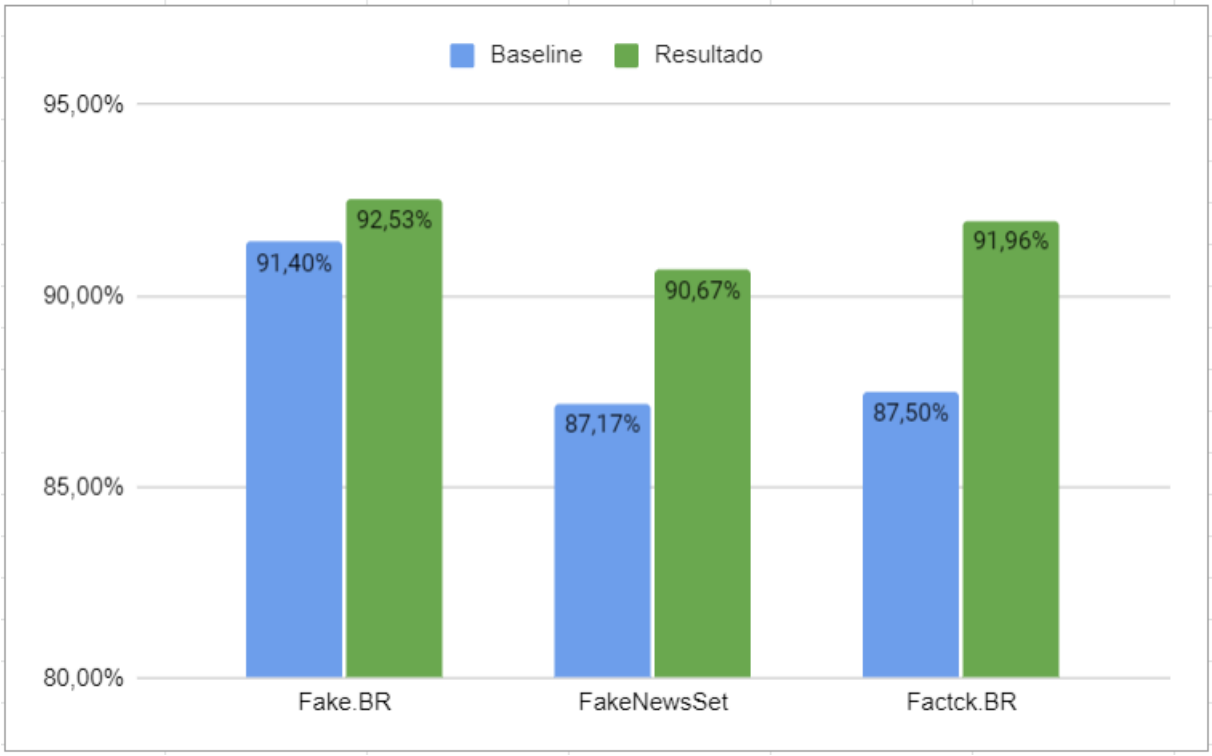


Figura 17 – Melhores resultados em cada Dataset

oito em dez e nove em dez para FakeNewsSet e Factck.BR, respectivamente.

Em todas as situações algum dos algoritmos de *boosting* foi o responsável pelos melhores resultados absolutos. Pode-se notar também, que embora o melhor valor de acurácia tenha sido obtido com o Fake.BR (92,53%), obteve-se uma expressiva melhora com a inclusão dos atributos de emoções nos outros *datasets* (3,50 p.p. para o FakeNewsSeet e 4,46 p.p. para Factck.BR como pode ser visto na Figura 17.

6 CONCLUSÃO

O problema das *Fake News*, embora não seja um fenômeno recente, tem atraído atenção da academia e da sociedade em geral. Todos os dias são mostradas evidências do poder devastador que pode ter uma notícia falsa. Revoltas na população, rejeição de governos, reputações destruídas, fortunas perdidas ou que mudam de mãos.

Os meios digitais fomentaram o acesso à informação e, se por um lado, democratizou o acesso, por outro, permitiu que notícias falsas se proliferassem de maneira muito rápida e atingissem um grande número de pessoas.

Este cenário levou ao desenvolvimento de abordagens para detecção de *Fake News* em várias frentes de atuação de forma a abranger diferentes aspectos para solução do problema e que permitissem sua detecção de forma rápida e confiável.

As abordagens no campo do processamento de linguagem natural, ou linguísticas, destacam-se por utilizarem informações que podem ser extraídas diretamente do texto da notícia, permitindo que se possa inferir se uma notícia é falsa ou não, sem que se necessite de outras informações além da própria noticia. Estudos em psicologia comportamental e linguística apontam existir evidências que fundamentam esta abordagem. Tais estudos apontam ser possível detectar intenções, sentimentos e emoções diferentes, relacionados aos objetivos determinado texto ou narrativa.

Baseados nessas abordagens, foram desenvolvidos métodos que identificam características linguísticas a partir do levantamento de atributos, entre eles a classificação gramatical das palavras dos textos, os relacionamentos entre estes atributos, as frequências como alguns elementos textuais são utilizados, ou até mesmo a análise dos sentimentos presentes na escrita das notícias em diferentes idiomas, mas mais especificamente na língua portuguesa.

Entretanto, até onde foi possível observar, a análise de sentimentos presente nesses trabalhos se limita a utilização da polaridade, sendo ela positiva, negativa ou neutra em relação a um assunto específico. Porém o campo da análise de sentimentos é mais amplo e permite a utilização de outras técnicas. Tendo como base essa limitação, este estudo levantou a hipótese de que a ampliação do uso de técnicas de análise de sentimentos, em particular com a inclusão da classificação de emoções, associadas a classificação gramatical dos textos das notícias pode viabilizar a construção de modelos de detecção de *Fake News* em língua portuguesa, mais robustos que os existentes na literatura.

Durante esse trabalho, foram identificadas ferramentas que permitem classificar emoções humanas e transformar essa informação em elementos que possam ser processados por algoritmos. Destacaram-se nesta pesquisa o LIWC e o Affect-br, com grande valor e

utilidade para esta tarefa.

Desta forma, este estudo propôs e aplicou um método estendido, onde os resultados obtidos com os experimentos realizados apresentaram evidências que apontam para a validade da hipótese levantada. Com este método foi possível realizar experimentos nos quais a inclusão dos atributos de emoção dos textos permitiu melhorar os resultados de acurácia dos modelos de aprendizado de máquina, na tarefa de classificação das *Fake News*.

Toda a pesquisa que foi realizada, assim como os experimentos executados e o resultados apresentados levam às seguintes contribuições:

Apresentação de um método que levanta atributos dos textos de notícias e promove a sua classificação em *fake* ou não *fake*. Este método tem a característica de ser flexível e adaptável, permitindo que as ferramentas utilizadas em sua implementação possam variar segundo a escolha do analista ou o surgimento de novas tecnologias. Além disso podem ser utilizados variados algoritmos de classificação ou mesmo a composição de comitês para a realização desta tarefa.

