Pós-Graduação em Ciência da Computação

Paulo de Assis Nascimento

Aplicando Ensemble para classificação de textos curtos em português do Brasil

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

Paulo de Assis Nascimento

Aplicando Ensemble para classificação de textos curtos em português do Brasil

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Aprendizagem de Máquina

Orientador: Leandro Maciel Almeida

Recife
2019

Catalogação na fonte
Bibliotecário Vimário Carvalho CRB4-1204

N244 Nascimento, Paulo de Assis.
Aplicando Ensemble para classificação de textos curtos em português do Brasil / Paulo de Assis Nascimento. – 2019.
98 f.: il., fig., tab.

Orientador: Leandro Maciel Almeida.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências.

1. Inteligência computacional. 2. Sentimentos. 3. Aprendizagem de máquinas. 4. Ensemble. I. Almeida, Leandro Maciel (orientador). II. Título.

006.3 CDD (23. ed.) UFPE-MEI 2019-155

Paulo de Assis Nascimento

"**Aplicando Ensemble para classificação de textos curtos em português do Brasil**"

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 29 de agosto de 2019.

BANCA EXAMINADORA

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática / UFPE

Prof. Dr. João Fausto Lorenzato de Oliveira
Escola Politécnica / UPE

Prof. Dr. Leandro Maciel Almeida
Centro de Informática / UFPE
(Orientador)

Dedico este trabalho à minha família, aos meus pais, irmãos, amigos que sempre me incentivaram nesta jornada, e em especial, ao meu orientador que me guiou neste percurso.

AGRADECIMENTOS

Este trabalho é composto por diversos incentivos dos meus familiares e dos amigos dos últimos anos, e também daqueles que fizeram parte dos primeiros anos da minha vida escolar.

Ao postar numa rede social que iniciei o curso de mestrado, uma amiga da adolescência me fez recordar de um antigo desejo com a seguinte mensagem: "Lembro exatamente como se fosse hoje. Na nossa adolescência você falava que iria fazer isso e que tinha esse objetivo. Fico muito feliz por você meu amigo". Nessa mesma época de adolescente fanático por informática e monitor de cursinho ganhei o apelido de "Mestre"Paulo. É às pessoas dessa época, dos primeiros anos da minha vida, a quem direciono os meus agradecimentos: Andréa Batista, André dos Santos, Ronaldo Cândido, Fabiano Fernando, Anderson Clayton, José Fábio e Paulo Vieira. Todos me incentivaram no sentindo de continuar caminhando rumo ao meu sonho de ser um profissional da tecnologia. Agradeço também ao meu amigo da graduação com quem alimento o sonho de crescimento no desenvolvimento de um trabalho no qual acreditamos juntos há anos, o meu amigo Arthur Pinangé, que sempre me incentiva com as suas palavras. Aos meus professores da graduação, em especial, à Professora Carla Cristina e ao Professor Lenin Abadié Otero nos quais me inspirei a dar sequência na busca por mais conhecimento acadêmico. Aos amigos que o mestrado meu trouxe: Débora Conceição, Felipe Navarro e Raimundo Martins. Todos contribuíram diretamente para que eu chegasse neste momento. Agradeço imensamente aos meus pais: minha mãe, Dona Luzinete H. de Assis Nascimento e ao meu pai o Sr. Antônio Ernesto do Nascimento. Aos meus irmãos André de Assis e David de Assis Nascimento. Agradeço também à minha esposa Salete Paiva da Silva, que assistiu a todo meu caminhar em direção a este dia. Por fim, agradeço aos Professores Profº. Dr. Cléber Zanchettin que viu a minha ideia ser testada em sua disciplina e, de forma bastante especial, agradeço ao meu orientador o Profº. Dr. Leandro Maciel Almeida, que me deu todo o suporte e acreditou na ideia um pouco audaciosa de atuar sobre um tema modestamente explorado. A todos, o meu muitíssimo obrigado!

Kennst du das Land, wo die Zitronen blühn, Im dunkeln Laub die Goldorangen glühn,
Ein sanfter Wind vom blauen Himmel weht, Die Myrte still und hoch der Lorbeer steht?
Kennst du es wohl? Dahin! Dahin möcht' ich mit dir, O mein Geliebter, ziehn
(GOETHE, 1873, p.141)

RESUMO

A popularização da internet no Brasil e o vasto uso das redes sociais permitem às pessoas a ter voz ativa onde suas opiniões não estão mais restritas a ambientes familiares. O constante uso da internet desencadeia a criação de conteúdos diversos e muito valiosos para negócios e tomadas de decisão. Estima-se que no Brasil haverá 99,4 milhões de usuários acessando a internet até o final do ano 2019. O conteúdo lançado na web desperta o interesse das empresas que desejam melhorar seus produtos e serviços. Reunir esses dados, processá-los e transformá-los em informação útil, é essencial para mapear os perfis de consumo dos usuários na web. Para isso, é necessário lançar mão de recursos automáticos de processamento de textos. O processamento automático desse tipo de informação está ligado à atividade de Análise de Sentimentos (AS), que trata do processamento automático de textos opinativos na web classificando-os em sentimentos. A aplicação dessa técnica em português do Brasil ainda é bastante modesta. Neste sentido, este trabalho explora a aplicação da técnica de *ensemble* para classificar textos curtos em português do Brasil, sobre o problema de múltiplas classes, utilizando a abordagem de Aprendizagem de Máquina (AM). *Ensembles*, em Aprendizagem de Máquina, são utilizados quando se deseja unir em um comitê os pontos fortes de cada algoritmo. Dessa forma eles atuam como algoritmos complementares para atingir melhores resultados em relação às suas capacidades de forma isolada. Para tal, sete classificadores clássicos de Aprendizagem de Máquina (AM) foram selecionados. Para os experimentos, os corpora 2000-tweets-BR e o TweetSentBR disponíveis na literatura recente foram utilizados, ambos contém três classes. Nos experimentos, os classificadores foram treinados e testados de forma isolada a fim de obter seus resultados médios em acurácia, F-Measure, Brier Score e tempo de execução por meio da técnica de validação cruzada para posterior comparação com os *ensembles*. O teste de Shapiro-Wilk foi utilizado sobre os dados a fim de verificar a normalidade, e assim decidir o tipo de teste de hipótese a ser aplicado. Todos os classificadores isolados foram combinados entre si formando oito *ensembles* dos quais uma combinação foi baseada na métrica Brier Score. Os testes com algoritmos clássicos obtiveram os resultados médios de 71% de acurácia, 46% F-Measure, e 93 segundos de tempo de execução sobre o córpus TweetSentBR. E sobre o córpus 2000-tweets-BR foram obtidos 68% de acurácia, 57% de F-Measure e 0,430 segundos de tempo de execução. Os resultados obtidos em valores médios nos testes combinando classificadores em *ensemble* juntamente com o voto majoritário foram de 71% de acurácia, 50% de F-Measure, e 189 segundos em tempo de execução sobre o corpus TweetSentBR. Sobre o córpus 2000-tweets-BR os resultados médios obtidos foram de 69% de acurácia, 52% F-Measure e 163 segundos de tempo de execução.

Palavras-chaves:Análise de Sentimentos. Ensemble. Classificação de textos curtos. Voto Majoritário. Português.

ABSTRACT

The popularization of the Internet in Brazil and the widespread use of social media enable people to have an active voice where their opinions are no longer restricted to familiar environments. The constant use of the Internet triggers the creation of diverse and very valuable content for business and decision making. It is estimated that in Brazil there will be 99.4 million users accessing the internet by the end of the year 2019. Content launched on the web arouses the interest of companies that wish to improve their products and services. Gathering this data, processing it and turning it into useful information, is essential for mapping user consumption profiles on the web. This requires the use of automatic word processing features. The automatic processing of this type of information is linked to the Sentiment Analysis activity, which deals with the automatic processing of opinion texts on the web by classifying them in feelings. The application of this technique in Brazilian Portuguese is still quite modest. In this sense, this work explores the application of the ensemble technique to classify short texts in Brazilian Portuguese, on the problem of multiclass classification, using the Machine Learning approach. Ensembles, in Machine Learning, are used when you want to combine the strengths of each algorithm. Thus they act as complementary algorithms to achieve better results in relation to their capacities in isolation. For this purpose, seven classic Machine Learning classifiers were selected. For the experiments, the corpora 2000-tweets-BR and TweetSentBR available in recent literature were used, both containing three classes. In the experiments, all classifiers were trained and tested in isolation to obtain their average results in accuracy, F-Measure, Brier Score and execution time through the cross validation technique for later comparison with the ensembles. The Shapiro-Wilk test was used on the data to verify their normality, and thus to decide the type of hypothesis test to be applied. All isolated classifiers were combined to form eight ensembles of which one combination was based on the Brier Score metric. The tests with classical algorithms obtained the average results of 71% accuracy, 46% F-Measure, and 93 seconds of runtime over the TweetSentBR corpus. And on the 2000-tweets-BR corpus, 68% accuracy, 57% F-Measure and 0.430 seconds runtime were obtained. The results obtained in average values in the tests combining ensemble classifiers together with the majority vote were 71% accuracy, 50% F-Measure, and 189 seconds at run time on the TweetSentBR corpus. Regarding the 2000-tweets-BR corpus, the average results obtained were 69% accuracy, 52% F-Measure and 163 seconds of execution time.

Keywords: Sentiment Analysis. Ensemble. Short Texts Classification. Majority Voting. Portuguese.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
AS	Análise de Sentimentos
CBOW	Continuous Bag-Of-Words
CGU	Conteúdo Gerado pelo Usuário
CNN	Convolutional Neural Network
FN	Falso Negativo
FP	Falso Positivo
GD	Gradiente Descendente
IA	Inteligência Artificial
KNN	K-Nearest Neighbor
LBFGS	Limited Memory Broyden Fletcher Goldfarb Shanno
LSTM	Long Short-Term Memory
MO	Mineração de Opinião
NB	Naïve Bayes
NILC	Núcleo Interinstitucional de Linguística Computacional
OVO	One-vs-One
PLN	Processamento de Linguagem Natural
POS	Part-of-Speech
RBF	Radial Basis Function
RI	Recuperação de Informação
RNA	Rede Neural Artificial
RNN	Recurrent Neural Network
RNR	Rede Neural Recorrente
SGD	Stochastic Gradient Descent
SVC	Support Vector Classification
SVM	Support Vector Machine

TF-IDF	Term Frequency-Inverse Document Frequency
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Nos últimos anos, com a popularização da internet no Brasil e o vasto uso das redes sociais, as pessoas passaram a ter voz ativa onde suas opiniões não estão mais restritas a ambientes familiares. O constante uso da Internet desencadeia a criação de conteúdos diversos e muito valiosos para negócios e tomadas de decisão. Segundo o site Statista[1], haverá no Brasil, até o final do ano de 2019, 99,4 milhões de usuários acessando a Internet (STATISTA, 2019).

O conteúdo lançado na web, chamado de Conteúdo Gerado pelo Usuário (CGU), desperta o interesse de muitas empresas que desejam melhorar seus produtos e serviços. O CGU normalmente é produzido em qualquer plataforma on-line onde é possível a livre expressão sobre qualquer assunto. Nas redes sociais, esses conteúdos passam a ter um papel fundamental, devido ao seu conteúdo carregado de opinião. Normalmente, as pessoas passam a consultar a opinião alheia nas redes antes de adquirir um produto. Esse comportamento é útil para as empresas melhorarem seus serviços. Ao lidar com CGU deve-se considerar diversas formas de expressão, normalmente informais, onde se encontra palavras como *vc* em vez de você, ou *kd?*, uma forma fonética para a expressão *aonde está?* Essas expressões não são uma particularidade do português, mas sim, dos CGU, que em alguns casos, como nas redes sociais, são modificados a fim de atingir o objetivo da compreensão e atender ao limite de caracteres exigido por uma plataforma como ocorre no microblog Twitter onde a quantidade de caracteres se limita a 280.

Para processar automaticamente as informações disponíveis aplica-se a técnica de Processamento de Linguagem Natural, uma área da computação que lança mão das técnicas de Inteligência Artificial para o processamento da língua de forma automática. Segundo NEY, o objetivo da PLN é projetar e construir sistemas capazes de analisar língua natural em qualquer idioma. Dentro da área de PLN estão as áreas de apoio como Recuperação de Informação (Information Retrieval), Análise de Sentimentos, e Compreensão da Linguagem (Language Understanding) (NEY, 2019).

A Análise de Sentimentos é o processamento automático de textos em relação a sentimentos, opiniões, e subjetividade (PANG; LEE et al., 2008). A AS se concentra na classificação e clusterização (clustering) de textos por meio da abordagem de Aprendizagem de Máquina, ou da abordagem por léxico. Alguns autores relacionam AS com Mineração de Opinião, enquanto WOLFGRUBER sugere que MO está mais relacionada com pesquisadores da área de Recuperação de Informação, enquanto AS é referenciada por pesquisadores que lidam com todo Processamento de Linguagem Natural.

[1] <https://www.statista.com>. Acessado em 30 de junho de 2019

Para aplicar AS baseada em léxico é necessário um dicionário de palavras previamente selecionadas e associadas às suas classes gramaticais (FREITAS, 2013). Para a construção dos dicionários de palavras se utiliza a técnica de POS-Tagger para etiquetar as palavras para posterior análise. (THAKKAR; PATEL, 2015) define a abordagem em léxico como amadora, porque, segundo o ele, há baixo desempenho quanto à acurácia exigindo, dessa forma, demasiado esforço para atingir 80% da métrica . Os métodos de AM, por sua vez, passaram a ser utilizados com mais frequência nos trabalhos devido aos bons resultados obtidos em tarefas de classificação de textos (THAKKAR; PATEL, 2015).

A Análise de Sentimentos (AS) tem várias aplicações a depender do contexto a ser aplicado como na avaliação de produtos, crítica de filmes (reviews), identificação de polaridade política, avaliação de entidades e serviços *etc*. No contexto da avaliação de produtos, por exemplo, (AVANÇO; BRUM; NUNES, 2016) analizaram opiniões extraídas de ambientes de compras on-line aplicando métodos de combinação de algoritmos. E (FRANÇA; OLIVEIRA, 2014) analisaram tweets sobre as manifestações políticas ocorridas no Brasil em 2013. No trabalho, os autores utilizaram a classificação binária para classificar polaridades.

Os estudos a respeito de emoções extraídas de textos por meio automático não são recentes. Há trabalhos dos anos 1990 onde se fala do reconhecimento computacional de emoções humanas (AVANÇO, 2015). Vários trabalhos foram publicados descrevendo atividades de extração de informação a partir de textos utilizando os mais variados métodos. Porém, a maioria desses trabalhos é desenvolvida no idioma inglês ao passo que no idioma português a produção ainda é modesta (DOSCIATTI; FERREIRA; PARAISO, 2015).

O português é um dos idiomas mais falados no mundo com mais de 291 milhões de falantes nas suas diversas variedades em 10 países[2]. Ele se torna mais um desafio para PLN por considerar as diversas variações formais sintáticas e morfológicas o que dificulta a identificação da relevância semântica das palavras. (ZAMPIERI; MALMASI; DRAS, 2016) destacam o exemplo das mensagens "*homem grande*" e "*grande homem*", onde a primeira está relacionada à altura do homem no sentido literal, enquanto a segunda utiliza o sentido figurado para se referir a uma pessoa com grandes qualidades. As variações do idioma podem influenciar no processo de classificação de textos conforme descreveu (VITÓRIO et al., 2017) ao identificar que as variações entre o português do Brasil e o europeu influenciam no desempenho de classificação.

Mesmo contando com poucos trabalhos voltados ao processamento da língua portuguesa na variante brasileira, algumas atividades foram feitas nos últimos anos sobre os mais variados contextos. BRUM; NUNES construiram um córpus de sentimentos a partir de mensagens extraídas da rede social Twitter, VITÓRIO et al. investigou as variações linguísticas entre o português brasileiro e o europeu, ZAMPIERI; MALMASI; DRAS investigou a classificação temporal de textos do português histórico., MORAES; MANSSOUR; SILVEIRA

[2] Brasil (212.378.921), Moçambique (31.391.653),Angola (31.828.092), Portugal (10.230.000), Guiné-Bissau (1.921.049), Timor Leste (1.351.757), Guiné Equatorial (1.360.000), Macau (642.090), Cabo Verde (538.535) e São Tomé e Príncipe (213.305)

classificou opiniões extraídas de tweets durante a copa do mundo de futebol no Brasil.

Nesse sentido este trabalho propõe a aplicação da técnica de *ensemble* baseada em classificadores convencionais de Aprendizagem de Máquina aplicados à Análise de Sentimentos em português do Brasil.Assim, se busca resposta para as seguintes perguntas: (1) Qual a representação de característica ideal para textos em português do Brasil, e (2) Qual classificador tem melhor performance em acurácia e F-Measure médias para atividades de Análise de Sentimentos (AS) em português do Brasil?

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Devido à pouca quantidade de trabalhos que relatam atividades de classificação de textos curtos (tweets) em português do Brasil, este trabalho tem como objetivo comparar diferentes algoritmos clássicos de Aprendizagem de Máquina com *ensemble* de classificadores para classificação desses textos, a fim de investigar se o *ensemble* de classificadores obtém melhor resultado na atividade de classificação em comparação com os classificadores executados isoladamente.

1.2.2 Objetivos Específicos

- Reunir um conjunto de córpus de textos, corpora, em português do Brasil utilizados em Análise de Sentimentos.
- Combinar os classificadores convencionais em *ensemble* para classificação de textos em português do Brasil.
- Identificar qual a melhor técnica para classificação de textos curtos em português do Brasil.
- Identificar qual o método de Word Embeddings apropriado para classificar textos em português do Brasil.
- Explorar trabalhos aplicados sobre bases de textos em Português do Brasil.

1.2.3 Estrutura da Dissertação

1. **Introdução:** Apresenta a motivação deste trabalho, introduz a Análise de Sentimentos (AS) e apresenta os objetivos deste trabalho.
2. **Fundamentação:** Reúne os conceitos básicos para o entendimento deste trabalho.
3. **Trabalhos Relacionados:** Apresenta os trabalhos desenvolvidos sobre atividades de Análise de Sentimentos (AS) em português do Brasil.

4. **Método Proposto:** Apresenta a metodologia do processo definido na arquitetura proposta nesta pesquisa.
5. **Experimentos:** Relata os experimentos e os resultados obtidos.
6. **Conclusão:** Retoma as questões levantadas neste trabalho e sugere trabalhos futuros.

2 FUNDAMENTAÇÃO

Esta pesquisa se concentra no estudo da classificação de textos em português do Brasil considerando as polaridades positiva, negativa e neutra sobre dois corpora de sentimentos construídos por pesquisadores que utilizaram o Twitter como fonte de dados. Neste capítulo são apresentadas as definições da Análise de Sentimentos e dos conceitos básicos para compreensão deste trabalho.

2.1 ANÁLISE DE SENTIMENTOS

Os estudos sobre a detecção de sentimentos a partir de textos são dos anos 2000, após a popularização da internet e o uso constante dos fóruns, salas de bate-papo, artigos de notícias e páginas da web (FRANÇA; OLIVEIRA, 2014). Nesse período, o estudo de sentimentos a partir de textos ganhou notoriedade devido ao crescente uso de técnicas de Aprendizagem de Máquina para auxiliar nas atividades de Processamento de Linguagem Natural e Recuperação de Informação para enfrentar os desafios da análise de mercado (WOLFGRUBER, 2015).