Comprovou que a classificação de emoções é um importante recurso para análise de estilos de escrita e para a identificação de intensões do texto e por conseguinte que pode ser usado para a identificação de *Fake News*, expandindo o senso comum de que análise de sentimentos se resume a classificação de polaridade.

Construiu um modelo funcional para classificação de *Fake News*, com baixa necessidade de recursos computacionais, que implementou o método proposto e que pode ser utilizado em ambiente produtivo para esta tarefa. Podendo inclusive ser parte integrante, como complemento em outros sistemas de detecção.

O presente trabalho estudou características globais dos textos analisados e, embora obtendo êxito na tarefa de classificação, não levou em consideração possíveis variações ao longo do texto. Não foram analisadas a alternância de atributos ao longo de frases e parágrafos ou entre estes elementos. Por se basear em análise linguística, o presente método apresenta limitação quando ao tamanho do texto utilizado no processo de detecção, sendo necessário textos maiores do que simples comentários de poucas palavras. Além disso, não leva em consideração de forma explicita, palavras fora do dicionário, abreviações ou símbolos.

Como trabalhos futuros, destaca-se a aplicação de algoritmos de classificação mais complexos, capazes de trabalhar com grande quantidade de dimensões de atributos e complexas fronteiras de separação, tais como o uso de redes neurais e *deep learning*. Adicionalmente, pode-se realizar a classificação de emoções em bases temporais, considerando a ordem que elas aparecem nos textos, além de considerar subdivisões menores dos textos, como frases, sentenças ou parágrafos. Nota-se também, espaço para a identificação ou criação de métodos dinâmicos de classificação de emoções, que atuem além da utilização

de simples léxicos, possivelmente com a criação de vetores de emoções ou outros tipos de ferramentas de identificação e classificação das emoções.

REFERÊNCIAS

1 BONDIELLI, A.; MARCELLONI, F. A survey on fake news and rumour detection techniques. *Information Sciences*, Elsevier Inc., v. 497, p. 38–55, 2019. ISSN 00200255.

2 RUSSELL, J. A. A circumplex model of affect. *Journal of personality and social psychology*, American Psychological Association, v. 39, n. 6, p. 1161, 1980.

3 SHU, K.; SLIVA, A.; WANG, S.; TANG, J.; LIU, H. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 19, n. 1, p. 22–36, set. 2017. ISSN 1931-0145. 2019. Disponível em: <http://doi.acm.org/10.1145/3137597.3137600>.

4 FREIRE, P. M. S.; GOLDSCHMIDT, R. R. Uma introdução ao combate automático às fake news em redes sociais virtuais. In: *Tópicos de Gerenciamento de Dados e Informação*. Fortaleza, CE, Brazil: SBC, 2019. (34th SBBD), p. 38–67. ISSN 2016-5170. Disponível em: <http://http://sbbd.org.br/2019/proceedings/>.

5 MEJOVA, Y.; KALIMERI, K. Advertisers jump on coronavirus bandwagon: Politics, news, and business. *ArXiv*, abs/2003.00923, 2020.

6 VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1146–1151, 2018. ISSN 0036-8075. Nov/2019. Disponível em: <https://science.sciencemag.org/content/359/6380/1146>.

7 CONROY, N. J.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, v. 52, n. 1, p. 1–4, 2015. ISSN 23739231.

8 AJAO, O.; BHOWMIK, D.; ZARGARI, S. Sentiment Aware Fake News Detection on Online Social Networks. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. v. 2019-May, p. 2507–2511. ISBN 9781479981311. ISSN 15206149.

9 MORAES, M. P.; de Oliveira Sampaio, J.; CHARLES, A. C. Data mining applied in fake news classification through textual patterns. In: *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web - WebMedia '19*. New York, New York, USA: ACM Press, 2019. p. 321–324. ISBN 9781450367639. Disponível em: <http://dl.acm.org/citation.cfm?doid=3323503.3360648>.

10 MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, Faculty of Engineering, Ain Shams University, v. 5, n. 4, p. 1093–1113, 2014. ISSN 20904479. Disponível em: <http://dx.doi.org/10.1016/j.asej.2014.04.011>.

11 TAUSCZIK, Y. R.; PENNEBAKER, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, v. 29, n. 1, p. 24–54, 2010. Disponível em: <http://jls.sagepub.com>.

12 NEWMAN, M. L.; PENNEBAKER, J. W.; BERRY, D. S.; RICHARDS, J. M. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, v. 29, n. 5, p. 665–675, 2003. PMID: 15272998. Disponível em: <https://doi.org/10.1177/0146167203029005010>.

13 FALLIS, D. A functional analysis of disinformation. In: . [S.l.: s.n.], 2014.

14 RUBIN, V. L.; CHEN, Y.; CONROY, N. J. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, v. 52, n. 1, p. 1–4, 2015. ISSN 23739231.

15 CONROY, N. J.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, v. 52, n. 1, p. 1–4, 2015. 2019. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082>.

16 WU, L.; LIU, H. 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*. New York, NY, USA: ACM, 2018. (WSDM '18), p. 637–645. ISBN 978-1-4503-5581-0. Disponível em: <http://doi.acm.org/10.1145/3159652.3159677>.

17 LIU, Y.; BROOKWU, Y. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: *AAAI Con on Artificial Intelligence*. [S.l.: s.n.], 2018. p. 354–361.

18 KNAPP, M. L.; HART, R. P.; DENNIS, H. S. An Exploration of Deception as a Communication Construct. *Human Communication Research*, v. 1, n. 1, p. 15–29, 03 2006. ISSN 0360-3989. Disponível em: <https://doi.org/10.1111/j.1468-2958.1974.tb00250.x>.

19 DEPAULO, B.; KASHY, D.; KIRKENDOL, S.; WYER, M.; EPSTEIN, J. Lying in everyday life. *Journal of personality and social psychology*, v. 70, p. 979–95, 06 1996.

20 RICHARDS, J.; GROSS, J. Emotion regulation and memory: the cognitive costs of keeping one's cool. *Journal of personality and social psychology*, v. 79 3, p. 410–24, 2000.

21 PINTO, A.; OLIVEIRA, H. G.; ALVES, A. O. Comparing the performance of different NLP toolkits in formal and social media text. *OpenAccess Series in Informatics*, v. 51, n. 3, p. 31–316, 2016. ISSN 21906807.

22 JURAFSKY, D.; MARTIN, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. [S.l.: s.n.], 2008. v. 2.

23 VIEIRA, R.; LIMA, V. L. S. Lingüística computacional: princípios e aplicações. *I Jornada de Atualização em Inteligência Artificial*, p. 47–86, 2001. Disponível em: <http://www.inf.unioeste.br/{~}jorge/MESTRADOS/LETRAS-MECANISMOSDOFUNCIONAMENTODALINGUAGEM-PROCESSAMENTODALINGUARTIGOSINTERESSANTES/lingu?sticacomputacional.p>.

24 NETO, J. M. d. O.; TONIN, S. D.; PRIETCH, S. S. Processamento de Linguagem Natural e suas Aplicacoes Computacionais. *Escola Regional de Informatica (ERIN)*, 2010. Disponível em: <https://drive.google.com/drive/folders/0B9rS4hK3zYsQdFUtMm9QUVJnMGs>.

25 Cambria, E.; White, B. Jumping nlp curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine*, v. 9, n. 2, p. 48–57, 2014.

26 MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2008. <https://nlp.stanford.edu/IR-book/>.

27 JURAFSKY, D.; MARTIN, J. H. Ch. 8: Part-of-speech tagging. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, v. 1, 2019.

28 PETROV, S.; DAS, D.; MCDONALD, R. A universal part-of-speech tagset. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. p. 2089–2096. 2019. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf>.

29 DAS, D.; MCDONALD, R. A Universal Part-of-Speech Tagset. 2009.

30 AKIRA, V.; GONçALVES, B.; COSTA, B.; SILVA, J. Modelos de predição estruturada em part-of-speech tagging para português do brasil. In: -. [s.n.], 2014. 2019. Disponível em: <http://>.

31 KLEIN, S.; SIMMONS, R. F. A computational approach to grammatical coding of english words. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 10, n. 3, p. 334–347, 1963.

32 GREENE, B. B.; RUBIN, G. M. *Automatic grammatical tagging of English*. [S.l.]: Department of Linguistics, Brown University, 1971.

33 KARLSSON, F. Designing a parser for unrestricted text. *Karlsson et al*, v. 1995, p. 1–40, 1995.

34 CHURCH, K. W. A stochastic parts program and noun phrase parser for unrestricted text. In: *Second Conference on Applied Natural Language Processing*. Austin, Texas, USA: Association for Computational Linguistics, 1988. p. 136–143. Disponível em: <https://www.aclweb.org/anthology/A88-1019>.

35 SCHMID, H. Part-of-speech tagging with neural networks. *arXiv preprint cmp-lg/9410018*, 1994.

36 RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: *Conference on empirical methods in natural language processing*. [S.l.: s.n.], 1996.

37 DAELEMANS, W.; ZAVREL, J.; BERCK, P.; GILLIS, S. Mbt: A memory-based part of speech tagger-generator. *arXiv preprint cmp-lg/9607012*, 1996.

38 BOYE, K.; BASTIAANSE, R. Grammatical versus lexical words in theory and aphasia: Integrating linguistics and neurolinguistics. *Glossa: a journal of general linguistics*, v. 3, n. 1, p. 29, 2018. ISSN 2397-1835.

39 LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, v. 5, n. 1, p. 1–167, 2012. Disponível em: <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>.

40 MÄNTYLÄ, M. V.; GRAZIOTIN, D.; KUUTILA, M. *The evolution of sentiment analysis—A review of research topics, venues, and top cited papers*. 2018. 16–32 p.

41 LIU, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. [S.l.]: Cambridge University Press, 2015.

42 MÁ, P. C. e.; LVA, R. J. . S. sentilex-pt: principais características e potencialidades. In: . [s.n.]. ISBN 978-82-91398-12-9. ISSN 1890-9639. Disponível em: <http://www.journals.uio.no/osla>.

43 BUECHEL, S.; HAHN, U. Emotion Representation Mapping for Automatic Lexicon Construction (Mostly) Performs on Human Level. p. 2892–2904, 2018. Disponível em: <http://arxiv.org/abs/1806.08890>.

44 MOHAMMAD, S. M. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *Emotion Measurement*, p. 201–237, 2016.

45 ACHEAMPONG, F. A.; WENYU, C.; NUNOO-MENSAH, H. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, Wiley, v. 2, n. 7, jul 2020. ISSN 2577-8196. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.12189>.

46 EKMAN, P. Facial expressions of emotion: New findings, new questions. *Psychological Science*, v. 3, n. 1, p. 34–38, 1992. Disponível em: <https://doi.org/10.1111/j.1467-9280.1992.tb00253.x>.

47 PLUTCHIK, R. *The emotions*. [S.l.]: University Press of America, 1991.

48 WATSON, D.; CLARK, L. A.; TELLEGEN, A. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, v. 54, n. 6, p. 1063–1070, 1988. ISSN 00223514.

49 CARVALHO, F.; SANTOS, G.; GUEDES, G. P. AffectPT-br: An Affective Lexicon based on LIWC 2015. *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, v. 2018-Novem, 2018. ISSN 15224902.

50 Weiss, E. A. Biographies: Eloge: Arthur lee samuel (1901-90). *IEEE Annals of the History of Computing*, v. 14, n. 3, p. 55–69, 1992.

51 BAHARUDIN, B.; LEE, L. H.; KHAN, K.; KHAN, A. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, v. 1, 02 2010.

52 FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, n. 1, p. 119 – 139, 1997. ISSN 0022-0000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002200009791504X>.

53 FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, The Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 10 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>.

54 BHUTANI, B.; RASTOGI, N.; SEHGAL, P.; PURWAR, A. Fake News Detection Using Sentiment Analysis. In: . [S.l.]: Institute of Electrical and Electronics Engineers (IEEE), 2019. p. 1–5. ISBN 9781728135915.

55 MORAIS, J. I. de; ABONIZIO, H. Q.; TAVARES, G. M.; FONSECA, A. A. da; BARBON, S. Deciding among Fake, Satirical, Objective and Legitimate news. In: . [S.l.]: Association for Computing Machinery (ACM), 2019. p. 1–8.

56 DURIER, F.; GARCIA, A. C. Fake News and Sarcasm, what is the limit of a critic and what is intentionally fake? In: *Anais do Simpósio Brasileiro de Sistemas Colaborativos*. Sociedade Brasileira de Computação - SBC, 2019. p. 58–61. Disponível em: <https://sol.sbc.org.br/index.php/sbsc/article/view/7807>.

57 Faustini, P.; Covões, T. Fake news detection using one-class classification. *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, p. 592–597, 2019.

58 SHU, K.; CUI, L.; WANG, S.; LEE, D.; LIU, H. dEFENDD: Explainable Fake News Detection. p. 395–405, 2019.

59 AWRAHMAN, B.; ALATAS, B. Sentiment analysis and opinion mining within social networks using Konstanz information miner. *Journal of Telecommunication, Electronic and Computer Engineering*, v. 9, n. 1, p. 15–22, 2017. ISSN 22898131.