Segundo BRUM, as primeiras referências ao termo Análise de Sentimentos foram feitas em NASUKAWA; YI ao relatar a "busca de expressão de sentimento a um dado sujeito e a determinação da polaridade deste sentimento".

As definições para Análise de Sentimentos ou Mineração de Opinião são diversas. ALMASHRAEE; DÍAZ; PASCHKE, por exemplo, preferiu dizer que"AS analisa textos on-line e declara os sentimentos ou opiniões dos escritores em relação a um objeto específico ou a qualquer um de seus recursos ou aspectos incluídos". Já a definição de PANG; LEE et al. pode ser resumida como o processamento automático de textos em relação a sentimentos, opiniões, e subjetividade. E MEDHAT; HASSAN; KORASHY pontuou que AS e MO são o estudo computacional da opinião, atitudes e emoções das pessoas a respeito de uma entidade.

Além das variadas definições, há também a utilização de duas expressões AS e MO como sinônimas. WOLFGRUBER define que a expressão Mineração de Opinião, normalmente encontrada nas definições de AS, é bastante utilizada por pesquisadores que lidam com Recuperação de Informação, do inglês, Information Retrieval, enquanto Análise de Sentimentos é utilizada por pesquisadores da área de Processamento de Linguagem Natural e da Linguística Computacional.

Embora cada pesquisador defina AS de acordo com a sua compreensão e forma de ver o relacionamento entre a área e os seus objetos de estudo, a definição mais aceita foi feita por LIU ao incluir os pontos aos quais a AS atinge: A Análise de Sentimentos (AS), também chamada de Mineração de Opinião, é o campo de estudo que analisa opiniões,

sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e seus atributos (LIU, 2012).

As atividades de AS normalmente são aplicadas sobre sentenças curtas a fim de identificar o sentido atribuindo-lhe, assim, uma polaridade. Essas sentenças normalmente são extraídas das redes sociais ou ambientes onde a expressão de opinião ocorre de forma natural. A esses textos, sejam eles longos ou curtos, porém provenientes do ambiente web, dá-se o nome de Conteúdo Gerado pelo Usuário, do inglês, User Generated Content. A rede social mais utilizada como fonte de dados para AS é o Twitter[1], porque nela há um limite de caracteres que obriga o usuário ser objetivo no que deseja expressar. Algumas técnicas para aplicar atividades de AS são empregadas para obter melhor desempenho no processo de classificação (THAKKAR; PATEL, 2015).

- Análise Léxica

  Baseia-se sobre um dicionário léxico de palavras previamente etiquetadas/marcadas com a classe gramatical à qual elas pertencem. Assim, a sentença é separada em palavras, essa operação é chamada de tokenização, e então cada palavra é comparada com cada palavra do léxico de palavras marcadas até que cada palavra encontre o seu par e aumente uma pontuação, caso contrário a pontuação diminui e ao final a pontuação restante define a polaridade da mensagem. THAKKAR; PATEL afirma que esta técnica parece amadora e que é necessário o uso de um léxico de adjetivos, cruciais para a definição da polaridade, marcados manualmente para atingir 80% de acurácia. Essa análise tem a limitação relacionada ao desempenho no que diz respeito à acurácia. Ela cai drasticamente se houver um crescimento exponencial do tamanho do dicionário utilizado para a comparação. BRUM relata alguns léxicos para o idioma português brasileiro devidamente anotados: SentiLex (SILVA; CARVALHO; SARMENTO, 2012), Opinion Lexicon (SOUZA et al., 2011), OntoPT-Sentic (OLIVEIRA; SANTOS; GOMES, 2014), Léxico do ReLi (FREITAS, 2013), e LIWC (FILHO; PARDO; ALUÍSIO, 2013).

- Análise baseada em Aprendizagem de Máquina

  Esta é a técnica mais utilizada por pesquisadores por ser adaptável e ter bom desempenho e por se sobrepor à técnica de análise léxica quanto a sua limitação no que diz respeito ao desempenho.

  Em AS essa técnica é empregada na forma de aprendizagem supervisionada que exige uma sequência de atividades sobre os dados. Essas atividades são inerentes ao PLN e são essenciais para AS baseada em Aprendizagem de Máquina. Normalmente, a preparação de uma base de dados de textos para essa técnica é composta por quatro estágios:

[1] <https://www.twitter.com>. Acessado em 04 de junho de 2019

Coleta dos Dados: ocorre ao extrair mensagens de textos das redes sociais (Facebook[2], Twitter, Instagram[3]), blogs, portais de avaliação de produtos ou serviços como o TripAdvisor[4] e fóruns considerando do contexto (Se político, filmes, produtos *etc*) a ser utilizado.

Pré-processamento: Neste estágio, o texto recebe um tratamento de normalização. O objetivo é remover qualquer ocorrência de caracteres especiais que pode criar resultados inesperados. Entre os caracteres especiais estão os vários tipos de emojis, esses ainda são utilizados em outra técnica de classificação, cerquilhas (hashtags), acentos no caso do português, stopwords, que são palavras que aparecem diversas vezes no texto, e pontuações. Dessa forma, a sentença se transforma em um texto uniforme.

Dados de treinamento: Para a técnica de Aprendizagem de Máquina (AM) é necessário que a base de dados esteja rotulada, isto é, etiquetada com as polaridades que se deseja classificar. Em seguida, deve-se dividir a base de dados em normalmente duas partes como base de treinamento do algoritmo e base de testes. O treinamento ocorre sobre a parte da base dedicada a esse fim. Dessa forma o algoritmo aprende as estruturas e os padrões dos tipos das mensagens.

Classificação: Como último passo, a classificação vem para etiquetar mensagens desconhecidas, que não foram utilizadas no processo de treinamento do algoritmo. Vale lembrar que as mensagens também foram uniformizadas pelo pré-processamento. Neste estágio, o modelo é aplicado e dá origem a resultados para análise. Esses resultados são exibidos de forma gráfica ou de modo a suportar uma análise rápida para que o modelo seja avaliado e, se necessário, sofra as modificações para outra fase de treinamento e testes de classificação. É importante esclarecer que essa classificação ocorre sobre a base de dados separada para testes, pois ela contém as polaridades que darão subsídios para avaliação do modelo quanto ao seu desempenho. Segundo THAKKAR; PATEL, esta é a principal fase do processo.

- Análise Híbrida

  Alia as duas técnicas anteriores utilizando as suas vantagens para buscar o melhor resultado. Na análise híbrida pode-se combinar a velocidade da técnica baseada em léxico com a precisão da técnica baseada em AM. Nesta forma, se cria dois léxicos de polaridades sendo um com mensagens positivas e outro com mensagens negativas. Assim, se pode aplicar o cálculo da similaridade dos cossenos entre uma mensagem, até o momento desconhecida, com o léxico positivo e também com o léxico negativo.

[2] <https://www.facebook.com>. Acessado em 08 de junho de 2019
[3] <https://www.instagram.com>. Acessado em 08 de junho de 2019
[4] <https://www.tripadvisor.com.br>. Acessado em 08 de junho de 2019

Neste processo se define e se treina o modelo a depender de qual léxico se o resultado se aproximará mais.

Segundo MEDHAT; HASSAN; KORASHY a AS tem três principais níveis de classificação de textos como classificação em nível de documento, que busca classificar a opinião (positiva ou negativa) ou sentimento expressados em um documento. Ela trata todo o documento como a unidade básica de informação de modo que a opinião classificada representa todo documento conforme na Figura 1[5]; em nível de sentença, que tem o objetivo de classificar sentimento a partir de uma sentença conforme a Figura 2. Nela é possível perceber a mensagem negativa em vermelho, mas também uma mensagem neutra com cor de fundo branca; e a classificação em nível de aspecto classifica o sentimento em relação a aspectos específicos da entidade conforme Figura 3. Mas para isso se deve identificar quais são as entidades e os seus aspectos.

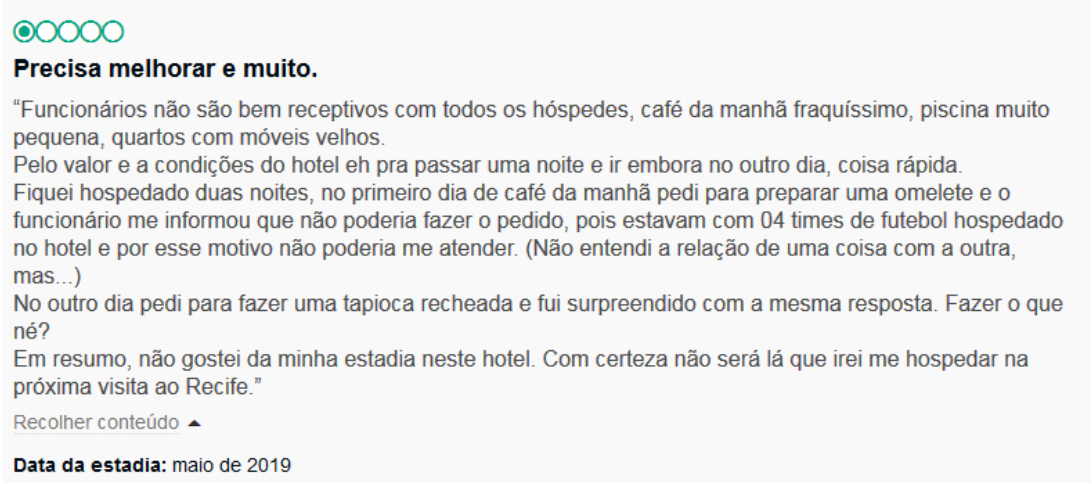
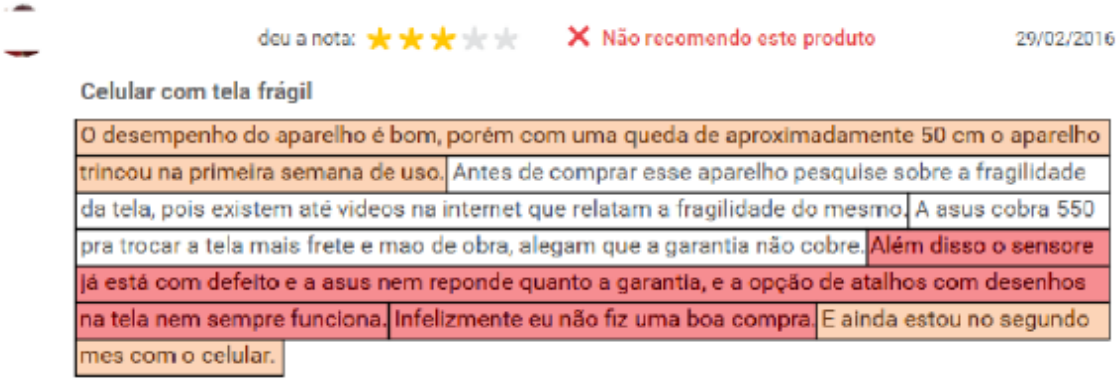


Figura 1 – Avaliação com polaridade negativa em nível de documento

Fonte: Elaborada pelo autor.

[5] <https://www.tripadvisor.com.br/Hotel_Review-g304560-d559495-Reviews-Golden_Park_Recife_Boa_Viagem-Recife_State_of_Pernambuco.html#REVIEWS>. Acessado em 03 de junho de 2019



Fonte: (BRUM, 2018)

Figura 2 – Avaliação com polaridade positiva em nível de aspecto

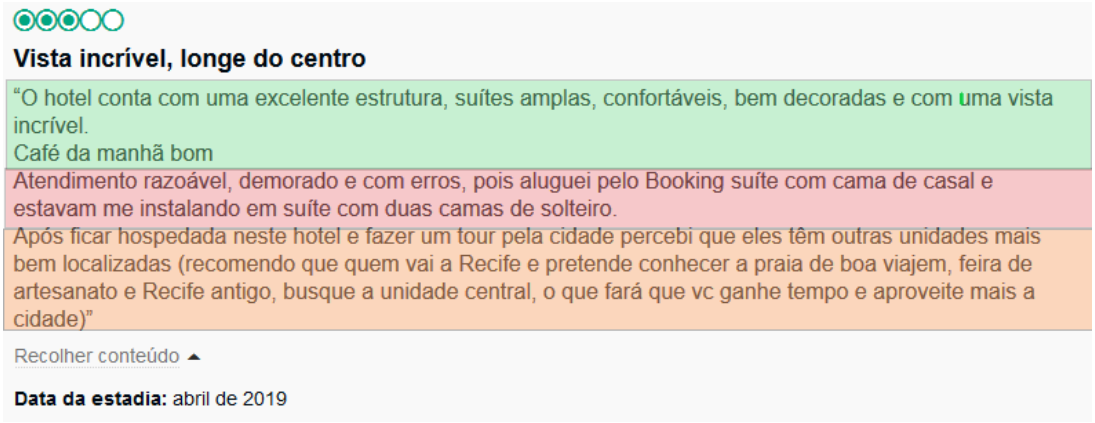


Figura 3 – Avaliação de produto em nível de sentença

Fonte:Elaborada pelo autor.

2.2 EMOÇÃO

A emoção está relacionada a sentimentos de modo que a área de AS poderia facilmente ser denominada por Análise de Emoções. Segundo MILLER et al., emoção é o sentimento que caracteriza o estado da mente humana e pode despertar os sentimentos de prazer, medo ou segurança. Influenciando diretamente na nossa tomada de decisão, as emoções são orientadas por aquilo que avaliamos como bom ou ruim. Por isso, sem avaliação não há emoção. Já que emoção positiva ou negativa depende de uma avaliação de um produto ou serviço como bom ou ruim (SHERKAT et al., 2018).

As emoções são divididas em emoções primárias, que compreendem alegria, tristeza, medo, raiva, surpresa e aversão; emoções de segundo plano, onde estão as sensações de bem e mal-estar, calma e tensão, dor e prazer, entusiasmo e depressão; e as emoções sociais, que compreendem a vergonha, o ciúme, a culpa e o orgulho (RAMOS; BERRY; CARVALHO, 2005).

Os vários sentimentos presentes nas divisões da emoção são formados de modo in-

consciente por três fatores que contribuem para que haja aceitabilidade de produto ou serviço pela avaliação inconsciente que ocorre de forma automática nos seres humanos. Esses fatores são: apreciação, preocupação e estímulo.

- **Avaliação:** A avaliação se refere à avaliação automática e inconsciente de um produto ou serviço (SHERKAT et al., 2018).
- **Preocupações:** Por trás de cada emoção há uma preocupação. Se as preocupações das pessoas puderem ser abordadas por meio das capacidades dos produtos ou serviços, então eles são avaliados como benéfico e leva à emoções positivas. No lado oposto, as capacidades de produtos ou serviços que não correspondem às preocupações das pessoas são avaliadas como prejudiciais e levam a emoções negativas (SHERKAT et al., 2018).
- **Estímulo:** O estímulo se refere ao objeto que representa qualquer produto ou serviço que possa abordar uma ou mais preocupações. Produtos físicos, serviços, itens recordados e imaginários podem ser considerados como um estímulo no modelo básico das emoções Figura 4 (SHERKAT et al., 2018).

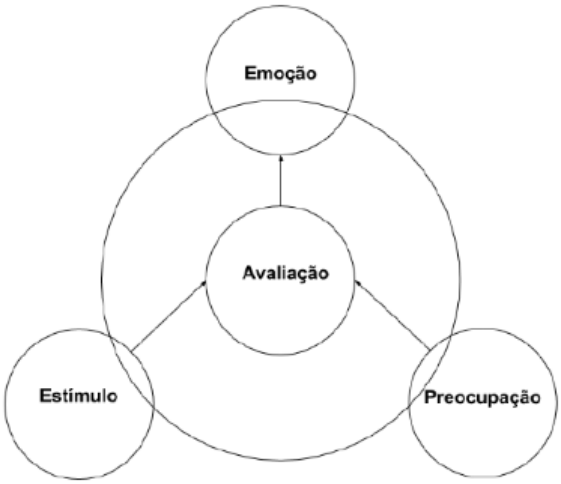


Figura 4 – Modelo básico das emoções proposto por Desmet

Fonte:Adaptada de (SHERKAT et al., 2018)

2.3 CÓRPUS

A expressão córpus, no plural corpora, está presente neste e em vários trabalhos desenvolvidos na área de Processamento de Linguagem Natural. Por isso, nesta seção se busca definir o que são corpora e como são produzidos ou identificados como tal. Segundo SARDINHA, a definição de córpus deve ser a mais abrangente possível e que possa envolver a língua na sua forma natural.

Há várias definições para córpus, e seguindo as diversas definições, ZACCARA sintetizou a definição mais completa ao definir córpus como "uma coleção de dados linguísticos, sejam eles textos ou partes de textos escritos ou a transcrição de fala, de uma determinada língua, escolhidos segundo um determinado critério, representando uma amostra desta língua ou uma variedade linguística".

De acordo com SARDINHA, há uma série de critérios para que um texto seja considerado um córpus: O córpus deve ser composto por texto na forma natural da língua. Porque dessa forma não há um viés específico para produção; o texto deve ser escrito ou ter sido produzido por um nativo; a escolha do córpus deve seguir alguns critérios de seleção. Entre esses critérios de seleção a autenticidade é o qual se deve estar mais atento.

O processo de criação de córpus demanda bastante tempo das pesquisas e é composto pelas etapas de projeto, compilação, anotação e uso. A etapa de projeto diz respeito à definição do objetivo do córpus, isto é, em qual contexto ele será aplicado. Os contextos podem ser os mais diversos como análise de notícias tendenciosas, avalição de produtos e serviços, análise de sentimentos sobre movimentos sociais nas redes sociais *etc.* A compilação se refere à extração do texto. E a última etapa, o uso, não se refere à aplicação do córpus em sistemas de classificação como o nome parece sugerir, mas sim, ao seu tratamento no que diz respeito à, se for o caso, anotação/rotulação/etiquetamento (ZACCARA, 2012).

```
niquesteffane| @rayzafelix tbm te amo!!!!|POSITIVO
Guimaraes1999_| @DudMartinss meu ultimo dia ne, tem q ter|NEUTRO
brunaerror| sexta feira e o lol dando RUIM|NEGATIVO
#BiCampeaoMundia|NEUTRO
brusakazaki| misturar estilo de tatuagem fica feio?|NEUTRO
TorresMonique| vamos usar as tag #BTS #?????|NEUTRO
```

Figura 5 – Fragmento do córpus 2000-tweets-br.csv

Fonte: Elaborada pelo autor.

2.4 PRÉ-PROCESSAMENTO DE TEXTOS

Em Processamento de Linguagem Natural o pré-processamento de textos é útil para normalizar as mensagens a fim de otimizar o custo computacional ao processar grande quantidade de informação. De acordo com (GUIMARÃES; MEIRELES; ALMEIDA, 2019), o objetivo do pré-processamento é extrair informação estruturada e manipulável de texto não estruturado. Nesta seção são exploradas as atividades do pré-processamento aplicadas neste trabalho.

2.4.1 Normalização, stopwords e Tokenização

A normalização de textos consiste na filtragem dos caracteres que são constantes numa mensagem. A aplicação desse artifício tem o objetivo de adequar os textos de modo a mantê-los puros, sem caracteres especiais de qualquer tipo visando a redução da dimensionalidade do córpus durante o processamento. Entre as atividades inerentes ao processo de normalização estão: 1) a filtragem (filtering), onde ocorre a remoção ou substituição de caracteres indesejados; 2) a normalização técnica, busca e substitui as diversas combinações de caracteres especiais que forma unicode específico, e 3) remoção de acentos (VITÓRIO et al., 2017).

As stopwords são palavras frequentemente utilizadas nos textos e tem pouco valor semântico. Como relata SAIF et al., são palavras não-discriminativas. O processo de remoção visa a redução do espaço de características dos classificadores ajudando-os, assim, a obter melhores resultados (SAIF et al., 2014). Diversos trabalhos relatam remoção dessas palavras afirmando que elas contribuem para o aumento de ruído durante o processamento textual SAIF et al.. Uma lista dessas palavras foi disponibilizada pelo framework NLTK (Natural Language Toolkit)[6]. As palavras “a”, “ao”, “aos”, “aquela”, “aquelas”, “aquele”, “aqueles”, “aquilo”, “as” e “até” são exemplos de stopwords em português.

Tokenização é a subtarefa de dividir um texto em palavras e tokens (ALMASHRAEE; DÍAZ; PASCHKE, 2016). Como resultado da tokenização é possível obter também pontuações. Essa fase é importante para atividades de AS, pois os tokens atuam como parte elementar das orações. A mensagem “O Brasil é o país do futuro!”, após a tokenização aparece da seguinte forma [O, Brasil, é, o, país, do, futuro, !].

2.5 CLASSIFICAÇÃO

Classificar dados é uma atividade largamente utilizada por diversas áreas como na área de Processamento de Imagens, Diagnóstico Médico e Organização de Documentos, não limitando-se somente a textos (ALLAHYARI et al., 2017). Conforme SHERKAT et al., a classificação é a parte mais importante na atividade de AS. Ela, a classificação, está relacionada à atividade de associar uma polaridade ou classe predefinida à uma sentença ou documento. Para cada tipo de problema há um tipo de classificação. Os problemas podem ser:

- Binário: que compreende as decisões do tipo sim/não, positivo/negativo, 0/1, verdadeiro/ falso... A classificação binária também pode ser chamada de valente e ser referida como classificação bivalente. MOHAMMAD, inclusive, cita ORTONY; CLORE; COLLINS ao argumentar que emoções são valentes, isto é, binárias, então postivias ou negativas e nunca neutras. Uma grande quantidade de trabalhos sobre AS utiliza a classificação binária por obter melhor desempenho no que diz respeito à precisão.

[6] <<http://www.nltk.org/howto/portuguese_en.html>>. Acessado em 04 de julho de 2019

- Categórico: é o problema que pode ser classificado em mais de duas categorias. Quando se refere aos problemas categóricos surge também o conceito de classe como sinônimo de categoria. Então, se passa a considerar juntamente das polaridades positiva e negativa também a polaridade neutra. Esse tipo de classificação de textos normalmente tem perda de desempenho já que as mensagens neutras estão no limiar entre as positivas. BRUM relata mais categorias como a divisão por quatro classes (muito positiva, positiva, negativa, muito negativa) ou cinco classes (muito positiva, positiva, neutra, negativa e muito negativa) que neste trabalho também se caracteriza como categóricas.

A classificação ocorre sobre um vetor de características que representa um padrão que caracteriza as classes. Esse vetor de características representa os atributos que diferencia uma sentença positiva da negativa e vice-versa. Há casos em que os vetores de características podem conter variabilidade (ruído), que influencia na associação de diferentes polaridades para a mesma classe, isto é, duas sentenças negativas são classificadas de forma diferente aumentando o grau de dificuldade de classificação (DUDA; HART; STORK, 2012).

Ao utilizar classificação de textos, é necessário utilizar métricas diferentes em relação à classificação aplicada a outros domínios. Essas métricas são úteis para avaliar o desempenho dos classificadores. A métrica acurácia, por exemplo, é utilizada normalmente em algumas áreas e corresponde ao total de classes corretamente classificadas em relação ao total de sentenças submetidas à classificação (ALLAHYARI et al., 2017). O seu resultado não se preocupa em saber qual classe foi classificada corretamente.

Na Análise de Sentimentos (AS) é comum utilizar as métricas de precisão, recall e F-Measure para medir o desempenho dos classificadores. A precisão é a métrica que corresponde ao resultado obtido pela correta classificação de uma sentença que, de fato, corresponde à classe originalmente associada. Um classificador é preciso quando ele consegue classificar corretamente o maior número de classes sejam elas positivas, negativas ou neutras. Recall é definida como a porcentagem de classes corretamente classificadas dentre as classes presentes no universo. E F-Measure ou F1-score é a métrica que se baseia na média geométrica entre precisão e recall (ALLAHYARI et al., 2017). A análise do resultado de uma classificação considerando a precisaão pode ser feita por meio da matriz de confusão. Esse recurso é essencial para apontar quais foram as classes corretamente classificadas como positivo (verdadeiro positivo), as classes errôneamente classificadas como positivo (falso positivo) e o mesmo para negativo e neutro.

A Figura 6 exibe o resultado da 4ª etapa (fold)[7] da classificação de textos curtos sobre uma base de dados previamente rotulada. As siglas MLP, REG e GAU representam Multilayer Perceptron, Regressão Logística e Gaussian Naïve Bayes respectivamente. No eixo X, 0 está para positivo, 1 para negativo e 2 para neutro, essas categorias representam o

[7] k-fold Cross-validation

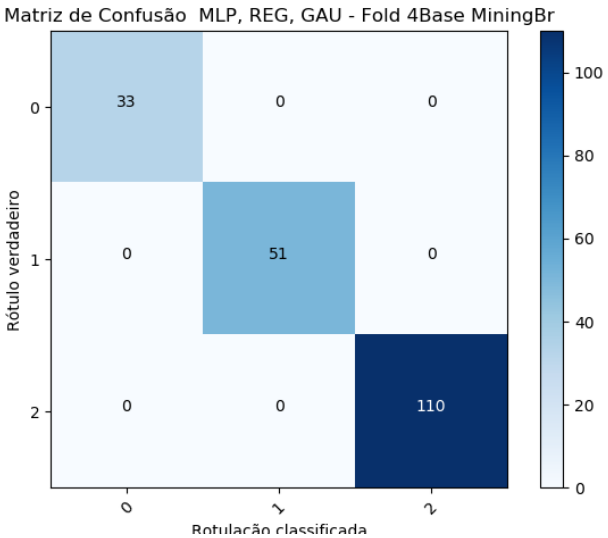


Figura 6 – Matriz de confusão

Fonte: Elaborada pelo autor.

resultado da classificação, e no eixo Y estão os verdadeiros rótulos seguindo a mesma definição para 0, 1 e 2. Na Figura 6 se pode constatar uma sequência de verdadeiros positivos, verdadeiros negativos e verdadeiros neutros. Nela lê-se que das 110 mensagens neutras, 110 delas foram classificadas como neutras. O mesmo ocorreu para as 51 mensagens negativas e para as 33 mensagens positivas.

2.6 WORD EMBEDDING

Word Embedding é um conjunto de técnicas para representação de características semânticas e sintáticas de textos em vetores. Essas técnicas são utilizadas em métodos de classificação de textos utilizando Aprendizagem de Máquina. Porque muitos algoritmos de AM exigem vetores de características de tamanhos fixos. Assim, as palavras ou sentenças são convertidas em um vetor de números reais para serem utilizados como entrada para algoritmos de AM, como por exemplo, a RNA. Para isso, os algoritmos de Word Embeddings utilizam técnicas de IA e cálculos estatísticos. Um bom exemplo utilizado em várias publicações é a aproximação entre duas palavras do mesmo contexto por meio do cálculo da distância entre essas palavras conforme a Figura 7.

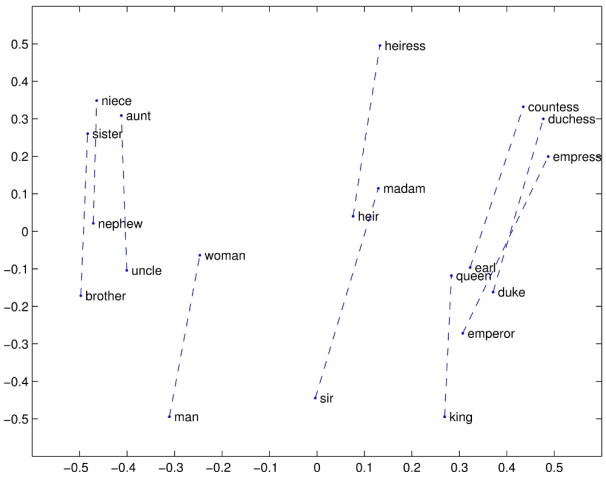


Figura 7 – Representação da similaridade entre palavras usando Glove

Fonte: (PENNINGTON; SOCHER; MANNING, 2019)

Na Figura 7 se pode ver as relações entre homem (*man*) e mulher (*woman*). Essa relação se dá por ambos serem seres humanos, mas se deve considerar a relação de oposição entre os sexos caracterizada pelo eixo Y. Segundo PENNINGTON; SOCHER; MANNING, a analogia é de que rei está para rainha assim como homem está para mulher.

Existem diversos algoritmos de Word Embedding disponíveis na literatura, mas os mais presentes nos trabalhos são GloVe[8] (Global Vectors for Word Representation) (PENNINGTON; SOCHER; MANNING, 2014), Continuous Bag-Of-Words (CBOW) e Skip-gram (MIKOLOV et al., 2013), Term Frequency-Inverse Document Frequency (TF-IDF). Segundo LE; MIKOLOV, um dos métodos mais utilizados para classificação de textos é o Bag of Words.

GloVe: Global Vectors for Word Representation combina os métodos Global Matrix Factorization e Local Context Window para capturar co-ocorrência de palavras. Assim é possível extrair significados e aspectos de palavras que co-ocorrem no mesmo contexto por meio do cálculo da probabilidade de co-ocorrência (PENNINGTON; SOCHER; MANNING, 2014).

CBOW: CBOW é um modelo que compartilha as palavras na mesma posição no espacço de características. Segundo LE; MIKOLOV, o CBOW tem desvantagens de não considerar a posição das palavras nas frases e ignorar a semântica. O objetivo do CBOW é prever palavras futuras dado o contexto (MIKOLOV et al., 2013).

Skip-gram: Originalmente, Continuous Skip-gram Model, é um modelo semelhante ao CBOW, mas em vez de prever uma palavra dado um contexto, ele tenta maximizar a classificação de uma palavra baseada em outra na mesma sentença (MIKOLOV et al., 2013).

TF-IDF: TF-IDF é um modelo estatístico baseado sobre o bag of words padrão. Calcula a frequência que um termo aparece em um documento pela sua frequência inversa. Segundo RAMOS et al., uma palavra com TF-IDF alto denota uma forte relação entre a

[8] <<https://nlp.stanford.edu/projects/glove/>>.Acessado em 13 de junho de 2016

palavra e o documento aonde ela está presente. O TF-IDF tem mostrado bons resultados em atividades de AS porque ele considera documentos e não apenas palavras.

FastText: recentemente criado, é um método de Word Embeddings baseado no algortimo skip-gram onde considera cada palavra um n-gram (BOJANOWSKI et al., 2017). Esse modelo busca extrair informações morfológicas para criar os Word Embeddings. Pois, segundo, BOJANOWSKI et al., os modelos presentes na literatura ignoram a morfologia das palavras.

Neste trabalho foram aplicados testes preliminares de desempenho em acurácia afim de selecionar o melhor modelo para extrair características dos textos utilizados. Após os testes, neste trabalho passou-se a utilizar o algoritmo TF-IDF por ter obtido melhor desempenho. O TF-IDF tem maior capacidade de extrair valor semântico das palavras conseguindo, dessa forma, representar bem a polaridade de uma sentença. Diversos trabalhos na literatura apontam o uso do TF-IDF para classificação de textos em português do Brasil como CASTRO et al., HARTMANN et al., AVANÇO, AVANÇO; NUNES.

2.7 APRENDIZAGEM DE MÁQUINA

Trata-se da forma como os computadores aprendem e identificam padrões a partir dos dados. A Aprendizagem de Máquina está presente em várias atividades como Reconhecimento de Imagens, Identificação de Sequências de DNA, Reconhecimento de Caracteres e Processamento de Linguagem Natural (PLN). Essas atividades envolvem padrões ou comportamentos previsíveis a partir de características associadas a uma classe, a um grupo ou categoria (DUDA; HART; STORK, 2012). O desenvolvimento do mecanismo para identificar automaticamente esses padrões é chamado de modelo. Assim, se diz que se desenvolve um modelo para identificar, por exemplo, se as mensagens são positivas, negativas ou neutras, não sendo possível identificar com o mesmo modelo se uma determinada mensagem é duvidosa.

As atividades de AM necessitam de processos de preparação para que o modelo gerado tenha o melhor desempenho possível no sentido de obter o melhor tempo de treinamento possível, isto é, ser computacionalmente viável e também no sentido de ter a melhor acurácia, além de ser capaz de generalizar. A generalização ocorre quando o modelo é treinado sobre características diferentes ao ponto de conseguir identificar a classe ou categoria de um objeto nunca visto. Isso signfica que tal objeto não fez parte da base de dados de treinamento e mesmo assim foi possível idenficar a sua classe (DUDA; HART; STORK, 2012).

Coleta de dados: Pode ser a atividade mais custosa no processo de AM. De acordo com DUDA; HART; STORK, é possível iniciar testes com uma pequena quantidade de dados, mas no decorrer do processo mais dados serão necessários para obter melhor performance do modelo.

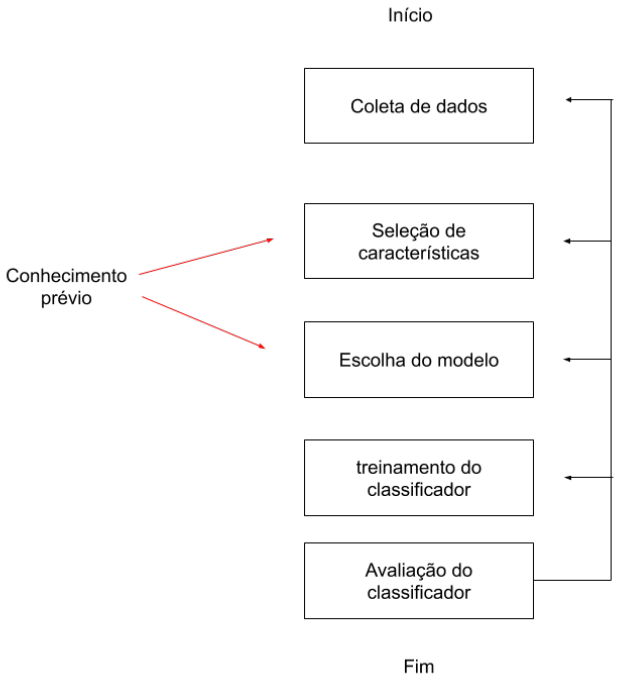


Figura 8 – Arquitetura de um sistema de reconhecimento de padrões

Fonte: Adaptado de (DUDA; HART; STORK, 2012)

Seleção de características: No contexto da AS, essa fase representa a aplicação de algum algoritmo de Word Embedding como CBOW, Skip-gram ou TF-IDF para extrair características dos textos. A seleção de características não está relacionada apenas à aplicacção de uma técnica de extração, mas também à seleção das melhores características para representar um documento. Uma boa seleção de características torna o trabalho do classificador mais trivial (DUDA; HART; STORK, 2012).

Escolha do modelo: Em caso de insatisfação com os resultados, o modelo pode ser ajustado para obter melhor performance. Nesta fase, se analisa se o modelo previu uma determinada classe a partir das características disponíveis. DUDA; HART; STORK levanta algumas questões importantes que se deve fazer nesta fase.

1. "Como saber se o modelo hipotético difere do modelo idealizado?"

2. "Como saber o momento de descartar uma classe do modelo e selecionar outra?"

Treinamento do classificador: Treinar classificador pode ser considerado como processar os dados obtidos para o treinamento do modelo por meio das características. Assume-se que esses dados são compostos por vetores de características que representam classes previamente definidas (conhecimento prévio). Nesse processo de treinamento, o classificador procura treinar o modelo para que seja possível prever as classes das mensagens, no caso de AS, até então não vistas.

Avaliação do classificador: Após o treinamento aplica-se o teste do classificador sobre um conjunto de dados ainda não vistos. A avaliação contempla análise da performance do classificador, taxa de erros, custo computacional e outros resultados inerentes ao problema. Nesta etapa é possível identificar também se o modelo obteve bons resultados durante o processo de treinamento, mas não obteve o mesmo sucesso sobre dados até então não vistos. Essa situação é conhecida como overfitting. Quando um modelo apresenta overfitting, significa que ele perdeu o poder de generalização e não consegue obter os bons resultados sobre dados nunca vistos.

2.7.1 Tipos de Aprendizagem

A aprendizagem ocorre na busca pela redução da taxa de erros do modelo. A taxa de erros pode ser reduzida atravé do método GD, porque ele altera os parâmetros do classificador para reduzir a taxa de erros buscando pelo mínimo local. Segundo DUDA; HART; STORK, no processo de criação de um modelo para reconhecimento de padrões se investe bastante tempo na busca por um classificador que obtenha melhor desempenho. Isso se encontra nos classificadores que aplicam GD. Os tipos de aprendizagem podem ser:

Aprendizagem Supervisionada, que ocorre sobre os dados com classes previamente definidas em um conjunto de dados definido como conjunto de treinamento. No contexto da AS, a aprendizagem supervisionada atua sobre o córpus após a última etapa descrita na seção 2.3.

Aprendizagem Não-Supervisionada ou, segundo DUDA; HART; STORK, **clustering**, em português agrupamento, não precisa de um conjunto de dados rotulados para aprender os padrões. Esse processo ocorre através da similaridade entre os objetos de um conjunto onde os mais semelhantes são agrupados. A Figura 9 ilustra o agrupamento de três classes de objetos onde os similares pertencem ao mesmo grupo. Considerando AS no problema categórico de três classes, a aprendizagem não supervisionada utilizaria grupos representando as classe, onde as mensagens semelhantes seriam agrupadas.

Aprendizagem por Reforço ocorre sob a crítica binária de certo ou errado. Se o classificador prevê uma classe diferente da presente no conjunto de treinamento, ele recebe um feedback de negativo. Sendo o erro desconhecido, então o classificador é obrigado a ajustar os seus parâmetros e tentar novamente até que ele receba um feedback positivo.

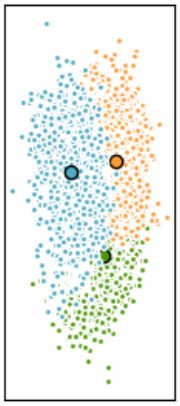


Figura 9 – Exemplo de Aprendizagem Não-Supervisionada (clustering), segundo DUDA; HART; STORK

Fonte: Elaborada pelo autor.

2.8 ABORDAGEM DE DECISÃO, VALIDAÇÃO E MÉTRICAS

2.8.1 Voto Majoritário

O sistema do Voto Majoritário, Majority Voting, se baseia originalmente na fórmula de maioria simples, onde se calcula a maior quantidade de votos para uma das opções concorrentes. Esse sistema é aplicado em diversas áreas como financeira e política. Talvez, esteja na área política, o maior exemplo da aplicação deste sistema, pois se trata da decisão eleitoral dos países democráticos.