60 CARVALHO, F.; RODRIGUES, R. G.; SANTOS, G.; CRUZ, P.; FERRARI, L.; GUEDES, G. P. Evaluating the brazilian portuguese version of the 2015 liwc lexicon with sentiment analysis in social networks. In: SBC. *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2019. p. 24–34.

61 MONTEIRO, R. A.; SANTOS, R. L.; PARDO, T. A.; ALMEIDA, T. A. de; RUIZ, E. E.; VALE, O. A. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [s.n.], 2018. v. 11122 LNAI, p. 324–334. ISBN 9783319997216. ISSN 16113349. Disponível em: <https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit->.

62 MORENO, J.; BRESSAN, G. FACTCK.BR. In: *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web - WebMedia '19*. New York, New York, USA: ACM Press, 2019. p. 525–527. ISBN 9781450367639. Disponível em: <http://dl.acm.org/citation.cfm?doid=3323503.3361698>.

63 SILVA, F. R. M. da; FREIRE, P. M. S.; SOUZA, M. P. de; PLENAMENTE, G. de A. B.; GOLDSCHMIDT, R. R. Fakenewssetgen: A process to build datasets that support comparison among fake news detection methods. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: Association for Computing Machinery, 2020. (WebMedia '20), p. 241–248. ISBN 9781450381963. Disponível em: <https://doi.org/10.1145/3428658.3430965>.

64 CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. ISSN 10769757.

APÊNDICE A – MATRIZ PARA SELEÇÃO DE ARTIGOS

Autor	Título	Que informações utiliza na identificação?	Abordagem Utilizada	Dataset	Considera identificação de Sentimentos?	Considera construção semantica dos textos?	Considera o texto como um todo ou separa em sentenças?	Depende de tópico específico?	Analiza o texto de meses ou somente a notícia?	Pode ser utilizado em outras línguas?	Testado em Portugês	Verifica a complexidade do texto analisado?	Utilizam estatísticas globais dos textos?	Utilizam redes neurais?	
De Morais J.I., Abonizio H.Q., Tavares	Deciding among fake, satirical, objective and le	Texto de Notícias	Classificar distintamente os textos em objective/satirical and legitimate/fake	Textos de sites de notícias	N	N	Separa	Política	Notícias	N	S	S	N	B2	
Monteiro R.A., Santos R.L.S., Pardo T.A.S., de Almeida T.A., Ruiz E.E.S., Vale O.A.	Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results	Texto	Criação de dataset em português para detecção de fake news.	Fake.br	N	N	Todo	N	Notícia	N	N	S	N		
Moreno, João Bressan, Graça	FACTCK.BR: a new dataset to study fake news	Artigos da agências de factcheck que usam o modelo do ClaimReview.	Criação de dataset em português para detecção de fake news.	Aos fatos Lupa Truco	N	N	N	N	N	N	N	N	N		
CARDOSO DURIER DA SILVA, Fernando; BICHARRA GARCIA, Ana Cristina.	Fake News and Sarcasm, what is the limit of a critic and what is intentionally fake?.	Artigos	Uso de redes neurais para diferenciar sarcasmo de notícias falsas.	Sites de Notícias, Sites de FakeNews e Sites de Sarcasmo	S	S		Política	Notícia			S	X		
Marcos Paulo Moraes Jonice de Oliveira Sampaio Anderson Cordeiro Charles	Data mining applied in fake news classification	Texto	Uso de NLP e ML para identificar padrões de linguagem usados em fake news.	Fakepedia Fake.br	S	N	Texto Todo		Notícia		S	S	N		

Tabela 16 – Matriz de perguntas para selecão de trabalhos relacionados

APÊNDICE B – RESULTADOS COMPLETOS DOS EXPERIMENTOS

	GRAMÁTICA SPACY			GRAMÁTICA LIWC			GRAMÁTICA SPACY + POLARIDADE			GRAMÁTICA LIWC + POLARIDADE			GRAMÁTICA LIWC + POLARI + MAIÚSCULAS			GRAMÁTICA LIWC + POLARI + MAIÚSCULAS + EMOÇÕES		
	F-measur	Accurac	Desv. Pad	F-measure	Accurac	Desv. Pad	F-measure	Accuracy	Std. deviation	F-measure	Accuracy	Desv. Padr	F-measur	Accurac	Desv. Padr	F-measure	Accurac	Desv. Pad
NAIVE BAYES																		
VERDADEIRO	65,98%			73,08%			75,93%			81,30%			84,92%			85,40%		
FALSO	76,17%			77,38%			82,02%			84,45%			86,91%			87,18%		
Overall		71,97%	10,81		75,42%	8,67		79,42%	946,10%		83,10%	14,49		85,99%	11,30		86,35%	12,96
GRADIENT BOOST																		
VERDADEIRO	88,27%			88,54%			90,84%			91,30%			91,69%			92,34%		
FALSO	88,48%			88,60%			90,94%			91,50%			91,84%			92,46%		
Overall		**88,38%**	7,06		**88,57%**	4,45		**90,89%**	634,56%		**91,40%**	6,30		**91,76%**	7,97		**92,40%**	6,17
ADABOOST																		
VERDADEIRO	83,40%			85,02%			87,83%			90,90%			89,27%			90,40%		
FALSO	84,27%			85,76%			88,33%			91,00%			89,78%			90,55%		
Overall		83,85%	6,73		85,40%	8,08		88,08%	927,12%		90,90%	6,91		89,53%	8,33		90,47%	7,24
SMV																		
VERDADEIRO	85,85%			76,97%			87,77%			83,20%			77,18%			77,18%		
FALSO	86,89%			79,86%			88,41%			84,80%			80,04%			80,14%		
Overall		86,39%	7,30		78,51%	8,19		88,10%	912,93%		84,10%	9,17		78,71%	9,67		78,76%	9,53
KNN																		
FALSO	70,78%			73,45%			73,73%			76,30%			76,81%			75,22%		
VERDADEIRO	79,80%			80,87%			80,78%			82,80%			82,29%			81,76%		
Overall		76,11%	12,00		77,76%	6,94		77,81%	11,38		80,10%	10,78		79,92%	6,15		78,99%	6,04