Em Aprendizagem de Máquina, esse sistema de voto tem três variações; 1) Voto Unânime, quando todos os classificadores base ou estimadores concordam na classificação; 2) Voto da Maioria Simples, quando uma classe recebe mais da metade dos votos; e 3) Voto Majoritário, quando uma classe recebe o maior número de votos (POLIKAR, 2006). O Voto Majoritário é utilizado por alguns classificadores, entre eles Random Forest, seção 2.9.4, e também é utilizado como técnica para regra de decisão pelos comitês ou *ensembles*, para atribuição de uma determinada classe a um objeto em análise (CASTRO et al., 2017).

Neste trabalho o Voto Majoritário é utilizado por meio do sistema de múltiplos classificadores utilizado para classificação de textos curtos sobre um problema de múltiplas classes ou categórico.

2.8.2 Validação Cruzada com k-fold

A validação cruzada, Cross-Validation, segundo (DUDA; HART; STORK, 2012), é uma abordagem empírica que testa um classificador,onde k ou m se refere ao número de partes de um Dataset D sobre as quais se treina um classificador. A motivação da validação cruzada é reduzir os erros do classificador durante o processo de treinamento, onde após cada iteração sobre uma parte do Dataset, os erros são analisados e os parâmetros do classifi-

cador são alterados. Busca-se, assim, aumentar o poder de generalização do classificador. Basicamente a técnica de validação cruzada com k-fold, assumindo $k = 10$, divide randomicamente um córpus em k partes para treinamento das quais uma parte é destinada ao teste. Normalmente, se utiliza a validação cruzada para alterar parâmetros de Rede Neural Artificial em decisão de quantas camadas deverão ser inseridas ou retiradas após uma iteração. Outro algoritmo que também se beneficia diretamente dessa validação para ajuste de parâmetros é o KNN quando se deseja encontrar o número ideal para definir o vizinho mais próximo também representado por k. (DUDA; HART; STORK, 2012).

2.8.3 Métricas

As métricas são as unidades de medida aplicadas à análise dos desempenhos dos classificadores. Em AM as métricas têm importância fundamental, porque delas surgem as certezas das aplicações das soluções dos problemas no mundo real. Em AM, as métricas F-Measure ou F1-Score, Precisão, Acurácia e Recall são constantes. As métricas se baseiam nos valores de Verdadeiro Positivo, que presenta os dados corretamente classificados, Falso Positivo, que corresponde aos dados erroneamente classificados como verdadeiros, Verdadeiro Negativo, os dados corretamente identificados como falso e Falso Negativo, os dados erroneamente classificados como falso.

	Positivo	**Negativo**
Positivo	VP	FP
Negativo	FN	VN

Tabela 1 – Matriz de confusão em um problema binário.

Fonte: Elaborada pelo autor.

Acurácia: Indica a quantidade de acertos feitos pelo modelo em relação ao todo informando assim o quanto um resultado se aproxima do real. A tabela 2 demonstra

	Positivo	**Negativo**
Positivo	5	2
Negativo	3	90

Tabela 2 – Exemplo de Acurácia em um problema binário.

Fonte: Elaborada pelo autor.

um exemplo simples da acurácia sobre um problema binário onde se obtém 95% de um conjunto de 100 exemplos. No exemplo da tabela 2 a fórmula utilizada para acurácia foi

$$\mathbf{ac} = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{2.1}$$

	Positivo	**Negativo**	**Neutro**
Positivo	VP	FP	FP
Negativo	FN	VN	FN
Neutro	FT	FT	VT

Tabela 3 – Matriz de confusão de um problema categórico

Fonte: Elaborada pelo autor.

Neste trabalho se utilizou a acurácia sobre problemas categóricos. A tabela 3 espelha as categorias utilizadas, onde é possível ver mais uma classe, a classe neutra.

Nos probelmas categóricos, a equação ideal para calcular a acurárica é demonstrada pela equação 2.2 (SOKOLOVA; LAPALME, 2009).

$$ac = \frac{\sum_{i=i}^{I} \frac{VP_i+VN_i}{VP_i+FN_i+FP_i+VN}}{I} \tag{2.2}$$

Onde I é a função indicadora se uma determinada classe foi corretamente classificada retornando 1 para positivo e 0 para negativo.

Precisão: Pode ser considerada um complemento da acurácia, porque responde a questão do quão preciso nas classificações é um modelo. Em linhas práticas, ela responde a questão: De um córpus composto por tweets cuja classificação da quantidade q como positivo, quantos deles, de fato, são positivos? Na Figura 10 é possível verificar o intervalo ocupado pela acurácia e dentro dele a região aonde a precisão é identificada.De acordo com SOKOLOVA; LAPALME, a equação 2.3 é ideal para calcular a precisão de um sistema de classificação de múltiplas classes.

$$p = \frac{\sum_{i=1}^{l} \frac{VP_i}{VP_i+FP_i}}{I} \tag{2.3}$$

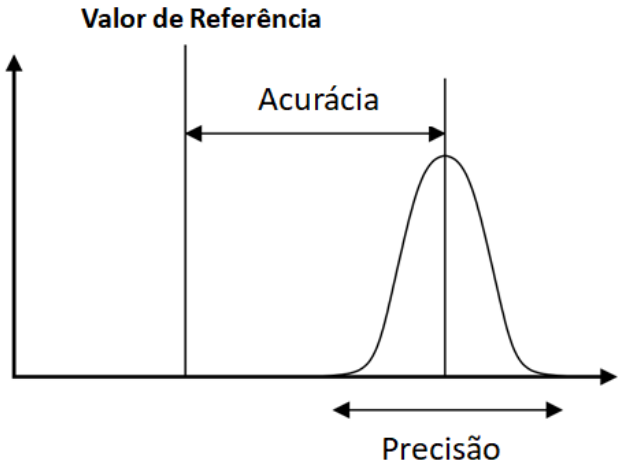


Figura 10 – Relação entre acurácia e precisão

Fonte: Elaborada pelo autor.

Recall: Para recall a definição de SOKOLOVA; LAPALME diz que se trata do número de exemplos positivos corretamente classificados dividido pelo número de exemplos positivos no córpus.

$$r = \frac{\sum_{i=1}^{l} \frac{VP_i}{VP_i + FN_i}}{I} \tag{2.4}$$

F-Measure: Combina precisão e recall numa média harmônica. Dessa forma, calcula-se recall e precisão numa só métrica. A F-Measure, então entrega, de forma mais precisa, os resultados referentes aos acertos dos classificadores e não somente dispõe a indicação de estar dentro da área de aceitabilidade de um resultado, como se obtém na acurácia. Em um sistema de classificação de múltiplas classes essa métrica é calculada conforme a equação 2.5 (SOKOLOVA; LAPALME, 2009) onde p está para precisão, r para recall e β abastrai a expressão $1^2 + 1$ que corresponde ao valor 2 padrão da fórmula.

$$f = \frac{(\beta^2 + 1)p.r}{\beta^2 p + r} \tag{2.5}$$

Brier Score: Essa métrica complementa a acurácia quanto a tomada de decisão sobre desempenho de classificadores em atividades de classificação. De acordo com GUL et al., o Brier Score, entre outras aplicações, pode ser utilizado para medir a capacidade de previsão de um classificador. GUL et al. utilizaram esta métrica para selecionar classificadores básicos para compor um *ensemble*. Para isso, considerou-se a regra de quanto menor o valor da probabilidade da perda, melhor é o classificador. Com esse resultado, se pode assumir que um classificador é um bom candidato a compor um *ensemble*.

2.9 MODELOS DE CLASSIFICADORES

Em AM existem as atividades de classificação, regressão e clusterização. Para atividades de classificação há vários algoritmos como Support Vector Machine, Multilayer Perceptron, Naïve Bayes, K-Nearest Neighbor entre outros; ao grupo de regressão, entre outros classificadores, está o algoritmo Regressor Logístico que implementa a função logística, e portanto chamado de Regressão Logística. Nesta seção são apresentados os modelos de classificadores utilizados nesta pesquisa.

2.9.1 Support Vector Machine

Este algoritmo foi desenvolvido por CORTES; VAPNIK para realizar problemas linearmente separáveis. Vários trabalhos relatam a sua aplicação em problemas binários como TELES; SANTOS; SOUZA e AVANÇO; BRUM; NUNES. Igualmente aos outros modelos baseados em Aprendizagem de Máquina, o SVM também depende do pré-processamento dos dados para representar padrões em alta dimensão. O objetivo desse modelo durante o processo de classificação consiste na busca do melhor hiperplano. Neste caso, o melhor hiperplano é aquele que tem a maior distância em relação ao mais próximo do padrão de treinamento.

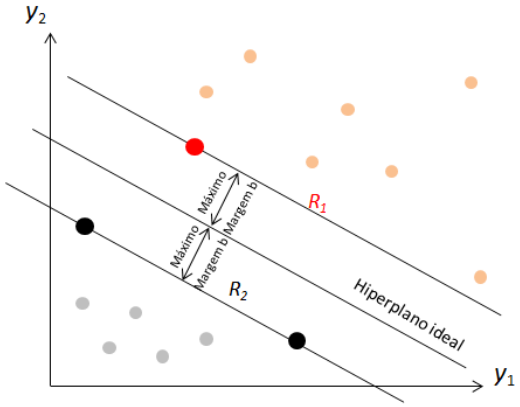


Figura 11 – Treinamento da Support Vector Machine (SVM)

Fonte: Reprodução de (DUDA; HART; STORK, 2012)

Na Figura 11, os Support Vectors, vetores de suporte que dão origem ao nome do classificador, são os exemplos do treinamento que definem o hiperplano de separação ideal e, segundo DUDA; HART; STORK, além de ser os padrões mais difíceis de classificar, eles também são os mais significativos para classificação.

O SVM também consegue lidar com problemas não-lineares, os problemas XOR, ao projetar o problema em um espaço de multidimensional por meio de uma função discriminante. A Figura 12 mostra um exemplo da função discriminante, onde à esquerda está o problema linearmente não separável, cujo o hiperplano ideal encontrado

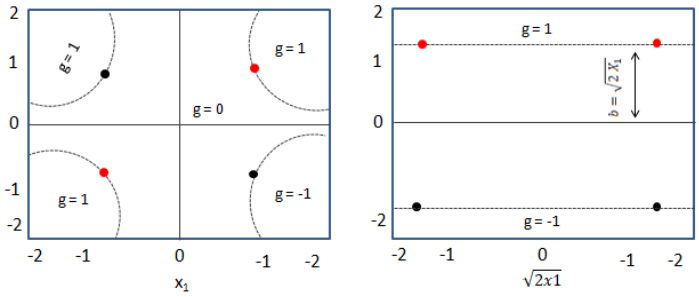


Figura 12 – Treinamento da Support Vector Machine (SVM) resulução do problema XOR

Fonte: Reprodução de (DUDA; HART; STORK, 2012)

é $g(x1, x2) = x1x2 = 0$, e a margem b é representada por $b = \sqrt{2}$, e à direita o resultado da equação é representado por $g(x) = g(x1, x2) = x1, x2$ tendo a o hiperplano $g = 0$. Nesta pesquisa SVM foi utilizado devido à sua capacidade de resolver problemas categóricos. Nos problemas binários, TELES; SANTOS; SOUZA utilizaram SVM para classificar textos entre positivo e negativo, e ALVES et al. aplicaram sobre o problema de múltiplas classes, onde envolve o problema do tipo XOR, para classificar mensagens entre positivas, negativas e neutras.

2.9.2 Multilayer Perceptron

O classificador Multilayer Perceptron é uma Rede Neural Artificial do tipo feedforward, comumente Feedforward Neural Network. Essa Rede Neural Artificial (RNA) contém pelo menos 3 camadas cujas conexões não são cíclicas, como as conhecidas Rede Neural Recorrente, do inglês, Recurrent Neural Network, Figura 13. Assim como o modelo SVM, o Multilayer Perceptron também implementa função discriminante linear para resolver problemas categóricos. Nesta pesquisa optou-se por aplicar o Multilayer Perceptron sobre o conjunto da dados em português do Brasil devido à sua capacidade de mapear grandes espaços de características. Essa capacidade ajuda a reduzir a região de ambiguidade entre as classes (DUDA; HART; STORK, 2012). Alguns trabalhos recentes relatam bons desempenhos aplicando o Multilayer Perceptron como (BRUM, 2018), LOPES et al. e (MONTEIRO et al., 2018). A Figura 13 [9] mostra um exemplo de uma RNA Feedforward à esquerda, e uma Rede Neural Recorrente.

2.9.3 Regressão Logística

Regressão Logística é um modelo estatístico utilizado pela área de AM inicialmente para lidar com problemas binários. Neste trabalho aplicou-se a Regressão Logística Multinomial para encontrar a probabilidade de cada classe a partir do texto. Essa classificação

[9] Disponível em <<https://missinglink.ai/guides/neural-network-concepts/recurrent-neural-network-glossary-uses-types-basic-structure>>. Acessado em 05 de julho de 2019

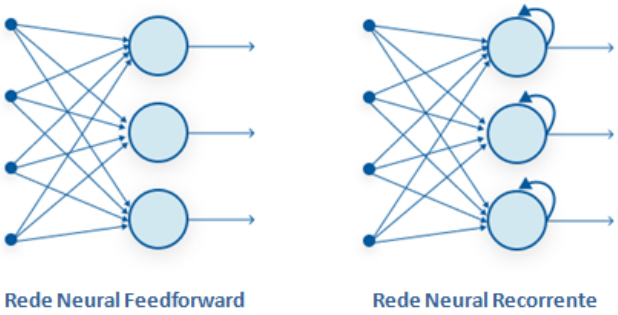


Figura 13 – Exemplo de uma Rede Neural Feedforward

Fonte: Elaborada pelo autor.

utiliza a função softmax. Outros trabalhos aplicam esse classificador para problemas categóricos como (VITÓRIO et al., 2017) e (NICULAE et al., 2014) utilizaram para classificação de textos português em múltiplas classes.

2.9.4 Random Forest

O Random Forest é um classificador que utiliza o sistema de técnicas de comitê (*ensemble*) para atividades de classificação ou regressão por meio de vários tipos de Árvores de Decisão (Decision Tree). No Random Forest cada Árvore de Decisão é desenvolvida durante do processo de treinamento sob diferentes tipos de randomização. A previsão das classes é dada pelo Voto Majoritário onde a classe que recebe o maior número de votos é atribuída ao objeto a ser classificado. As Árvores de Decisão, por sua vez, são algoritmos não-métricos[10], isto é, que utilizam valores nominais. DUDA; HART; STORK descrevem um valor nominal como como "descrições que são discretas e sem qualquer noção natural de similaridade ou mesmo ordenação". Problemas com esses tipos de valores não podem ser resolvidos por algoritmos que reconhecem padrões por meio de vetores de números reais. Para atividades de classificação de textos, Random Forest foi utilizado por MONTEIRO et al.. A Figura 14[11]

2.9.5 Gaussian Naïve Bayes

Gaussian Naïve Bayes é um classificador que assume que as características informadas se assemelham a uma distribuição gaussiana (REGALADO et al., 2018). Esse classificador é uma variação do algoritmo probabilístico Naïve Bayes. O GaussianNB é diferente dos outros modelos de algoritmos por ser capaz de processar a entrada de dados contínuos. Quando aplicado em atividades de classificação de textos em inglês, esse classificador

[10] Algo que não tem medida ou valor métrico.

[11] Disponível em <<https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>>. Acessado em 06 de julho de 2019

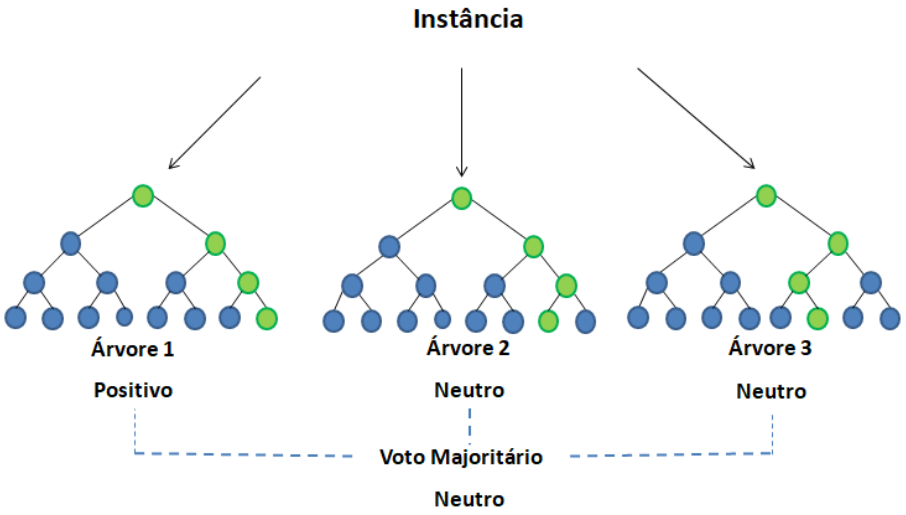


Figura 14 – Exemplo de uma Random Forest

Fonte: Adaptada de (KOEHRSEN, 2017)

mostrou bons resultados como reportado por (GOKHALE; FASLI, 2018). Alguns trabalhos aplicaram esse classificador em atividades de classificação de textos em português do Brasil como (TELES; SANTOS; SOUZA, 2016), (AVANÇO; BRUM; NUNES, 2016).

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{X}) = \frac{p(X|\omega_i).P(\omega_i)}{\sum_{j=1}^{c} p(x|\omega_j).P(\omega_j)} \tag{2.6}$$

$$g_i(\mathbf{X}) = p(\mathbf{X}|\omega_i).P(\omega_i) \tag{2.7}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{X}|\omega_i) + \ln P(\omega_i) \tag{2.8}$$

As equações 2.6, 2.7 e 2.8 demonstram a aplicação da função discriminante para problemas de classificação para múltiplas classes utilizadas por este algoritmo (DUDA; HART; STORK, 2012).

2.9.6 K-Nearest Neighbor (KNN)

O K-Nearest Neighbor, em português, Vizinho mais Próximo, é um método de classificação no qual uma atribuição de uma classe é feita levando-se em consideração seus vizinhos mais próximos. A métrica de proximidade utilizada pelo KNN normalmente é o cálculo da distância euclidiana. A parte do aprendizado consiste no armazenamento simples dos exemplos de treinamento, que também é chamado de lazy learning. A definição final da classe ocorre por meio do Voto Majoritário como pode ser visto na Figura 15. Embora ROSA; CARVALHO; BATISTA tenham afirmado que o KNN é um dos algoritmos mais usados para classificação de textos, poucos trabalhos foram encontrados utilizando esse classificador sobre textos em português. Na Figura 15 se o algoritmo inicia a busca no

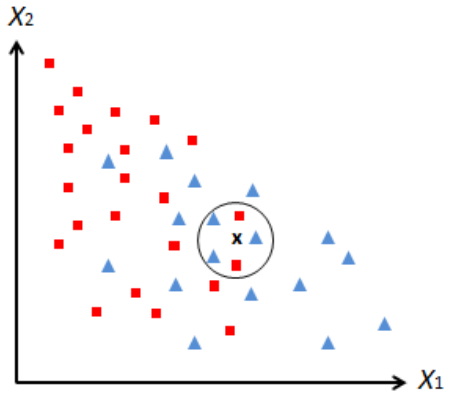


Figura 15 – Exemplo de classificação com KNN onde K=5

Fonte: Elaborada pelo autor.

ponto **x** e utiliza o voto majoritário simples para atribuir a classe ao objeto representado por **x**. Neste exemplo, a classe dos triângulos azuis é atribuída a **x**.

2.9.7 Stochastic Gradient Descent

O "Método" Gradiente Descendente Estocástico, reúne duas funções importantes para atividades de classificação e regressão. Em AS, principalmente, esse algoritmo tem sido utilizado em diversos trabalhos. Uma das suas funções que motiva sua aplicação é a forma randômica que o dado de treinamento é selecionado. Isto é, dentre os exemplos de treinamento aplicados ao SGD, a busca ainda é aleatória. A sua principal função é a função de minimização representada pela Figura 16[12]. Essa função é aplicada sobre vetores esparsos e de grandes dimensões.Nas suas configurações é possível definir as funções que influenciarão na atividade do algoritmo, se ele atuará como uma função de regressão ou de classificação.

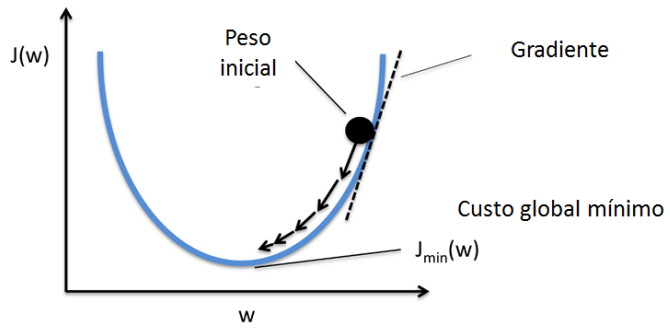


Figura 16 – Gráfico da Função Gradiente Descendente

Fonte: Adaptada de SURYANSH

[12] Disponível em <<https://hackernoon.com/gradient-descent-aynk-7cbe95a778da>>. Acessado em 12 de julho de 2019

2.9.8 *Ensemble*

Em AM, *ensemble* é um comitê de classificadores, chamados de classificadores base (base classifiers), utilizado para obter melhor desempenho em acurácia. HAN; PEI; KAMBER define esse sistema de múltiplos classificadores como "um modelo composto que consiste na combinação de classificadores". Assim, os *ensembles*, tendem a ser mais performáticos que os modelos isolados. Isso ocorre porque o resultado final de associação de uma classe c a um dado d é definido por voto majoritário. Nesse sistema de votação se procura evitar erros de classificação. Ainda de acordo com HAN; PEI; KAMBER, o *ensemble* erra na classifcação se a maioria dos classificadores base do comitê cometerem erros. Por isso, aconselha-se a utilizar diferentes paradigmas de modelos para construir um *ensemble*. O ideal é haver o mínimo de relação possível entre os classificadores base. Aconselha-se evitar de incluir, por exemplo, em um conjunto de três classificadores base os modelos que utilizam o mesmo algoritmo de classificação como (KNN, K-Means, Multilayer Perceptron) onde KNN e K-Means utilizam a distância euclidiana para classificar objetos no espaço, dessa forma se tem um comitê heterogênio. Outro fator importante para composição de *ensembles* é a diversidade, que pode ser obtida a partir do conjunto de dados. Há algumas técnicas para se obter diversidade, e uma delas é aplicação da técnica de treinamento com validação cruzada utilizada neste trabalho, ou através de técnicas de criação de dados como a reamostragem (resempling). Vários trabalhos relatam a aplicação de *ensemble* em AS para diversos contextos. Em português do Brasil há os trabalhos de AVANÇO; BRUM; NUNES e CASTRO et al.. Na linha de classificação de textos curtos em português, nesta pesquisa se aplicou *ensemble* efetivamente para classificar textos em português do Brasil sobre o problema de múltiplas classes.

2.10 ANÁLISES ESTATÍSTICAS

A análise estatística conta com o processo de teste de hipóteses sobre os dados, afim de assegurar que eles provém da mesma distribuição. O teste de hipótese, segundo COUCH et al., verifica se a medida de um determinado conjunto está de acordo com uma teoria. O teste de hipóse se baseia sobre duas variáveis:

H_0: hipótese nula ou da existência onde se baseia a referida teoria.

H_1: hipótese alternativa.

O teste de hipóse pode ser elaborado conforme recomenda MORETTIN:

- Definir das hipóteses nula, representada por H_0, e alternativa H_1.
- Fixar o nível de significância α.
- Obter de uma amostra de tamanho n e estimar o parâmetro θ.

- Calcular o valor do parâmetro θ_0, dado por H_0, calcular também o valor crítico, o valor observado $V_{calculado}$.
- Definir a região de não rejeição da hipótese nula e a da não rejeição.
- Se $V_{calculado}$ estiver dentro da região de não rejeição de H_0, então considera-se H_0. Caso contrário rejeita-se H_0

2.10.1 Teste de Shapiro-Wilk

Esse teste é aplicado para verificar se os dados pertencem a uma distribuição gaussiana, normalmente conhecida como distribuição normal. Esse teste é importante para decidir qual o tipo de teste será aplicado sobre os dados. Então, as hipóteses nula H_0 e alternativa H_1, neste teste, indicam que os dados pertencem a uma distribuição normal e não pertencem a uma distribuição normal respectivamente.

O nível de significância é importante para os testes de hipóteses para indicar o grau de confiabilidade dos testes. Por isso, no teste de Shapiro-Wilk essa definição também é importante. O nível de significância é estabelecido pela definição de α, onde normalmente se assume 0.05. Esse valor representa 95% de confiança.

2.10.2 Teste t-Student

Este teste é aplicado sobre distribuições normais com a finalidade de comparar se há diferença significativa entre as amostras em análise. Neste teste se define as hipóteses conforme abaixo: define-se a hipótese nula H_0:$\mu = \mu_0$ onde μ representa a média das amostras. Já a hipótese alternativa H_1, neste tipo de teste, pode ser unilateral à direita, unilateral à esquerda, e bilateral.

H_1: $\mu \neq \mu_0$ teste bilateral.
H_1:$\mu > \mu_0$ teste unilateral à direita.
H_1: $\mu < \mu_0$ teste unilateral ‘a esquerda.

2.10.3 Teste de Wilcoxon

Este teste é aplicado sobre amostras que não obedecem uma distribuição normal. Por isso, em alguns casos, aplica-se este teste como uma alternativa ao teste t-Student. Ainda de acordo com COUCH et al., para aplicar o teste de Wilcoxon não é necessário ter conhecimento da distribuição dos dados. Nem mesmo se eles provém de uma distribuição normal.

Neste trabalho os testes de Shapiro-Wilk, t-Student (bilateral) e o de Wilcoxon foram aplicados sobre os resultados dos desempenhos em acurácia dos classificadores clássicos em execução isolada, e também sobre o *ensemble* desses classificadores.

3 TRABALHOS RELACIONADOS

Nos últimos anos, o idioma português despertou o interesse dos pesquisadores da área de PLN devido ao grau de dificuldade que ele impõe às atividades de Análise de Sentimentos. A dificuldade de lidar com o português está na preservação da sua estrutura, por possuir várias conjugações verbais e diferentes estruturas morfológicas. CASTRO et al. relatam de forma sucinta que o português é um idioma pluricêntrico que mantém variedades que dificultam o processo de reconhecimento do idioma por algumas ferramentas.

Em Análise de Sentimentos (AS) há inúmeros trabalhos que relatam atividades de processamento e classificação de texto em vários idiomas. Mas no idioma português, embora haja avanços nas pesquisas, ainda há carência de recursos como dicionário de léxicos, corpora, palavras etiquetadas com suas classes gramaticais (taggers), etc (ZAMPIERI; BECKER, 2013). Contudo, alguns trabalhos relatam progresso nas atividades de AS aplicando as mais variadas técnicas de classificação e processamento.

Neste capítulo são apresentados os trabalhos desenvolvidos sobre córpus em português do Brasil aplicado às atividades de Análise de Sentimentos (AS). Nessa seção se contribui para que todos tenham informações sobre onde buscar por córpus em português.

3.1 ANÁLISE DE SENTIMENTO EM PORTUGUÊS

ZAMPIERI; MALMASI; DRAS investiram os esforços em atividades sobre vários tipos de português ao longo do tempo. No trabalho, se buscou, segundo os autores, modelar alterações históricas no português. E para isso, foi necessário utilizar um córpus de textos literários para o objetivo de classificação temporal de textos. Os autores trataram as variações léxicas aplicando a técnica de n-gram. Outra variação da língua, a semântica, foi representada por etiquetamento por partes de palavras, do inglês, POS, largamente expressado como POS-Tagger ou POS-tagging. As características das palavras foram obtidas por uni-gramas e foram aplicadas como entrada para o classificador SVM. O resultado obtido foi de 99,98% de acurácia na atividade de prever o ano da publicação dos documentos no intervalo de tempo entre um e meio século. Outros experimentos seguiram aplicando POS-tagger como extração de características para obtenção de outros resultados. Os autores desenvolveram e testaram os algoritmos sobre parte do córpus de texto histórico em português chamado Colonia[1], que contém textos que abrangem do século XVI ao início do século XX. De acordo com ZAMPIERI; BECKER, o córpus Colonia é o maior córpus histórico em português que se tem conhecimento. Contudo, nesse córpus há escassez de documentos referentes aos séculos iniciais variando de 13 documentos do século XVI a 38 documentos do século XIX. Essa quantidade de documentos é insuficiente para atividades

[1] <<http://corporavm.uni-koeln.de/colonia/index.html>>. Acessado em 17 de junho de 2019

de classificação de textos utilizando métodos de Aprendizagem de Máquina (AM). Então, para mitigar o problema de escassez de documentos históricos para treinamento, os autores aplicaram a técnica de composição de documentos. Essa técnica é feita de sentenças de outros documentos da mesma classe (período de tempo em que o documento foi escrito). Dessa forma, foi possível criar 330 tokens que deram origem a novos documentos para teste. ZAMPIERI; MALMASI; DRAS afirmam que nessa técnica se garante que os textos são misturados em diferentes estilos sem perder as características da época, evitando, então, que o "recém-composto" texto seja classificado em outro período. Os resultados reportados em acurácia foram validados com a técnica de validação cruzada com k-fold onde $k = 10$. Após a validação, os resultados mostraram que a proposta de ZAMPIERI; MALMASI; DRAS é capaz de prever as datas de publicação dos documentos em intervalos de cem e cinquenta anos. Segundo os autores, ao utilizar uni-gramas como características, a performance chegou a 99,8% de acurácia; da mesma forma as datas puderam ser previstas utilizando apenas POS-tagger onde a performance atingiu 90,7% para intervalos de cem anos, e 90,1% para intervalos de cinquenta anos.

CASTRO; SOUZA; OLIVEIRA apresentaram a técnica de suavização, do inglês smoothing, juntamente com a técnica de n-gram para classificar tweets no idioma português. Entre as diversas configurações de n-gram com suavização aplicadas ao experimento, o melhor desempenho atingido, no sentido de acurácia, foi 92,7% sobre o problema de identificar a qual português uma variação corresponde. Para atingir esse resultado, os autores aplicaram *ensemble* das técnicas Lindstone (0.1), 6-gram, Good-Turing de uni-grama de palavras. A técnica de suavização é bem explorada pelo MiningBR Research Group[2] quando aplicaram a mesma técnica no trabalho "Uma análise comparativa de técnicas supervisionadas para mineração de opinião de consumidores brasileiros no twitter" para obter melhores resultados nos experimentos de classifcação de tweets (TELES; SANTOS; SOUZA, 2016). O MiningBR Research Group se dedica à Mineração de textos em português contribuindo ativamente na produção de conhecimento na área de PLN e também na construção de córpus para português nas variantes brasileira e européia.

[2] <<http://miningbrgroup.com.br/>>. Acessado em 18 de junho de 2019

CASTRO et al. estenderam o trabalho Discriminating between Brazilian and European Portuguese national varieties on Twitter texts (CASTRO; SOUZA; OLIVEIRA, 2016) ao aplicar as técnicas de suavização e n-gram para classificar textos em português considerando as veriações do português brasileiro e o europeu. Os autores obtiveram o resultado de 92,71% de acurácia. Nessa extensão, a pesquisa foi direcionada à identificação do idioma português em tweets considerando as variações entre o português brasileiro e o europeu (CASTRO et al., 2017). A pesquisa se concentrou na aplicação e análise de desempenho das várias técnicas de suavização, do inglês smoothing, juntamente com n-gram para identificar textos nas duas variações, resultando em um extenso trabalho. Os experimentos foram feitos sobre textos curtos originados da rede social Twitter. Além do bom resultado, os autores reportam a aplicação do TF-IDF com esquema de pesos para extração de características na fase de pré-processamento de tweets. Os autores também relataram a importância da aplicação do SVM linear para identificar em qual das variantes do português uma determinada mensagem pertence.

FRANÇA; OLIVEIRA investigou a classificação de textos em português do Brasil a fim de obter o resultado satisfatório em relação à capacidade humana de avaliação da subjetividade que chega, segundo WIEBE; WILSON; CARDIE, a 72%. O trabalho se concentrou na extração de tweets referentes às manifestações ocorridas no ano de 2013 no Brasil. As mensagens foram rotuladas manualmente delimitando o tipo de problema binário. Para a atividade de classificação, o classificador probabilístico Naïve Bayes foi utilizado sobre os dados obtendo acurácia de 90% na classficação das mensagens do córpus de mensagens positivas e 72% no córpus de mensagens negativas. Outras métricas como Recall, Precisão, F1-Score também foram utilizadas. A Precisão, segundo o autor, foi de 79% e 85% para o corpora postivo e negativo respectivamente. A análise da precisão é fundamental para descobrir do universo de mensagens avaliadas como positivas, de fato, pertencem à tal classe. No trabalho de FRANÇA; OLIVEIRA verificou-se que o classificador foi mais preciso sobre as mensagens negativas.

DOSCIATTI; FERREIRA; PARAISO relataram o processo de construção de um córpus de notícias em português do Brasil para Análise de Sentimentos (AS). No trabalho foram consideradas as emoções básicas alegria, tristeza, raiva, medo, repugnância e surpresa definidas por EKMAN, e a sua construção foi composta por fragmentos de textos em linguagem formal. O processo de anotação do córpus contou com três anotadores onde todos os textos do córpus foi anotado por dois anotadores diferentes ficando o terceiro anotador para atuar em caso de divergência. Para validar a qualidade das anotações feitas pelos anotadores, os autores aplicaram o coeficiente Kappa (COHEN, 1960) para medir o grau de concordância. Segundo os autores, o limite de aceitabilidade de um córpus anotado manualmente deve ser o coeficiente kappa superior a 0,67, e as anotações de qualidade

devem superar 0,80. No trabalho, DOSCIATTI; FERREIRA; PARAISO reportam um coeficiente kappa 0,38. Um índice muito abaixo do sugerido por KRIPPENDORFF. Na pesquisa ainda se buscou aplicar a anotação automática dos textos utilizando o método de AM para classificação utilizando Support Vector Machine sobre um conjunto de mensagens cujas anotações tinham boa concordância. Nesse experimento os autores obtiveram acurácia de 60,3%.

FREITAS descreveu o processo de construção de um léxico de textos afetivos para auxiliar nas atividades de Processamento de Linguagem Natural e principalmente nas tarefas de Análise de Sentimentos. Segundo a autora, os léxicos são fundamentais para atividades de extração de conhecimento, principalmente da subjetividade presente nos textos, pois não há processamento de informação sem informação prévia disponível por meio dos léxicos. O trabalho se baseou no córpus de resenhas de livros anotado com informação de opinião (FREITAS et al., 2013). No trabalho, FREITAS explorou com detalhes os diferentes significados que algumas palavras têm e qual a influência delas na definição de polaridades. Como ocorre em vários trabalhos da literatura de PLN, o léxico foi anotado manualmente, pois o objetivo era eliminar as palavras que, apenas em contexto restrito, simbolizavam opinião. O objetivo dessa eliminação foi a construção de um léxico enxuto onde ambiguidades referentes ao emprego de verbos e alguns adjetivos não sejam possíveis. Sobre os adjetivos, FREITAS esclareceu que nem todos são importantes para um léxico devido ao seu comportamento flutuante que pode depender de um domínio específico ou de um objeto. O autor citou o exemplo do adjetivo "*pequeno*" que pode ser, para um caso, positivo se relacionado à mensagem "*celular pequeno*", mas também pode ser de caráter negativo na sentença "*memória pequena*". Em outro caso, essas polaridades podem ser contrárias. Onde "memória pequena" pode ser positivo para um consumidor, da mesma forma que um "celular pequeno" pode ser negativo.

No trabalho de TELES; SANTOS; SOUZA se avaliou as técnicas de Aprendizagem de Máquina para avaliação de opinião de consumidores extraída do Twitter. Os algoritmos utilizados foram SVM, Regressão Logística e o algoritmo probabilístico Naïve Bayes para classificar tweets sobre um córpus com classes positiva e negativa, e outro córpus com as classes positiva, negativa e neutra. Os autores relatam 95% de precisão e acurácia de 94,87% sobre dados balanceados utilizando SVM sobre o córpus com duas classes, e 82,73% sobre o córpus com três classes. Os demais resultados foram tabelados e exibem os algoritmos Naïve Bayes e SVM com 82% e 84% de acurácia, e F-Measure de 0,819 e 0,821 respectivamente. O algoritmo de Regressão Logística, segundo os autores, teve boa performance juntamente com a técnica de suavização obtendo 90,57% de acurácia e 90,44% de precisão. O córpus utilizado no trabalho é composto por 2940 tweets extraídos do córpus desenvolvido pelo MiningBR Research Group. TELES; SANTOS; SOUZA aplicaram

no trabalho a técnica de suavização para maximizar a precisão.

AVANÇO; BRUM; NUNES aplicaram *ensemble* para classificar opiniões extraídas de textos dos ambientes de compras on-line Mercado Livre[3] e do córpus originado do BuscaPé[4] disponibilizado por HARTMANN et al.. Ambos corpora foram utilizados para aplicação da técnica de classificação baseada em léxico e em AM. Os classificadores baseados em léxico foram: a) Baseline, um método simples que apenas combina polaridades de palavras de um léxico de sentimento; b) Baseline-VSM, Linha de Base com Modelo de Espaço Vetorial para prever palavras que não estão no léxico; c) LBC-VSM, uma adaptação do modelo de espaço vetorial definidor por AVANÇO; NUNES. Os autores reportam F-Measures de 84% na técnica de classificação baseada em léxico e 95% como resultado da técnica de classificação utilizando métodos de AM.

JÚNIOR L. BARBOSA aplicaram a técnica de aprendizagem profunda, do inglês, Deep Learning, por meio da Rede Neural Artificial do tipo Long Short-Term Memory criando uma Convolutional Neural Network, comumente definidas como LSTM-CNN, para classificar entidades nomeadas em corpora de língua portuguêsa na variante do Brasil. Segundo os autores, as redes neurais profundas aprendem relações semânticas, morfológicas e sintáticas o que permite classificar melhor a entidade das palavras. Os experimentos foram feitos sobre quatro diferentes corpora sendo um deles dedicado ao treinamento dos modelos e os três restantes utilizados para testes. Os autores desenvolveram os seguintes modelos: ParamopamaCWNN, uma arquitetura de rede neural profunda, uma Deep Learning, que recebe Word Embeddings e Char Embeddings na sua camada de entrada; e também um classificador baseado em Conditional Random Forest, denominado pelos autores com ParamopamaCRF. Paramopama se refere ao córpus disponibilizado por JÚNIOR C. M.; BARBOSA. As métricas comuns na literatura de PLN como precisão, Recall e F-Measure foram aplicadas sobre os cenários de testes. No trabalho, se concluiu que é recomendável utilizar Deep Learning para classificação de entidades nomeadas. Segundo os autores, houve bom desempenho sem necessidade de definir manualmente as características.