Tabela 17 – Resultados incrementais com Fake.BR - parte 1

	GRAM LIWC + POL + MAIUSC EMOÇÃO AFFECT-BR			GRAM SPACY + POL + EMO LIWC			GRAM SPACY+ POL EMOÇÃO AFFECT-BR			TODOS OS ATRIBUTOS LIWC + POLARIDADE			GRAM. LIWC + POL + EMOÇÃO LIWC			GRAM. LIWC + POL + EMOÇÃO AFFECT-BR		
	F-measure	Accurac	Desv. Pad	F-measure	Accuracy	Desv. Padr	F-measure	Accuracy	Desv. Padr	F-measur	Accuracy	Desv. Pad	F-measure	Accuracy	Desv. Padr			
NAIVE BAYES																		
VERDADEIRO	83,09%			78,60%			78,95%			84,33%			82,50%			82,80%		
FALSO	85,62%			83,34%			83,61%			86,63%			85,20%			85,60%		
Overall		84,46%	11,27		81,26%	7,88		81,57%	10,04		85,57%	9,92		84,00%	10,78		84,30%	10,79
GRADIENT BOOST																		
VERDADEIRO	92,50%			92,09%			91,87%			94,00%			92,30%			92,50%		
FALSO	92,56%			92,16%			91,93%			94,03%			92,40%			92,60%		
Overall		**92,53%**	6,43		**92,13%**	5,95		**91,90%**	5,62		**94,01%**	5,20		**92,30%**	5,01		**92,50%**	6,42
ADABOOST																		
VERDADEIRO	89,85%			89,23%			89,19%			92,24%			92,10%			91,80%		
FALSO	89,90%			89,27%			89,26%			92,23%			92,10%			91,80%		
Overall		89,88%	4,61		89,25%	5,25		89,22%	6,33		92,24%	5,04		92,10%	4,19		91,80%	7,21
SMV																		
VERDADEIRO	82,42%			87,82%			87,71%			84,86%			83,30%			83,20%		
FALSO	84,16%			88,47%			88,38%			86,29%			84,90%			84,80%		
Overall		83,33%	9,80		88,15%	8,39		88,06%	7,62		85,61%	7,00		84,20%	9,30		84,00%	7,77
KNN																		
FALSO	77,63%			72,46%			72,03%			64,45%			70,10%			69,00%		
VERDADEIRO	83,17%			80,68%			80,37%			78,56%			80,50%			79,90%		
Overall		80,79%	8,59		77,29%	12,94		76,93%	9,73		73,25%	9,42		76,40%	12,45		75,60%	13,80

Tabela 18 – Resultados incrementais com Fake.BR - parte 2

NAIVE BAYES									
	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
VERDADEIRO	226	3374	1123	68,81%	91,64%	68,81%	93,72%	78,60%	
FALSO	1123	2477	226	93,72%	75,03%	93,72%	68,81%	83,34%	
Overall									81,26%
GRADIENT BOOST									
VERDADEIRO	267	3333	300	91,67%	92,51%	91,67%	92,58%	92,09%	
FALSO	300	3300	267	92,58%	91,74%	92,58%	91,67%	92,16%	
Overall									92,13%
ADABOOST									
VERDADEIRO	381	3219	393	89,08%	89,38%	89,08%	89,42%	89,23%	
FALSO	393	3207	381	89,42%	89,12%	89,42%	89,08%	89,27%	
Overall									89,25%
SMV									
VERDADEIRO	329	3271	524	85,44%	90,34%	85,44%	90,86%	87,82%	
FALSO	524	3076	329	90,86%	86,19%	90,86%	85,44%	88,47%	
Overall									88,15%
KNN									
FALSO	186	3414	1449	59,75%	92,04%	59,75%	94,83%	72,46%	
VERDADEIRO	1449	2151	186	94,83%	70,20%	94,83%	59,75%	80,68%	
Overall									77,29%

Tabela 19 – Fake.BR - FNE(SpaCy,FNE-CSR,Sentilex-PT,LIWC)

NAIVE BAYES										
	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
VERDADEIRO	2489	216	3384	1111	69,14%	92,01%	69,14%	94,00%	78,95%	
FALSO	3384	1111	2489	216	94,00%	75,28%	94,00%	69,14%	83,61%	
Overall										81,57%
GRADIENT BOOST										
VERDADEIRO	3295	278	3322	305	91,53%	92,22%	91,53%	92,28%	91,87%	
FALSO	3322	305	3295	278	92,28%	91,59%	92,28%	91,53%	91,93%	
Overall										91,90%
ADABOOST										
VERDADEIRO	3200	376	3224	400	88,89%	89,49%	88,89%	89,56%	89,19%	
FALSO	3224	400	3200	376	89,56%	88,96%	89,56%	88,89%	89,26%	
Overall										89,22%
SMV										
VERDADEIRO	3069	329	3271	531	85,25%	90,32%	85,25%	90,86%	87,71%	
FALSO	3271	531	3069	329	90,86%	86,03%	90,86%	85,25%	88,38%	
Overall										88,06%
KNN										
FALSO	2139	200	3400	1461	59,42%	91,45%	59,42%	94,44%	72,03%	
VERDADEIRO	3400	1461	2139	200	94,44%	69,94%	94,44%	59,42%	80,37%	
Overall										76,93%

Tabela 20 – Fake.BR - FNE(SpaCy,FNE-CSR,Sentilex-PT,Affect-BR)

NAIVE BAYES	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
VERDADEIRO	2874	257	3343	726	79,83%	91,79%	79,83%	92,86%	85,40%	
FALSO	3343	726	2874	257	92,86%	82,16%	92,86%	79,83%	87,18%	
Overall										86,35%
GRADIENT BOOST										
VERDADEIRO	3299	246	3354	301	91,64%	93,06%	91,64%	93,17%	92,34%	
FALSO	3354	301	3299	246	93,17%	91,76%	93,17%	91,64%	92,46%	
Overall										92,40%
ADABOOST										
VERDADEIRO	3229	315	3285	371	89,69%	91,11%	89,69%	91,25%	90,40%	
FALSO	3285	371	3229	315	91,25%	89,85%	91,25%	89,69%	90,55%	
Overall										90,47%
SMV										
VERDADEIRO	2586	515	3085	1014	71,83%	83,39%	71,83%	85,69%	77,18%	
FALSO	3085	1014	2586	515	85,69%	75,26%	85,69%	71,83%	80,14%	
Overall										78,76%
KNN										
FALSO	2296	209	3391	1304	63,78%	91,66%	63,78%	94,19%	75,22%	
VERDADEIRO	3391	1304	2296	209	94,19%	72,23%	94,19%	63,78%	81,76%	
Overall										78,99%

Tabela 21 – Fake.BR - FNE(LIWC,FNE-CSR,Sentilex-PT,LIWC)

NAIVE BAYES									
	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
VERDADEIRO	2749	268	3332	851	76,36%	91,12%	76,36%	92,56%	83,09%	
FALSO	3332	851	2749	268	92,56%	79,66%	92,56%	76,36%	85,62%	
Overall										84,46%
GRADIENT BOOST										
VERDADEIRO	3316	254	3346	284	92,11%	92,89%	92,11%	92,94%	92,50%	
FALSO	3346	284	3316	254	92,94%	92,18%	92,94%	92,11%	92,56%	
Overall										92,53%
ADABOOST										
VERDADEIRO	3226	355	3245	374	89,61%	90,09%	89,61%	90,14%	89,85%	
FALSO	3245	374	3226	355	90,14%	89,67%	90,14%	89,61%	89,90%	
Overall										89,88%
SMV										
VERDADEIRO	2813	413	3187	787	78,14%	87,20%	78,14%	88,53%	82,42%	
FALSO	3187	787	2813	413	88,53%	80,20%	88,53%	78,14%	84,16%	
Overall										83,33%
KNN										
VERDADEIRO	3417	1200	2400	183	94,92%	74,01%	94,92%	66,67%	83,17%	
FALSO	2400	183	3417	1200	66,67%	92,92%	66,67%	94,92%	77,63%	
Overall										80,79%