BRUM propôs uma abordagem semi-supervisionada para anotação de um córpus a partir de um córpus anotado manualmente. No trabalho, BRUM utilizou o córpus TweetSentBR de domínio televisivo. O córpus, composto por três classes positiva, negativa e neutra, foi anotado por sete anotadores. No trabalho, o autor se refere à expansão de córpus se referindo ao aumento da quantidade de mensagens rotuladas. Para executar a expansão, o autor desenvolveu um framework baseado em seis classificadores, onde cada um deles foi treinado sobre córpus manualmente rotulado. Para as atividades de AM como treinamento e validação, no trabalho lançou-se mão de bases de dados existentes na

[3] <<https://www.mercadolivre.com.br>>. Acessado em 21 de junho de 2019
[4] <<https://www.buscape.com.br>>. Acessado em 21 de junho de 2019

literatura como ReLi (FREITAS et al., 2012), Copa das Confederações no Brasil em 2013 (ALVES et al., 2014), o córpus 7x1-PT, referente à atuação da seleção brasileira contra a Alemanha na Copa do Mundo de 2014 no Brasil (MORAES; MANSSOUR; SILVEIRA, 2015), e o Computer-BR, resultado do trabalho de (MORAES ANDRÉ L. L. SANTOS, 2016). Esses corpora serviram de córpus de treinamento para os classificadores, que foram utilizados para expansão a partir de sentenças não anotadas extraídas do *Twitter*. O autor reporta o percentual de 62,14% de F-Measure média para o classificador semi-supervisionado de melhor desempenho sobre o córpus com três polaridades. BRUM também relata experimentos com córpus com polaridade binária onde se obteve 83,11% de F-Measure média sobre o córpus semi-supervisionado e 79,80% sobre o córpus anotado manualmente. O resultado mais expressivo do trabalho foi relatado numa F-Measure que atingiu 93,15% com dados binários sobre um córpus de avaliação (review) de produtos.

VITÓRIO et al. se dedicaram às atividades de análise dos resultados da Análise de Sentimentos quanto à classificação de textos curtos nas variações brasileira e européia do português. Durante o trabalho, os autores construíram o córpus 2000-tweets-br (VITÓRIO et al., 2017). Os autores identificaram que essas variações influenciam nos resultados de classifição em caso de treinamento e classificação sobre diferentes corpora.

HARTMANN et al. avaliaram diferentes modelos de word embeddings utilizados na literatura como GloVe, Word2Vec, Wang2Vec, e o FastText. Os word embeddings foram criados a partir de textos em português considerando as variantes do português brasileiro e o europeu. O objetivo do trabalho foi de identificar quais dos bons modelos sobre analogias morfo-sintáticas terá melhor performance sobre a técnica de POS-Tagging. Os autores reportam a extração de diversos corpora de diferentes autores. Entre os modelos avaliados, segundo os autores, o Wang2Vec pareceu ser o mais robusto em termos de acurácia (90,94%) utilizando o algoritmo *skip-gram* no vetor de mil posições. Além dos resultados, os autores também disponibilizaram no repostório do Núcleo Interinstitucional de Linguística Computacional[5] todos os word embeddings pré-treinados.

PAGANO; PARAISO desenvolveram um trabalho analítico trazendo importantes questões sobre as atividades de AS quanto à eficácia das representações gramaticais. No estudo, segundo os autores, as atividades de classificação de textos utilizando métodos de AM são ad-hoc por não estarem pautadas em teorias linguísticas que justifiquem a classificação da emoção. Essa afirmação pode ser justificada ao considerar o objetivo do trabalho que está na exploração de uma metodologia para Análise de Sentimentos utilizando a teoria da linguística. Os autores relatam que textos jornalísticos, por exemplo, não são utilizados para atividades de AS, porque eles não contém informações (léxicas) que indicam uma

[5] <<http://nilc.icmc.usp.br/embeddings>>. Acessado em 30 de junho de 2019

associação de afetividade. No exemplo, “Andorinhas mudam rotina em cidade paraense: Elas chegam a Parauapebas e dão espetáculo no céu. Entretanto, sujeira deixada pelas aves incomoda moradores.”, os autores afirmam que os métodos de AM aplicados à Análise de Sentimentos (AS) deveriam identificar os indicadores linguísticos das emoções. No caso da mensagem de exemplo, os autores relatam a emoção *surpresa*. A mensagem, se analisada humanamente, passa a ter vários identificadores de emções, e talvez poderiam ser avaliados por pesos a partir de um léxico. PAGANO; PARAISO classificaram a mensagem da seguinte maneira: “Andorinhas mudam rotina em cidade paraense (*surpresa*): Elas chegam a Parauapebas e dão espetáculo no céu (*alegria*). Entretanto, sujeira deixada pelas aves incomoda moradores (*repugnância*).”. Os autores fizeram os experimentos sobre o córpus de notícias anotado por DOSCIATTI; FERREIRA; PARAISO por meio das técnicas de AM. O experimento utilizou 10% do córpus onde os anotadores tiveram condordância total. Os autores extraíram as frequências e as relações das categorias, sendo elas as emoções difinidas por EKMAN, onde foi possível identificar padrões candidatos a informar algoritmos para Análise de Sentimentos (AS). Os resultados obtidos indicam, de acordo com os autores, alta frequência para emoção raiva nos seguintes tipos de processos: material (60,4%), como na mensagem, "Garota de 14 anos **engravidou** do próprio pai em Itariri, no interior de SP.”; verbal (17%), como em, “Mãe **confessa** [ter matado recém-nascido à tesouradas.].”; e Relacional Atributivo (13,25%), como na mensagem, “Filhos são suspeitos [de abandonar pai idoso]”. Os autores concluem que inicialmente encontraram características de emocionais nos textos rotulados. Essas características podem ser utilizadas juntamente com Análise de Sentimentos (AS). O trabalho de PAGANO; PARAISO permite sugerir aplicação da técnica de POS-tagging juntamente com o trabalho apresentado a fim de obter uma classificação baseada em léxico.

3.2 CORPORA EM PORTUGUÊS

Existe cada vez mais trabalhos relatando construção de córpus para as mais variadas atividades de Análise de Sentimentos (AS). Nos últimos anos, alguns pesquisadores relataram a escassez de bases de treinamento para atividades de classificação de textos baseada em métodos de Aprendizagem de Máquina (AM). DOSCIATTI; FERREIRA; PARAISO, por exemplo, afirmaram que a literatura ainda não é farta quanto à produção de córpus para o idioma português Brasileiro para atividades de Análise de Sentimentos (AS). E dois anos antes, ZAMPIERI; BECKER viram que o número de corpora em português vem se desenvolvendo nos últimos vinte anos e a maioria deles está disponível on-line.

BRUM; NUNES construíram o TweetSentBR[6], um córpus de sentimentos a partir de mensagens curtas extraídas da rede social Twitter para atividades de AS. No trabalho, os

[6] <<https://bitbucket.org/HBrum/tweetsentbr>>.Acessado em 27 de junho de 2019

autores utilizaram o processo de anotação manual onde 15.000 tweets foram rotulados. No processo de anotação foram utilizados 7 anotadores para identificar a polaridade das mensagens entre as classes positiva, negativa e neutra. Na contrução, os autores aplicaram os métodos de AM para testar a credibilidade das mensagens manualmente anotadas. Para isso, utilizaram as métricas acurácia e F-Measure que resultaram em 82,06% e 80,99% respectivamente.

VITÓRIO et al. elaborou o córpus 2000-tweets-br como atividade da pesquisa do MiningBR Research Group[7] e conta com 2000 mensagens em português do Brasil extraídas da rede social Twitter[8]. Esse córpus foi rotulado manualmente em quatro classes positiva, negativa, neutra e ambas. E por ser anotado manualmente, o referido córpus contou com três anotadores onde todos anotaram a mesma mensagem obtendo o grau de 54,25% de concordância entre anotadores . O grau de concordância entre os anotadores é definido pelo coeficiente Kappa (FLEISS, 1971).

MORAES; MANSSOUR; SILVEIRA criaram o córpus 7x1-PT referente à partida de futebol masculino ocorrida em 2014 entre Brasil e Alemanha durante a copa do mundo no Brasil. No trabalho, os autores relatam o processo de coleta e anotação do córpus. Os tweets presentes no córpus 7x1-PT são parte da base de dados WorldCupBrazil2014. Essa base contém 851.292 tweets coletados nos idiomas português, inglês e espanhol para posterior processamento e separação em um córpus menor de interesse dos autores. No trabalho, os autores relatam o processo de limpeza e nomalização das mensagens devido à forte presença de caracteres especiais característicos de CGU na web. Segundo os autores o maior desafio nesse sentido foi o tratamento das hashtags, porque elas normalmente representavam sujeitos nos tweets. Sabe-se que manter caracteres especiais em mensagens resulta em ruídos, prejudicando assim, o processo de classificação de textos curtos. Por esse motivo, os autores relatam a remoção do caractere cerquilha, símbolo das hashtags, das mensagens e mantiveram todo o corpo do texto. MORAES; MANSSOUR; SILVEIRA esclarecem que as anotações foram feitas manualmente e tinham como referência a seleção brasileira. Isso signfica que as anotações positiva, negativa e neutra foram feitas tomando as mensagens direcionadas ao Brasil. Como resultado do trabalho, os autores disponibilizaram o córpus com 157 mensagens positivas (6% do córpus), 1.771 mensagens neutras (65% do córpus) e 800 mensagens negativas (29% do córpus).

ZAMPIERI; BECKER desenvolveram um recurso linguístico para o português denominado Colonia[1]: Um córpus de português histórico. Esse recurso linguístico representa uma coleção de documentos do século XVI até o início do século XX etiquetados com POS-

[7] <<http://miningbrgroup.com.br>>. Acessado em 15 de junho de 2019
[8] <<www.twitter.com>>. Acessado em 15 de junho de 2019

taggers por meio da ferramenta TreeTagger[9]. Colonia contém 5.1 milhões de tokens que se divide em cinco sub-corpora[10] separados por século. A compilação do córpus ocorreu através da coleta dos textos de diferentes fontes como: Domínio Público[11], uma biblioteca digital mantida pelo Ministério da Educação Brasileiro, e textos de outras duas fontes históricas como Grupo de Morfologia Histórica do Português (GMHP) da Universidade de São Paulo e Tycho Brahe[12]

ROCHA; SANTOS criaram o córpus linguístico jornalístico para o idioma português na variante portuguêsa denomiado CETEMPúblico[13]. O córpus criado não contém polaridades anotadas por se tratar de uma base de textos linguísticos. Para criação do córpus os autores reportaram uma sequência de atividades para o tratamento dos textos, que consistiu majoritariamente na separação do texto seguindo a regra de inserir marcações, tags, para identificar se um bloco de texto, definido como *extracto*, é uma sentença, artigo, título *etc*. Essa separação, segundo os autores, foi necessária devido ao acordo com o jornal fornecedor dos textos jornálisticos no sentido da não reprodução das matérias por meio automático. Então, em seguida, todo material coletado foi separado em palavras, tokens, para compor o córpus. O processo de separação do texto em palavras, resultou em um córpus com 180 milhões de palavras. Os autores, não consideraram o descarte de pontuações como ocorre nas atividades de pré-processamento de textos para classificação.

Córpus	**Positiva**	**Negativa**	**Neutra**	**Total**
TweetSentBR	6.648	4.426	3.926	15.000
7x1-PT	157	800	1.771	2.728
Colonia[14]	-	-	-	-
2000-tweets-br[15]	390	509	1040	1939
CETEMPúblico*	-	-	-	-

Tabela 4 – Corpora listados na seção 3.2

Fonte: Elaborada pelo autor.

De acordo com os trabalhos resumidos nesta seção, se pode afirmar que a escassez de corpora em português do Brasil para atividades de AS para métodos de AM é de domínio ou contexto específico para trabalhos em PLN. Os diversos trabalhos relatados se basearam em córpus para seus próprios domínios, isto significa que a maioria deles tinha um contexto específico: Uns se aplicavam nas avaliações de produtos ou serviços

[9] <https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html>. Acessado em 28 de junho de 2019

[10] <http://corporavm.uni-koeln.de/colonia/inventory.html>.Acessado em 28 de junho de 2019

[11] <http://www.dominiopublico.gov.br/>. Acessado em 17 de junho de 2019

[12] <http://www.tycho.iel.unicamp.br/hotsite/index.html>. Acessado em 17 de junhho de 2019

[13] <https://www.linguateca.pt/CETEMPublico/informacoes.html>. Acessado em 29 de junho de 2019

(reviews), outros foram aplicados na identificação de emoções em textos curtos, alguns tratavam da extração de nomes de entidades *etc*.

4 MÉTODO PROPOSTO

Neste capítulo é apresentada a metodologia do processo definido na arquitetura proposta nesta pesquisa. Nele, todo o processo está organizado por seção onde se esclarece, de forma metodológica, todo processo de preparação dos algoritmos e dos dados a ser investigados.

4.1 ANÁLISE DO PROBLEMA

A Análise de Sentimentos (AS) se dedica à classificação automática de textos. Essa atividade se baseia nos métodos de Análise Léxica, onde a classificação ocorre basicamente por meio de um dicionário de léxico composto por palavras marcadas com suas respectivas classes gramaticais, e na Análise baseada em Aprendizagem de Máquina, onde se utiliza o córpus anotado com classes que simbolizam categorias. A abordagem baseada em AM é a mais utilizada por pesquisadores por obter melhor desempenho.

Vários trabalhos relatam aplicação da AS baseada em AM sobre córpus em vários idiomas como (WOLFGRUBER, 2015) sobre o idioma alemão, (ALMASHRAEE; DÍAZ; PASCHKE, 2016) sobre o inglês e SOUZA et al. sobre o idioma português. Muitos desses trabalhos aplicam a classificação baseada em classificadores únicos onde a atenção especial recai sobre a fase de pré-processamento dos dados. Outros muitos lançam mão dos artifícios mais arrojados da AM para obter melhor desempenho ao aplicar várias configurações aos classificadores ao utilizar a técnica de *ensemble*, o comitê de classificadores ou sistema de múltiplos classificadores, como fez o BRUM.

No tocante a AS aplicada sobre córpus em português brasileiro, a literatura não é farta. Consideráveis trabalhos foram desenvolvidos para esse idioma, mas ainda há escassez devido ao contexto ao qual se busca aplicar. ZAMPIERI; BECKER criou um dicionário do português histórico onde considerou as duas variantes a européia e a brasileira (cotexto histórico, datação de documentos), TELES; SANTOS; SOUZA avaliou opinião de consumidores com um sistema de classificação binário (contexto varejista). Em face dos resultados sobre classificadores únicos utilizados na literatura baseada sobre o português do Brasil e nos resultados com sistemas de múltiplos classificadores também no nosso idioma, esta pesquisa propõe a construção de um *ensemble* de classificadores tradicionais aplicado a um córpus de textos curtos em português do Brasil sobre o problema de múltiplas classes a fim de obter melhor desempenho em acurácia média em relação aos trabalhos relatados no capítulo 3.

4.2 ARQUITETURA

Neste trabalho a arquitetura proposta é definida em alguns passos que representam as atividades executadas para atingir os resultados desejados. A primeira delas é uma das etapas essenciais para toda e qualquer atividade de Análise de Sentimentos, pois corresponde à coleta dos dados, onde se obteve duas bases de textos curtos em português do Brasil. A segunda etapa corresponde ao pré-processamento do córpus. Nessa fase de pré-processamento as técnicas de remoção de stopwords, tokenização e normalização dos textos foram aplicadas. Em seguida, na terceira etapa, se dedica à configuração dos classificadores clássicos utilizados em AM aplicados à atividade de classificação de textos. Na quarta fase se aplica o treinamento e teste dos classificadores primários por meio da técnica de validação cruzada com k-fold para selecionar o método de extração de características adequado para esta pesquisa. E na quinta, ocorre a seleção do método de Word Embeddings ideal para os experimentos. A seleção do método para extração de características de textos ocorreu por meio da análise das acurácias médias obtidas a partir dos resultados dos testes de três classificadores primários sobre o córpus. A sexta etapa busca treinar todos classificadores utilizados neste trabalho. O treinamento ocorre por meio da técnica de validação cruzada com k-fold, onde k corresponde ao número de partes em que o córpus é dividido como aplicado por TELES; SANTOS; SOUZA. Segundo HAN; PEI; KAMBER, a construção e avaliação de um classificador se baseia na divisão da base de dados em treinamento e testes. Para isso, se pode lançar mão de algumas técnicas de particionamento. Por isso, neste trabalho optou-se por utilizar a técnica de validação cruzada com k-fold. Nesta fase, definiu-se $k = 10$, onde a décima parte é utilizada como conjunto de teste. Dessa forma, se obtém os resultados a partir do conjunto de testes para comparação entre o conjunto de classificadores e o sistema de múltiplos classificadores na sétima etapa. Na sétima e última etapa se define quais os classificadores devem compor as combinações que formarão o *ensemble* de classificadores clássicos.

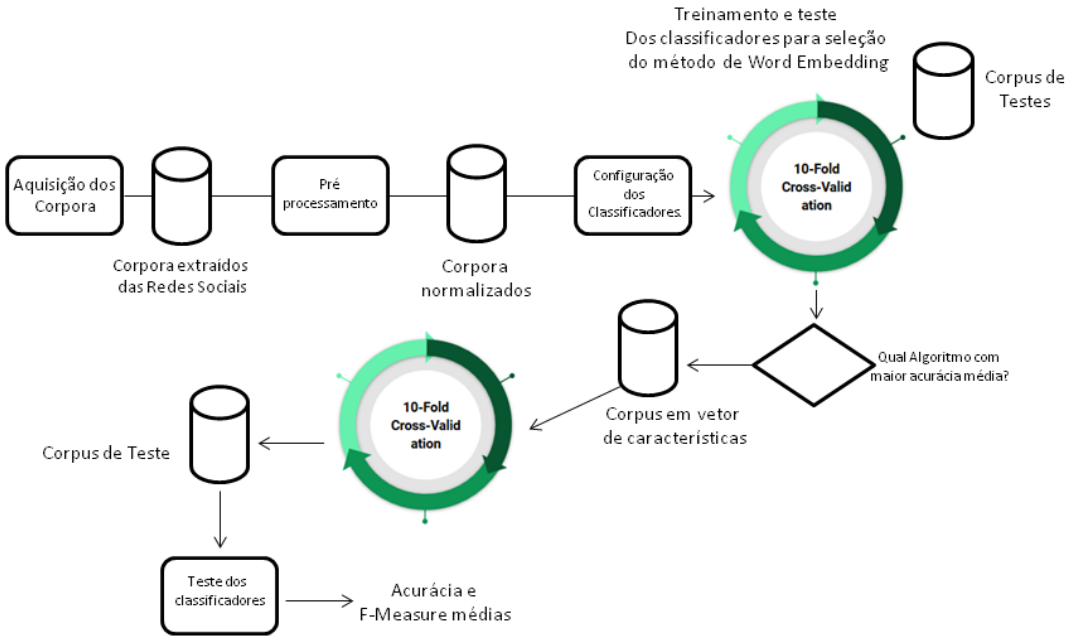


Figura 17 – Arquitetura Proposta

Fonte: Elaborada pelo autor.

4.3 AQUISIÇÃO DOS CORPORA

O estudo das opiniões por meio da Análise de Sentimentos baseada em métodos de Aprendizagem de Máquina (AM) necessita de um conjunto de textos para treinamento. A esse conjunto de textos se dá nome de córpus cujo plural é corpora. Um córpus pode ser um fragmento de texto que representa uma informação completa ou, também, uma unidade documental com várias mensagens. A definição de córpus mais completa foi elaborada por SARDINHA, ao esclarecer que se trata de "uma coleção de dados linguísticos, sejam eles textos ou partes de textos escritos ou a transcrição de fala, de uma determinada língua, escolhidos segundo um determinado critério, representando uma amostra desta língua ou uma variedade linguística".

Neste trabalho foram utilizados os seguintes corpora 2000-tweets-br[1], TweetSentBr[2] disponíveis na literatura.

[1] Disponível em <<http://miningbrgroup.com.br/index.php/resources/>>. Acessado em 02 de julho de 2019

[2] Disponível em <<https://bitbucket.org/HBrum/tweetsentbr/>>. Acessado em 02 de julho de 2019

4.3.1 2000-tweets-br

O 2000-tweets-br é um córpus com 2000 tweets opinativos construído por VITÓRIO et al. para investigar as variações entre o português brasileiro e o europeu. A construção desse conjunto de dados obteve coeficiente Kappa (FLEISS, 1971) de 52,3% de concordância entre os anotadores. Esse coeficiente avalia o grau de alinhamento entre os anotadores quanto à definição da polaridade à uma mensagem.

Classes				**Total**
Positivo	Negativo	Ambos	Neutro	
390	509	61	1040	2000

Tabela 5 – Desbalanceamento das classes no córpus 2000-tweets-br.

Fonte: Elaborada pelo autor.

Na tabela 5 é evidente o alto desbalanceamento das classes, o que significa a diferença de quantidade de exemplos no córpus. Segundo (VITÓRIO et al., 2017), essa divergência reflete o sentimento dos brasileiros no instante da extração dos tweets. Em geral, os métodos de AM utilizados para classificação tendem a favorecer as classes com maior ocorrência no córpus. Esse é um dos problemas que podem ocorrer ao se trabalhar com um conjunto de dados desbalanceado. Os problemas de desbalanceamento podem ser contornados por meio da técnica de aumento de dados, do inglês, Data Augmentation. ZAMPIERI; BECKER aplicou a mesma técnica sobre o desbalanceamento dos documentos históricos para aumento dos exemplos em menor quantidade.

As classes positivo e negativo presentes no córpus dispensam definição, enquanto a classe “ambos” e neutro são dependentes de esclarecimento devido ao contexto do córpus utilizado. A classe “ambos”, segundo VITÓRIO et al., indica que a mensagem contém sentimento ou opinião positiva e negativa. E neutro, representa um texto subjetivo onde não há qualquer tipo de sentimento ou opinião.

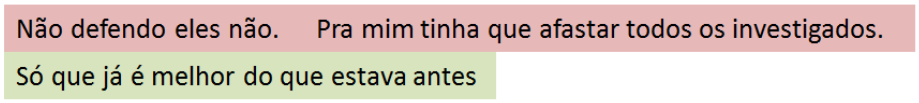


Figura 18 – Exemplo de uma mensagem de classe “ambos” (VITÓRIO et al., 2017)

Fonte: Elaborada pelo autor.

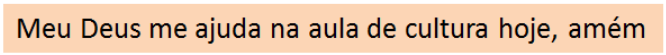


Figura 19 – Exemplo de uma mensagem de classe neutro (VITÓRIO et al., 2017)

Fonte: Elaborada pelo autor.

O tweet presente na Figura 18 é um exemplo de uma mensagem pertencente à classe ambos, que tem a configuração sintática em nível de aspecto. Esse nível ocorre quanto há classificação de partes específicas de uma entidade (BRUM, 2018). As sentenças na Figura 18 foram realçadas de modo a demonstrar com clareza os sentimentos que compõem a informação. As duas mensagens em vermelho demonstram a polaridade negativa, enquanto a sentença realçada em verde demonstra pertencer à classe positiva.

4.3.2 TweetSentBR

O TweetSentBR é um córpus construído a partir de mensagens extraídas da rede social Twitter para avaliar programas de TV devido ao alto nível de interação (BRUM; NUNES, 2018). O córpus contém 15.000 tweets e 17.166 tokens nas polaridades positiva, negativa e neutra com um leve grau de desbalanceamento entre as classes. Compreende-se como *token* as partes de uma mensagem após a sua normalização incluindo pontuação.

Os tweets foram selecionados empiricamente de nove programas de três canais de TV cujo processo de anotação contou com sete anotadores que concordaram entre si em 52,9%, seguindo a escala de (KRIPENDORFF, 2004) onde 100% simboliza concordância plena entre os anotadores.

Classe	Conjunto de Treino	Conjunto de Testes	Total
Positivo	5.745	903	6.648
Negativo	3.840	586	4.426
Neutro	3.414	512	3.926
Total	**12.999**	**2.001**	**15.000**

Tabela 6 – Distribuição das classes do córpus TweetSentBR.

Fonte: (BRUM; NUNES, 2018)

4.4 PRÉ-PROCESSAMENTO DOS TEXTOS

Em Processamento de Linguagem Natural o pré-processamento de textos é útil para normalizar as mensagens a fim de otimizar o custo computacional ao processar grande quantidade de informação. O pré-processamento de um córpus é composto por várias atividades partindo das mais simples, como a substituição ou a remoção de caracteres indesejados, como aplicação de métodos línguísticos de adequação textual, como a aplicação lematização. Neste trabalho foram aplicadas as técnicas simples de pré-processamento sem modificar estruturas tanto das frases como das palavras.

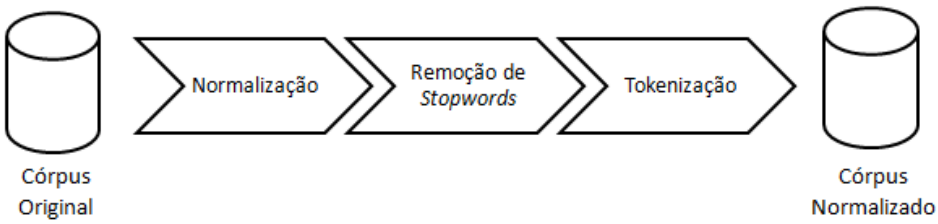


Figura 20 – Pré-processamento dos dados

Fonte: Elaborada pelo autor.

4.4.1 Aplicação da Normalização Textual

- Remoção de nomes de usuários e referências: Todos os nomes de usuários e referências, conhecidos como marcações, foram removidos do texto. As partes em negrito na mensagem abaixo representam o tipo de dado que é excluído no processo de normalização descrito neste item.

  "***inteligemcia***/ *Yoki e Telecine realizam ação inédita de naming rights https://t.co/T7pEWoAtlQ* ***@redetelecine*** *https://t.co/Z5VNPyJlBr*|*NEUTRO*"

- As *hashtags*, representadas pelo símbolo cerquilha "#", foram removidas e as suas expressões mantidas conforme a mensagem:

  "*@Flamengo esse novo uniforme ficou muito bonito, parabéns #SouFlamengoSouMaior #omantoehmeu*"

- Substituição de endereços *web*: Todos os endereços da web bem como e-mails foram substituídos por URL e EMAIL respectivamente (HARTMANN et al., 2017). Conforme exemplo abaixo:

"*EXISTEM MOMENTOS QUE SÃO INESQUECÍVEIS. @LuceroMexico @antoniohogaza* ***https://t.co/xMz5PLMDXf***"

"*EXISTEM MOMENTOS QUE SÃO INESQUECÍVEIS. @LuceroMexico @antoniohogaza URL*"

- Remoção de acentos: Todos os acentos e pontuações foram removidos, e palavras separadas por hífen ou underline foram unidas em uma só palavra.
- Normalização de caracteres unicode: Em Conteúdo Gerado pelo Usuário é normal conter símbolos criados de diferentes formas a partir dos códigos unicode. A princípio os códigos unicode indentificados por expressão regular foram substituídos por um espaçamento. Contudo, na ocorência de outros tipos de combinações de caracteres, aplicou-se o método de normalização para unificar as novas ocorrências dos símbolos conforme a documentação[3] da linguagem python.
- Substituição de números: A fim de evitar ruídos ou mensagens que representem outliers, isto é, um registro com características muito fortes dentro do seu grupo, todos os números presentes nas mensagens foram normalizados para zero (HARTMANN et al., 2017).
- Conversão de texto em letra minúscula: Todas as mensagens foram postas em letras minúsculas exceto as palavras substituídas pelas tags incluídas durante o processo de normalização.

As etapas de normalização foram aplicadas utilizando a linguagem de programação Python[3]. Para essas atividades foi aplicado o conceito de reconhecimento de padrão por meio das expressões regulares.

4.4.2 Stopwords e Tokenização

Visando reduzir a dimensionalidade dos dados para obter melhor desempenho no processo de classificação de textos, várias atividades inerentes ao Processamento de Linguagem Natural são aplicadas nas atividades de AS.

A Remoção de Stopwords é uma prática comum e necessária nos trabalhos de AS, devido às várias ocorrências de palavras bastante comuns no cópus (TELES; SANTOS; SOUZA, 2016). E a Tokenização compõe a parte elementar das mensagens e auxilia agregar mais significado aos textos ALMASHRAEE; DÍAZ; PASCHKE.

Nesta pesquisa, ambas as técnicas foram aplicadas de formas diferentes. A primeira forma se trata da aplicação da tokenização sobre os corpora a cujas características foram extraídas a partir da aplicação dos algoritmos de Word Embeddings CBOW e Skip-gram.

[3] <<https://docs.python.org/2/library/unicodedata.html>>. Acessado em 4 de julho de 2019

A segunda, se trata apenas da remoção de Stopwords sobre os corpora cuja vetorização seria provida pelo TF-IDF.

4.5 EXTRAÇÃO DE CARACTERÍSTICAS DOS TEXTOS

A maioria dos classificadores baseados na abordagem de AM utiliza vetores de características como dado de entrada para processamento. Os textos, embora existam classificadores aptos a processá-los (DUDA; HART; STORK, 2012, p .413), quando utilizados sobre os métodos tradicionais de AM, normalmente são submetidos ao processo de Extração de Características por meio das técnicas de Word Embedding, conforme abordado na seção 2.6. As técnicas mais presentes na literatura são CBOW (MIKOLOV et al., 2013), Skip-gram (MIKOLOV et al., 2013), FastText (BOJANOWSKI et al., 2017), e o Term Frequency-Inverse Document Frequency (TF-IDF) TF-IDF. Neste trabalho os métodos de Word Embeddings CBOW, Skip-gram e TF-IDF foram comparados a fim selecionar qual deles é adequado para atividades de classificação de textos em português do Brasil.

4.6 CONFIGURAÇÃO DOS CLASSIFICADORES

Neste trabalho foram utilizados os classificadores Support Vector Machine, Multilayer Perceptron, Regressão Logística, K-Nearest Neighbor, Random Forest, Gaussian Naïve Bayes, Stochastic Gradient Descent, e o sistema de múltiplos classificadores baseado no voto majoritário. Esses classificadores são bastante recorrentes na literatura e apresentam bons resultados sobre classificação de textos como apresentados em CASTRO et al.. Nesta seção são apresentadas as configurações aplicadas em cada um dos classificadores disponibilizados pelo framework scikit-learn[4].

4.6.1 Support Vector Machine

Para o classificador SVM, se definiu a versão Support Vector Classification juntamente com a função de decisão One-vs-One[5] e kernel do tipo Radial Basis Function por se tratar de atividades de classificação sobre um problema categórico, e o número máximo de mil épocas de treinamento com critério de parada de 1e-3. Em português do Brasil os seguintes trabalhos são referências na utilização do algoritmo SVM (BRUM, 2018) e (TELES; SANTOS; SOUZA, 2016).

4.6.2 Multilayer Perceptron

Nesse classificador poucos parâmetros foram definidos como o solver "lbfgs", mais adequado para bases de dados com poucos registros para treinamento, e cinco camadas

[4] <<https://scikit-learn.org/stable/index.html>>. Acessado em 10 de julho de 2019

[5] <<https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsOneClassifier.html>>. Acessado em 10 de julho de 2019

escondidas e também cinco neurônios. Os Multilayer Perceptron são aplicados em diferentes trabalhos. (MONTEIRO et al., 2018) aplicou o Multilayer Perceptron juntamente com outros classificadores para identificar Fake News.

4.6.3 Gaussian Naïve Bayes

Para esse classificador optou-se por utilizar as configurações padrão do scikit-learn. Pois, neste trabalho se busca a aplicação do algoritmo Naïve Bayes sobre um problema de múltiplas classes, onde o interesse, nesta questão, está na função gaussiana juntamente com o algoritmo probabilístico. Os trabalhos que reportam bons resultados aplicando o Naïve Bayes, normalmente se referem ao problema binário como em (BRUM, 2018).

Neste trabalho, a aplicação deste algoritmo contribui com os seus resultados sobre um problema do tipo categórico, e também na utilização de um algoritmo probabilístico como estimador de um *ensemble*.

4.6.4 K-Nearest Neighbor

Poucos trabalhos relatam a aplicação do KNN para atividades de AS em português do Brasil. O objetivo na aplicação desse algoritmo nesta pesquisa é de utilizar mais de um classificador baseado em espaço vetorial além do SVM.

4.6.5 Regressão Logística

BRUM utilizou o algoritmo de Regressão Logística para classificação de textos em português do Brasil e reportou bons resultados. A motivação de selecionar este algoritmo parte da referência do (BRUM, 2018) onde bons resultados foram apontados. Por isso, decidiu-se analisar o comportamento desse classificador como membro de um comitê de classificadores.

4.6.6 Random Forest

O Random Forest não foi aplicado como classificador base por ser por si só um próprio comitê de classificadores. Neste caso, há um comitê de várias Árvores de Decisão com diferentes configurações. O resultado deste classificador foi comparado com as outras configurações dos comitês de classificadores tradicionais.

4.6.7 Stochastic Gradient Descent

Se trata de uma função de otimização que pode ser utilizada por outros classificadores. A função gradiente é, por exemplo, utilizada no algoritmo GloVe (PENNINGTON; SOCHER; MANNING, 2014). Neste trabalho, essa função recebe parâmetros que influenciam no seu comportamento. Dessa forma ela passa a agir como um classificador. A função gradiente

cuida para que o classificador consiga ter boa performance sobre problemas de múltiplas classes.

5 EXPERIMENTOS E RESULTADOS

Neste capítulo serão apresentados os experimentos feitos a fim de responder às questões levantadas nesta pesquisa. As seções deste capítulo incluem a seleção do método de Word Embeddings para a condução do restante das pesquisa, treinamento e teste dos algoritmos clássicos de AM, e a combinação entre os classificadores. Essa combinação forma um conjunto composto por três estimadores (classificadores clássicos) para configurar o *ensemble* baseado no voto majoritário onde os estimadores devem estar em número ímpar para evitar empate. O conjunto de classificadores unidos para em um mesmo processo caracteriza um sistema de múltiplos classificadores, também chamado de *ensemble* ou comitê. Os algoritmos que compõem os comitês são agrupados entre si sem haver repetições de combinação, e um dos *ensembles* será construído sob a recomendação da métrica Brier Score conforme GUL et al.. A análise de desempenho dos classificadores clássicos e *ensembles* neste trabalho, está baseada nas maiores acurácias e F-Measure médias, juntamente com o menor tempo médio de execução do treinamento. E os dados para as análises são obtidos através dos folds da validação cruzada. A fim de garantir que todos os classificadores foram treinados sobre os mesmos dados, aplicou-se o teste de normalidade de Shapiro Wilk onde se definiu o nível de confiança de 95% ao definir parâmetro $\alpha = 0,05$.

Os testes foram feitos sobre a arquitetura de hardware Intel Core i5 3ª geração (3317U) 1.70 GHz , 8GB de memória RAM.

5.1 CONFIGURAÇÃO DOS EXPERIMENTOS

As atividades de classificação baseadas em técnicas de Aprendizagem de Máquina (AM) normalmente dependem de configuração de parâmetros para adequar os algoritmos ao tipo de problema que está sendo tratado. Nesta seção são descritas as configurações utilizadas para realizar os experimentos.

5.1.1 Support Vector Machine (SVM)

Neste classificador disponível no framework scikit-learn alguns parâmetros foram modificados. O valor do parâmetro "C" referente à penalidade do algoritmo em caso de erros, aplicou-se o valor 1 (um). Quanto maior o valor definido para "C", menor a margem de decisão definida como "b" na Figura 11. **??**p .95]brumexpansao relata que o valor ideal para penalidade é 1 (um) após ter aplicado alguns testes. Esse valor também é o padrão recomendado pelo framework utilizado.

5.1.2 Multilayer Perceptron

Algumas configurações padrão do framework foram modificadas para adequar a rede neural ao problema. A quantidade de 100 camadas escondidas padrão utilizada pelo framework foi mantido e o mesmo número é utilizado para a quantidade de neurônios. A função de ativação ReLU foi mantida na configurção. Entre os solvers disponíveis, selecionou-se o Limited Memory Broyden Fletcher Goldfarb Shanno em vez do padrão adam, ideal para grandes quantidades de registros, e não é utilizada neste trabalho, porque se trata de uma função de otimização baseada no algoritmo SGD, já utilizado neste trabalho com outras configurações. Para a quantidade máxima de iterações decidiu-se manter as 200 iterações propostas pelo framework, mas esse valor pode ser reduzido de acordo com a o parâmtro de tolerância definido como $tol = 1e - 4$. A tolerância indica em qual momento o classificador deve parar o treinamento devido à convergência.

5.1.3 K-Nearest Neighbor (KNN)

Para este classificador os parâmetros são mínimos pois se trata de um algoritmo simples que utiliza a distância euclidiana, definida como $p = 2$, para agrupar as mensagens pertencentes ás mesmas classes. O seu funcionamento depende da definição da quantidade de vizinhos mais próximos em relação á mensagem em análise. Esse parâmetro é definido por $k = 5$.

5.1.4 Regressão Logística

No algoritmo Regressão Logística dois parâmetros foram definidos. O primeiro multiclass como multinomial e o segundo, o *solver*, como LBFGS. Neste algoritmo, o número de iteração definido como padrão é 100, o que se achou razoável para executar os testes.

5.1.5 Random Forest

Para esse classificador, se definiu no primeiro momento o valor 5 para ser o valor máximo de profundidade, e depois esse valor foi desconsiderado. Também foi necessário intervir no número de processadores para executar o processamento onde ficou definido como 4 (quatro). Em seguida, outras cinco configurações diferentes do Random Forest foram testadas sobre o conjunto de testes extraído por meio da técnica de validação cruzada com k-fold variando o número de estimadores de 3, 30, 60, 100 e 200 da mesma forma que foi utilizado por BRUM.