Tabela 22 – Fake.BR - FNE(LIWC,FNE-CSR,Sentilex-PT,Affect-BR)

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	10,00%	88,24%	10,00%	98,67%	17,96%	
	VERDADEIRO	98,67%	52,30%	98,67%	10,00%	68,36%	
	Overall						54,33%
GRADIENT BOOST	FALSO	91,67%	87,86%	91,67%	87,33%	89,72%	
	VERDADEIRO	87,33%	91,29%	87,33%	91,67%	89,27%	
	Overall						89,50%
ADA BOOST	FALSO	90,00%	88,82%	90,00%	88,67%	89,40%	
	VERDADEIRO	88,67%	89,86%	88,67%	90,00%	89,26%	
	Overall						89,33%
SVM	FALSO	88,67%	80,12%	88,67%	78,00%	84,18%	
	VERDADEIRO	78,00%	87,31%	78,00%	88,67%	82,39%	
	Overall						83,33%
KNN	FALSO	78,33%	66,20%	78,33%	60,00%	71,76%	
	VERDADEIRO	60,00%	73,47%	60,00%	78,33%	66,06%	
	Overall						69,17%

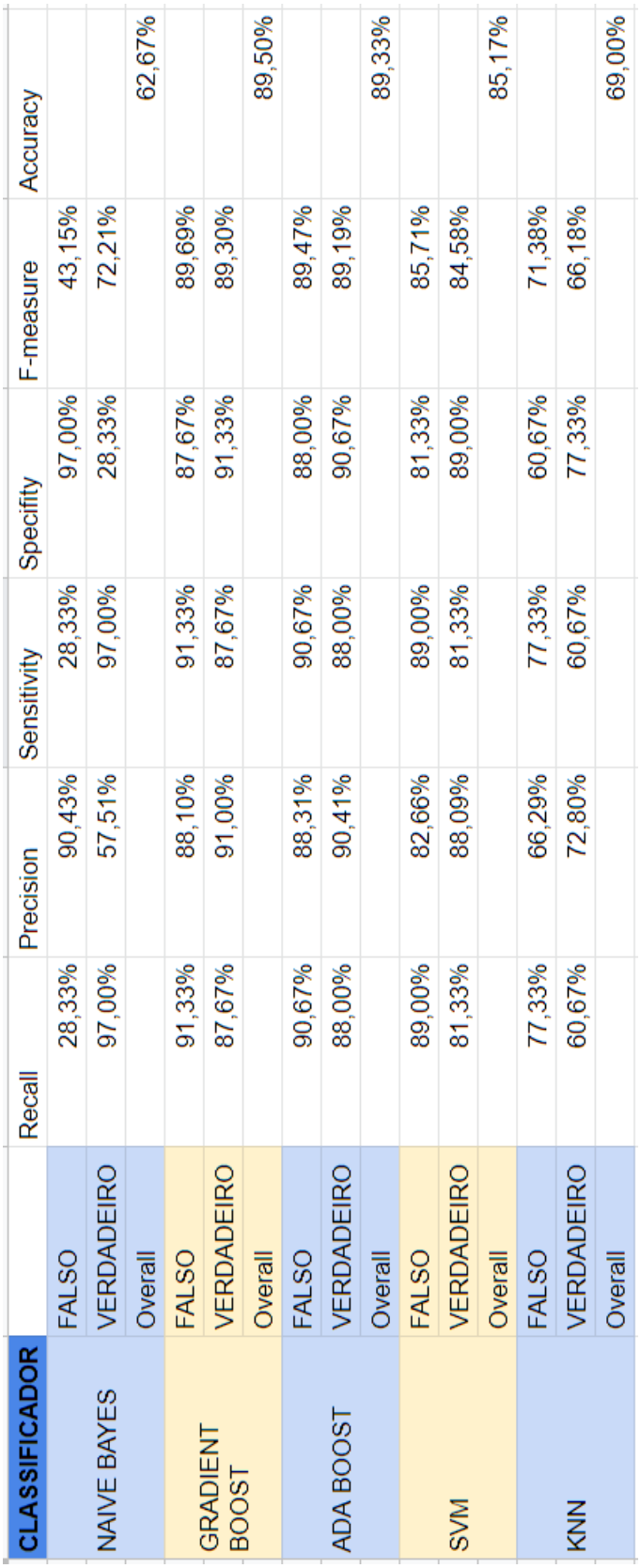
Tabela 23 – FakeNewsSet Baseline 1

CLASSIFICADOR	Row ID	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	50,00%	83,80%	50,00%	90,33%	62,63%	
	VERDADEIRO	90,33%	64,37%	90,33%	50,00%	75,17%	
	Overall						70,17%
GRADIENT BOOST	FALSO	88,00%	85,16%	88,00%	84,67%	86,56%	
	VERDADEIRO	84,67%	87,59%	84,67%	88,00%	86,10%	
	Overall						86,33%
ADA BOOST	FALSO	90,00%	85,17%	90,00%	84,33%	87,52%	
	VERDADEIRO	84,33%	89,40%	84,33%	90,00%	86,79%	
	Overall						87,17%
SVM	FALSO	84,33%	79,81%	84,33%	78,67%	82,01%	
	VERDADEIRO	78,67%	83,39%	78,67%	84,33%	80,96%	
	Overall						81,50%
KNN	FALSO	72,00%	70,82%	72,00%	70,33%	71,40%	
	VERDADEIRO	70,33%	71,53%	70,33%	72,00%	70,92%	
	Overall						71,17%

Tabela 24 – FakeNewsSet Baseline 2

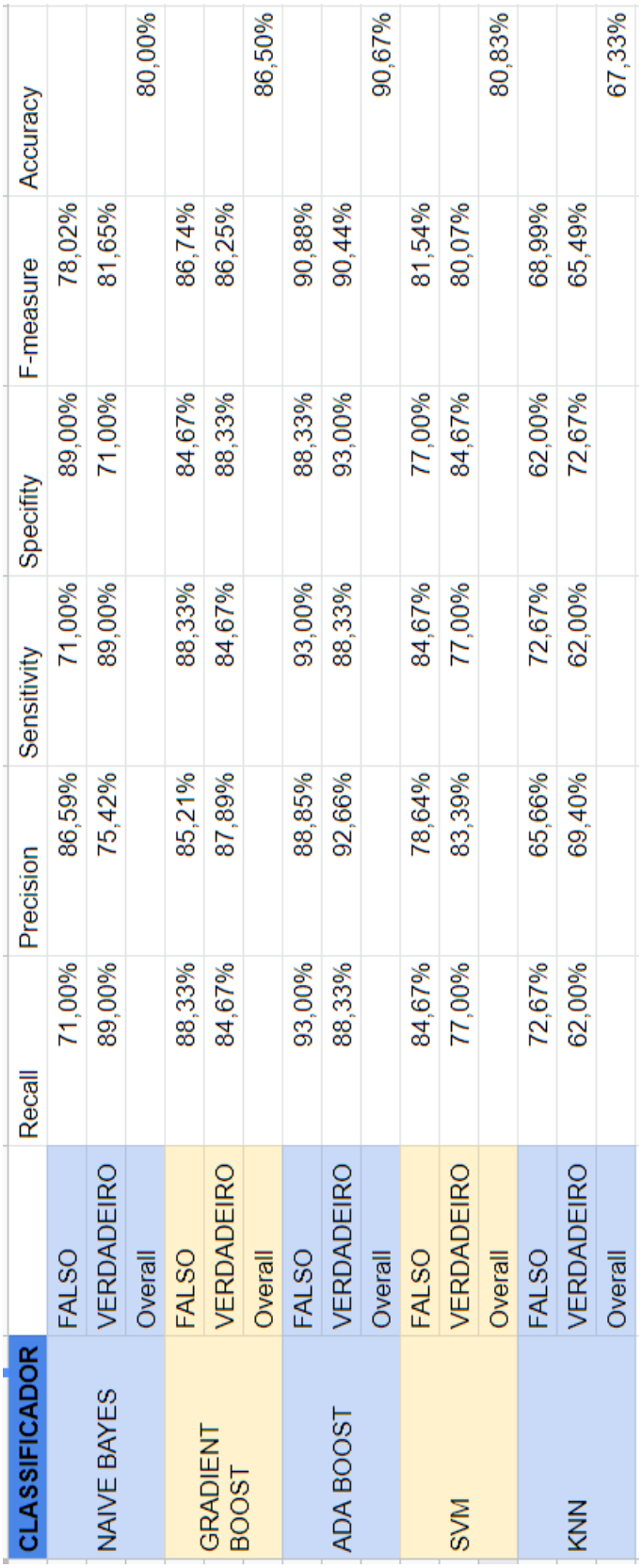
CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	27,67%	91,21%	27,67%	97,33%	42,46%	
	VERDADEIRO	97,33%	57,37%	97,33%	27,67%	72,19%	
	Overall						62,50%
GRADIENT BOOST	FALSO	90,33%	88,27%	90,33%	88,00%	89,29%	
	VERDADEIRO	88,00%	90,10%	88,00%	90,33%	89,04%	
	Overall						89,17%
ADA BOOST	FALSO	90,67%	89,77%	90,67%	89,67%	90,22%	
	VERDADEIRO	89,67%	90,57%	89,67%	90,67%	90,12%	
	Overall						90,17%
SVM	FALSO	89,67%	83,28%	89,67%	82,00%	86,36%	
	VERDADEIRO	82,00%	88,81%	82,00%	89,67%	85,27%	
	Overall						85,83%
KNN	FALSO	78,33%	66,01%	78,33%	59,67%	71,65%	
	VERDADEIRO	59,67%	73,36%	59,67%	78,33%	65,81%	
	Overall						69,00%

Tabela 25 – FakeNewsSet - FNE(SpaCy, FNE-CSR,Sentlex-PT, LIWC)



CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	28,33%	90,43%	28,33%	97,00%	43,15%	
	VERDADEIRO	97,00%	57,51%	97,00%	28,33%	72,21%	
	Overall						62,67%
GRADIENT BOOST	FALSO	91,33%	88,10%	91,33%	87,67%	89,69%	
	VERDADEIRO	87,67%	91,00%	87,67%	91,33%	89,30%	
	Overall						89,50%
ADA BOOST	FALSO	90,67%	88,31%	90,67%	88,00%	89,47%	
	VERDADEIRO	88,00%	90,41%	88,00%	90,67%	89,19%	
	Overall						89,33%
SVM	FALSO	89,00%	82,66%	89,00%	81,33%	85,71%	
	VERDADEIRO	81,33%	88,09%	81,33%	89,00%	84,58%	
	Overall						85,17%
KNN	FALSO	77,33%	66,29%	77,33%	60,67%	71,38%	
	VERDADEIRO	60,67%	72,80%	60,67%	77,33%	66,18%	
	Overall						69,00%

Tabela 26 – FakeNewsSet - FNE(SpaCy, FNE-CSR,Sentlex-PT, Affect-BR)



CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	71,00%	86,59%	71,00%	89,00%	78,02%	
	VERDADEIRO	89,00%	75,42%	89,00%	71,00%	81,65%	
	Overall						80,00%
GRADIENT BOOST	FALSO	88,33%	85,21%	88,33%	84,67%	86,74%	
	VERDADEIRO	84,67%	87,89%	84,67%	88,33%	86,25%	
	Overall						86,50%
ADA BOOST	FALSO	93,00%	88,85%	93,00%	88,33%	90,88%	
	VERDADEIRO	88,33%	92,66%	88,33%	93,00%	90,44%	
	Overall						90,67%
SVM	FALSO	84,67%	78,64%	84,67%	77,00%	81,54%	
	VERDADEIRO	77,00%	83,39%	77,00%	84,67%	80,07%	
	Overall						80,83%
KNN	FALSO	72,67%	65,66%	72,67%	62,00%	68,99%	
	VERDADEIRO	62,00%	69,40%	62,00%	72,67%	65,49%	
	Overall						67,33%

Tabela 27 – FakeNewsSet - FNE(LIWC, FNE-CSR,Sentlex-PT, Affect-BR)

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	67,67%	87,50%	67,67%	90,33%	76,32%	
	VERDADEIRO	90,33%	73,64%	90,33%	67,67%	81,14%	
	Overall						79,00%
GRADIENT BOOST	FALSO	88,00%	86,27%	88,00%	86,00%	87,13%	
	VERDADEIRO	86,00%	87,76%	86,00%	88,00%	86,87%	
	Overall						87,00%
ADA BOOST	FALSO	91,00%	88,64%	91,00%	88,33%	89,80%	
	VERDADEIRO	88,33%	90,75%	88,33%	91,00%	89,53%	
	Overall						89,67%
SVM	FALSO	87,00%	81,31%	87,00%	80,00%	84,06%	
	VERDADEIRO	80,00%	86,02%	80,00%	87,00%	82,90%	
	Overall						83,50%
KNN	FALSO	75,00%	67,77%	75,00%	64,33%	71,20%	
	VERDADEIRO	64,33%	72,01%	64,33%	75,00%	67,96%	
	Overall						69,67%

Tabela 28 – FakeNewsSet - FNE(LIWC, FNE-CSR,Sentlex-PT, LIWC)

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	0,71%	100,00%	0,71%	100,00%	1,42%	
	VERDADEIRO	100,00%	50,18%	100,00%	0,71%	66,83%	
	Overall						50,36%
GRADIENT BOOST	FALSO	83,57%	86,35%	83,57%	86,79%	84,94%	
	VERDADEIRO	86,79%	84,08%	86,79%	83,57%	85,41%	
	Overall						85,18%
ADA BOOST	FALSO	86,07%	88,60%	86,07%	88,93%	87,32%	
	VERDADEIRO	88,93%	86,46%	88,93%	86,07%	87,68%	
	Overall						87,50%
SVM	FALSO	66,07%	74,60%	66,07%	77,50%	70,08%	
	VERDADEIRO	77,50%	69,55%	77,50%	66,07%	73,31%	
	Overall						71,79%
KNN	FALSO	66,43%	85,32%	66,43%	88,57%	74,70%	
	VERDADEIRO	88,57%	72,51%	88,57%	66,43%	79,74%	
	Overall						77,50%

Tabela 29 – Factck.BR Baseline 1

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	14,29%	100,00%	14,29%	100,00%	25,00%	
	VERDADEIRO	100,00%	53,85%	100,00%	14,29%	70,00%	
	Overall						57,14%
GRADIENT BOOST	FALSO	81,79%	77,63%	81,79%	76,43%	79,65%	
	VERDADEIRO	76,43%	80,75%	76,43%	81,79%	78,53%	
	Overall						79,11%
ADA BOOST	FALSO	79,29%	74,75%	79,29%	73,21%	76,95%	
	VERDADEIRO	73,21%	77,95%	73,21%	79,29%	75,51%	
	Overall						76,25%
SVM	FALSO	83,93%	65,64%	83,93%	56,07%	73,67%	
	VERDADEIRO	56,07%	77,72%	56,07%	83,93%	65,15%	
	Overall						70,00%
KNN	FALSO	70,36%	73,23%	70,36%	74,29%	71,77%	
	VERDADEIRO	74,29%	71,48%	74,29%	70,36%	72,85%	
	Overall						72,32%

Tabela 30 – Factck.BR Baseline 2

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	100,00%	50,00%	100,00%	0,00%	66,67%	
	VERDADEIRO	0,00%		0,00%	100,00%		
	Overall						50,00%
GRADIENT BOOST	FALSO	86,79%	87,73%	86,79%	87,86%	87,25%	
	VERDADEIRO	87,86%	86,93%	87,86%	86,79%	87,39%	
	Overall						87,32%
ADA BOOST	FALSO	90,00%	89,68%	90,00%	89,64%	89,84%	
	VERDADEIRO	89,64%	89,96%	89,64%	90,00%	89,80%	
	Overall						89,82%
SVM	FALSO	79,29%	68,94%	79,29%	64,29%	73,75%	
	VERDADEIRO	64,29%	75,63%	64,29%	79,29%	69,50%	
	Overall						71,79%
KNN	FALSO	71,43%	90,09%	71,43%	92,14%	79,68%	
	VERDADEIRO	92,14%	76,33%	92,14%	71,43%	83,50%	
	Overall						81,79%

Tabela 31 – Factck.BR (SpaCy, FNE-CSR,Sentlex-PT, LIWC)

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	1,07%	100,00%	1,07%	100,00%	2,12%	
	VERDADEIRO	100,00%	50,27%	100,00%	1,07%	66,91%	
	Overall						50,54%
GRADIENT BOOST	FALSO	86,43%	90,64%	86,43%	91,07%	88,48%	
	VERDADEIRO	91,07%	87,03%	91,07%	86,43%	89,01%	
	Overall						88,75%
ADA BOOST	FALSO	91,07%	92,73%	91,07%	92,86%	91,89%	
	VERDADEIRO	92,86%	91,23%	92,86%	91,07%	92,04%	
	Overall						91,96%
SVM	FALSO	79,29%	74,50%	79,29%	72,86%	76,82%	
	VERDADEIRO	72,86%	77,86%	72,86%	79,29%	75,28%	
	Overall						76,07%
KNN	FALSO	68,93%	93,69%	68,93%	95,36%	79,42%	
	VERDADEIRO	95,36%	75,42%	95,36%	68,93%	84,23%	
	Overall						82,14%

Tabela 32 – Factck.BR (SpaCy, FNE-CSR,Sentlex-PT, Affect-BR)

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	0,71%	100,00%	0,71%	100,00%	1,42%	
	VERDADEIRO	100,00%	50,18%	100,00%	0,71%	66,83%	
	Overall						50,36%
GRADIENT BOOST	FALSO	83,93%	81,88%	83,93%	81,43%	82,89%	
	VERDADEIRO	81,43%	83,52%	81,43%	83,93%	82,46%	
	Overall						82,68%
ADA BOOST	FALSO	80,71%	80,71%	80,71%	80,71%	80,71%	
	VERDADEIRO	80,71%	80,71%	80,71%	80,71%	80,71%	
	Overall						80,71%
SVM	FALSO	80,00%	66,67%	80,00%	60,00%	72,73%	
	VERDADEIRO	60,00%	75,00%	60,00%	80,00%	66,67%	
	Overall						70,00%
KNN	FALSO	67,50%	77,78%	67,50%	80,71%	72,28%	
	VERDADEIRO	80,71%	71,29%	80,71%	67,50%	75,71%	
	Overall						74,11%

Tabela 33 – Factck.BR (LIWC, FNE-CSR,Sentlex-PT, Affect-BR)

CLASSIFICADOR		Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
NAIVE BAYES	FALSO	1,43%	100,00%	1,43%	100,00%	2,82%	
	VERDADEIRO	100,00%	50,36%	100,00%	1,43%	66,99%	
	Overall						50,71%
GRADIENT BOOST	FALSO	81,43%	81,43%	81,43%	81,43%	81,43%	
	VERDADEIRO	81,43%	81,43%	81,43%	81,43%	81,43%	
	Overall						81,43%
ADA BOOST	FALSO	82,86%	80,28%	82,86%	79,64%	81,55%	
	VERDADEIRO	79,64%	82,29%	79,64%	82,86%	80,94%	
	Overall						81,25%
SVM	FALSO	57,86%	76,42%	57,86%	82,14%	65,85%	
	VERDADEIRO	82,14%	66,09%	82,14%	57,86%	73,25%	
	Overall						70,00%
KNN	FALSO	66,79%	71,10%	66,79%	72,86%	68,88%	
	VERDADEIRO	72,86%	68,69%	72,86%	66,79%	70,71%	
	Overall						69,82%

Tabela 34 – Factck.BR (LIWC, FNE-CSR,Sentlex-PT, LIWC)