5.1.6 Stochastic Gradient Descent (SGD)

O algoritmo Stochastic Gradient Descent (SGD) é uma função de otimização utilizada por vários algoritmos. Neste trabalho, essa função disponibilizada pelo framework

do scikit-learn permite a modificação de alguns parâmetros para que ela atue como um classificador. Para isso, três parâmetros foram modificados como a função de perda, definida como "*hinge*", que resulta numa SVM linear, a penalidade foi definida como L2 e 5 iterações.

5.2 SELEÇÃO DO MÉTODO DE WORD EMBEDDINGS

Os métodos de Word Embeddings são técnicas eficazes para representação de textos em vetores de características. Inúmeros trabalhos na literatura relatam bons resultados na aplicação desses recursos em atividades de classificação de textos. Neste trabalho os algoritmos Skip-gram, CBOW e TF-IDF foram aplicados em atividades de classificação de tweets em português do Brasil por meio dos classificadores primários SVM, Multilayer Perceptron e SGD sobre a base de dados 2000-tweets-BR.

A base de dados de tweets contém 2000 tweets distribuídos de forma desbalanceada nas quatro classes: positivo, negativo, neutro e ambos (ver capítulo 4 tabela 5). A classe "ambos" foi removida do córpus já que neste trabalho a pesquisa se concentra sobre as classes positiva, negativa e neutra. As mensagens passaram pelo processo de remoção de stopwords, normalização e tokenização conforme definido no capítulo 4. O desbalanceamento desta base de dados não foi tratado para estar de acordo com as condições utilizadas por CASTRO et al..

O objetivo desta etapa é identificar qual dos algoritmos de extração de características obtém desempenho médio em acurácia superior aos seus concorrentes especificados nesta seção. Para isso, aplicou-se o método de validação cruzada com k-fold onde k se refere ao número de iterações. Neste cenário, definiu-se $k = 10$ cujo resultado é um conjunto de acurácias obtidas a cada iteração. Desse conjunto, se obtém a média das acurácias que será utilizada para selecionar o algoritmo de Word Embeddings com o melhor desempenho neste contexto.

5.2.1 Configuração do CBOW e Skip-gram:

Os dois algoritmos utilizaram as mesmas configurações. Para ambos definiu-se o tamanho do vetor $|\vec{a}| = 50$, pois o aumento de $|\vec{a}| = 100$ e $|\vec{a}| = 300$ não mostrou mudanças significativas no desempenho de acurácia. Também se definiu para a quantidade mínima da frequência de uma palavra para ser ignorada, portanto não foi descartada palavra alguma, pois considerou-se que até uma palavra pode ser determinante para diferenciar um padrão. Logo, se definiu $f = 1$ onde f é a frequência de uma palavra. Outro parâmetro importante a ser definido para esses algoritmos é o tamanho da janela w a ser utilizado. O valor de w representa a distância máxima entre a palavra atual e a prevista em uma frase, assim definiu-se $w = 5$ por ser o padrão adotado no framework utilizado. Por fim,

se definiu as épocas de treinamento e como $e = 300$. Esses valores foram definidos levando em consideração o valor padrão do framework.

O CBOW e o Skip-gram são algoritmos desenvolvidos para previsão de palavras a partir de um contexto. Para isso, ambos algoritmos projetam as características dos tokens em vetores de tamanho $|\vec{a}| = s$. Suponhamos um tweet com 100 tokens após o pré-processamento e normalização. Logo, se tem para cada token t um vetor v de tamanho $|\vec{a}| = 50$ para representar um único tweet. Esse é o problema de representação de mensagens utilizando ambos algoritmos. A solução adotada para esse problema foi a aplicação da soma dos vetores representada pela equação 6.1.

$$m_n = \sum_{i=1}^{t_i} \tag{5.1}$$

Onde m_n representa a enésima mensagem resultante do somatório dos vetores de t tokens.

Outra forma seria a média aritmética dos vetores. Isso resultaria em vetores de dimensões variadas e não atingiriam o $|\vec{a}| = 50$. Isso obrigaria o preenchimento das lacunas por zero e tornaria semelhante todas as mensagens de menores dimensões vetoriais devido a quantidade de zeros representando similaridade. A solução aplicada neste procedimento parte do princípio utilizado pelo algoritmo doc2vec desenvolvido por (LE; MIKOLOV, 2014), não adotado nesta pesquisa.

O método que obteve melhor desempenho em acurácia no processo de classificação utilizando os classificadores primários foi selecionado para dar seguimento a pesquisa.

Algoritmo	**SVM**	**Multilayer Perceptron**	**SGD**	**Média**
Skip-gram	38%	41%	40%	39,66%
CBOW	48%	49%	43%	46,66%
TF-IDF	38%	61%	73%	57,33%

Tabela 7 – Desempenho dos algoritmos sobre os diferentes métodos de Word Embeddings

Fonte: Elaborada pelo autor.

Os resultados presentes na tabela 7 demonstram os desempenhos dos classificadores primários sobre os três métodos de Word Embeddings. De acordo com os dados, o TF-IDF obteve melhor desempenho em acurácia média no processo de extração de características e aplicação numa atividade de classificação de tweets em português do Brasil. Com esse resultado, o TF-IDF passou a ser utilizado no restante da pesquisa.

5.3 EXPERIMENTOS COM CLASSIFICADORES CLÁSSICOS

A fim de investigar quais dos algoritmos têm melhor performance em atividades de Análise de Sentimentos (AS), os classificadores mais utilizados nos trabalhos presentes na literatura sobre AS foram utilizados, exceto os algoritmos de aprendizagem profunda, Deep Learning, devido à necessidade de grande volume de dados. Eles foram treinados e

testados sobre os corpora utilizados nesta pesquisa. Os resultados obtidos são acurácia, F-Measure e Brier Score médias resultantes do processo de validação cruzada aplicado neste trabalho. O Brier Score é utilizado nesta fase da pesquisa para complementar análise de definição de quais algoritmos devem compor um *ensemble*.

Classificador	**Acurácia**	**Desvio Padrão**	**F-Measure**	**Brier Score**	**Tempo de Execução em segundos**
Regressão Logísitca	**69%**	0,03	48%	37%	81
Gaussian NB	54%	0,04	46%	57%	18
KNN	52%	0,02	51%	44%	100
Multilayer Perceptron	66%	0,03	57%	57%	74
SGD	**68%**	0,02	57%	57%	0.430
SVM	28%	0,05	23%	41%	89

Tabela 8 – Desempenho dos classificadores sobre o Córpus 2000-Tweets-br do MiningBR Research Group.

Fonte: Elaborada pelo autor.

Na tabela 8 os classificadores Regressão Logística e Stochastic Gradient Descent obtiveram a melhor performance entre todos os classificadores em acurácia média sobre o córpus 2000-tweets-BR do MiningBR Research Group. Os tempos de execução também apontam baixo custo computacional. No gráfico da Figura 21 é possível verificar o quanto os classificadores estão próximos em desempenho.

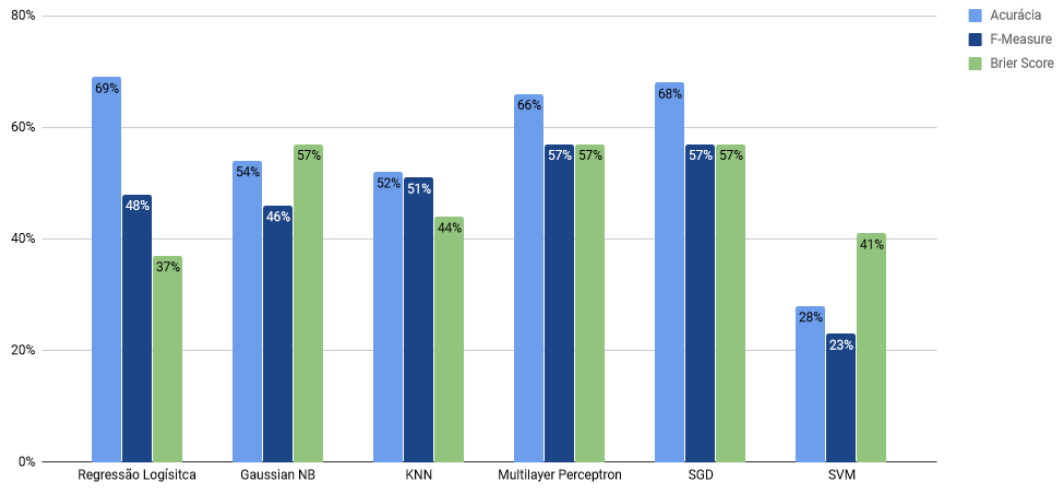


Figura 21 – Desempenho dos classificadores sobre o Córpus 2000-Tweets-br do MiningBR Research Group

Fonte: Elaborada pelo autor.

No córpus TweetSentBr, os resultados dos testes dos mesmos classificadores, represen-

tados na tabela 9, os classificadores Regressão Logística e Stochastic Gradient Descent demonstraram aumento de 2% de acurácia média em relação aos dados apresentados na tabela 8, mas obtiveram resultados diferentes na métrica F-Measure. Nessa métrica o SGD obteve o melhor resultado no córpus 2000-tweets-BR.

Classificador	**Acurácia**	**Desvio Padrão**	**F-Measure**	**Brier Score**	**Tempo em Segundos**
Regrassão Logísitca	**71%**	0,15	46%	29%	93
Gaussian NB	56%	0,16	36%	39%	35
KNN	51%	0,25	40%	31%	43
Multilayer Perceptron	68%	0,16	40%	40%	191
SGD	**70%**	0,14	40%	33%	0,889
SVM	20%	0,07	0.09%	29%	440

Tabela 9 – Desempenho dos classificadores sobre o Córpus TweetSentBR.

Fonte: Elaborada pelo autor.

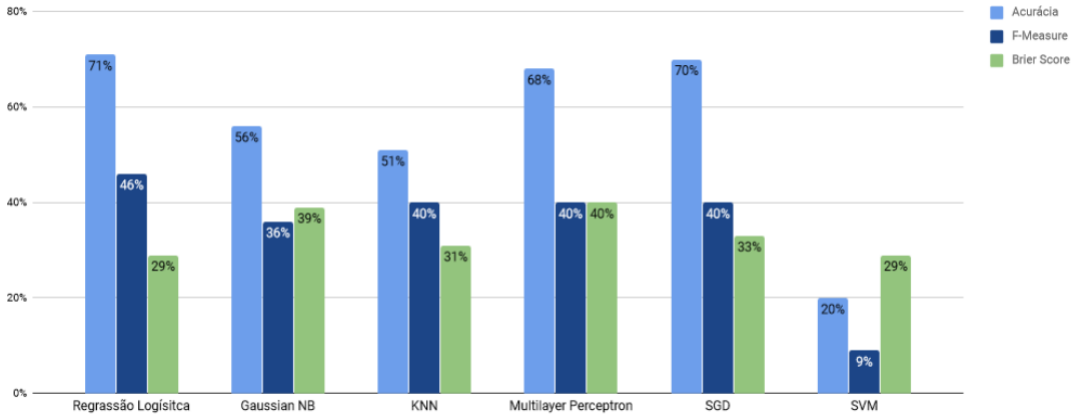


Figura 22 – Desempenho dos classificadores sobre o Córpus TweetSentBR

Fonte: Elaborada pelo autor.

Todos os resultados em acurácia média foram submetidos ao teste de normalidade de Shapiro-Wilk a fim de ter conhecimento da distribuição dos dados, que neste trabalho se trata das características extraídas das mensagens. Para esse teste, se define a hipótese nula, ou da existência, H_0 partindo da premissa de que os dados provém de uma distribuição normal. O teste baseado sobre o α indicará a rejeição ou a aceitabilidade de H_0, quando $\alpha = 0.05$. A partir do resultado obtido é possível indicar o teste de hipótese adequado.

Na distribuição dos dados das acurácias médias sobre o córpus 2000-tweets-br foi possível identificar que apenas os dados do algoritmo KNN não provém de uma distribuição normal, rejeitando portanto a hipótese nula H_0.

Os resultados obtidos sobre o córpus TweetSentBR apresentam alto desvio padrão, o que representa dispersão dos dados. Os algoritmos cujos resultados não descartam a

Classificador	**p-value**	**Desvio Padrão**	**p-value $<\alpha$**	**Resultado**
Regressão Logística	0,474	0,03	Não	Não é possível rejeitar H_0
Gaussian Naïve Bayes (NB)	0,623	0,04	Não	Não é possível rejeitar H_0
KNN	0,047	0,02	Sim	Rejeita-se H_0
Multilayer Perceptron	0,331	0,03	Não	Não é possível rejeitar H_0
SGD	0,623	0,02	Não	Não é possível rejeitar H_0
SVM	0,802	0,05	Não	Não é possível rejeitar H_0

Tabela 10 – Teste de Normalidade dos dados sobre o córpus 2000-tweets-BR.

Fonte: Elaborada pelo autor.

Classificador	**p-value**	**Desvio Padrão**	**p-value $<\alpha$**	**Resultado**
Regressão Logística	0,305	0,15	Não	Não é possível rejeitar H_0
Gaussian Naïve Bayes (NB)	0,026	0,16	Sim	Rejeita-se H_0
KNN	0,156	0,25	Não	Não é possível rejeitar H_0
Multilayer Perceptron	0,069	0,16	Não	Não é possível rejeitar H_0
SGD	0,026	0,14	Sim	Rejeita-se H_0
SVM	0,0	0,07	Sim	Rejeita-se H_0

Tabela 11 – Teste de Normalidade dos dados sobre o córpus TweetSentBR.

Fonte: Elaborada pelo autor.

hipótese nula, indicando que seus dados provém de uma distribuição normal, apesar da irregular distribuição dos dados, foram Regressão Logística, KNN e o Multilayer Perceptron.

5.3.1 Teste de Hipótese sobre os classificadores clássicos

Neste passo aplica-se os testes de hipótese t-Student e Wilcoxon sobre os resultados em acurácia obtidos do conjunto de testes executados durante a aplicação da validação cruzada. O teste de t-Student é aplicado sobre os dados que atendem à distribuição normal, não rejeitando a hipótese nula H_0 de acorco com as tabelas 10 e 11, onde se avalia se o

Classificador		**Tipo de Teste**	**p-value**	**p-value $<\alpha$**	**Resultado**
Regressão Logística	Gaussian NB	t-Student	7.3366e-07	Sim	Não é possível rejeitar H_0
	Multilayer Perceptron	t-Student	0.1057	Não	Rejeita-se H_0
	SGD	t-Student	0.1452	Não	Rejeita-se H_0
	SVM	t-Student	0.0015	Sim	Não é possível rejeitar H_0
	KNN	Wilcoxon	0.0092	Sim	Não é possível rejeitar H_0

Tabela 12 – Comparação de desempenho do classificador Regressão Logística em relação aos demais sobre o córpus 2000-tweets-BR.

Fonte: Elaborada pelo autor.

valor do p-value obtido está dentro da região de aceitação, isto é, menor que α; e o teste de Wilcoxon para os dados que não estão distribuídos na forma normal, acatando então, a hipótese alternativa H_1. Acata-se H_0 quando o valor obtido for menor que $\alpha = 0.05$.

De acordo com as tabelas 8 e 9 os classificadores Regressão Logística e Stochastic Gradient Descent obtiveram melhor desempenho em acurácia média sobre os demais. Assim, o teste t-Student foi aplicado comparando se há divergências entre eles e os demais classificadores. A hipótese nula H_0 procura saber se a há diferença significativa entre os classificadores com melhor desempenho em relação aos outros modelos.

Como se pode confirmar na tabela 12 não há grande divergência estatística entre o desempenho do Regressão Logística, o Multiplayer Perceptron e o SGD. Já na comparação com o KNN e o Gaussian Naïve Bayes (NB), a superioridade do Regressão Logística é maior, embora em termos de acurácia média eles não estejam distantes.

Na tabela 13 se pode ver que apenas a comparação que não apresentou grande divergência estatística entre as médias foi entre SGD e Regressão Logística. Um resultado já apontado na tabela 12. Já em relação aos demais classificadores, o SGD mostrou ser superior estatisticamente.

Os testes de hipóteses indicam que os algoritmos Regressão Logística e Stochastic Gradient Descent são estatisticamente superiores. Para isso, considerou-se a hipótese nula cuja premissa é de que existe diferença significativa entre as médias dos classificadores com melhor desempenho de acurácia. Os testes com os classificadores clássicos em execução isolada foram feitos sobre o córpus 2000-tweets-BR.

O próximo teste de hipótese a ser aplicado é sobre o córpus TweetSentBR. O teste de normalidade de Shapiro-Wilk sobre esse córpus indicou alta disperção dos dados indicando que eles não provém de uma distribuição normal. A comparação dos desempenhos dos modelos foi feita conforme na comparação anterior onde se considera os classificadores

Classificador		**Tipo de Teste**	**p-value**	**p-value $<\alpha$**	**Resultado**
SGD	Gaussian NB	t-Student	1.8556e-08	Sim	Não é possível rejeitar H_0
	Multilayer Perceptron	t-Student	0.0019	Sim	Não é possível rejeitar H_0
	Regressão Logistica	t-Student	0.1452	Não	Rejeita-se H_0
	SVM	t-Student	6.4974e-05	Sim	Não é possível rejeitar H_0
	KNN	Wilcoxon	0.0050	Sim	Não é possível rejeitar H_0

Tabela 13 – Comparação de desempenho do classificador SGD em relação aos demais sobre o córpus 2000-tweets-BR.

Fonte: Elaborada pelo autor.

com melhor desempenho em acurácia sobre o córpus. Atendem a esse critério o Regressão Logística e o SGD.

Classificador		**Tipo de Teste**	**p-value**	**p-value $<\alpha$**	**Resultado**
Regressão Logística	Gaussian NB	Wilcoxon	0.0050	Sim	Não é possível rejeitar H_0
	Multilayer Perceptron	t-Student	0.7139	Não	Rejeita-se H_0
	SGD	Wilcoxon	0.2411	Não	Rejeita-se H_0
	SVM	Wilcoxon	0.0050	Sim	Não é possível rejeitar H_0
	KNN	t-Student	0.0218	Sim	Não é possível rejeitar H_0

Tabela 14 – Comparação de desempenho do classificador Regressão Logística em relação aos demais sobre o córpus TweetSentBR.

Fonte: Elaborada pelo autor.

Os testes de hipóteses sobre os resultados dos algoritmos sobre o córpus TweetSentBR apontam que há diferença estatística entre as médias do algoritmo Regressão Logística e os outros classificadores, exceto em relação ao Multilayer Perceptron e o Stochastic Gradient Descent cujos resultados rejeitam H_0.

Por último, a comparação do SGD com os demais classificadores sobre o córpus TweetSentBR por meio dos testes de hipóteses de Wilcoxon e t-Student indicam que a hipótese nula H_0 foi rejeitada nos resultados de comparação com o Multilayer Perceptron e o Regressão Logística. Nos demais resultados H_0 não foi rejeitada e demonstrou que existe diferença significativa entre as médias.

Classificador		Tipo de Teste	p-value	p-value $<\alpha$	Resultado
SGD	Gaussian NB	Wilcoxon	0.0050	Sim	Não é possível rejeitar H_0
	Multilayer Perceptron	t-Student	0.8805	Não	Rejeita-se H_0
	Regressão Logística	Wilcoxon	0.2411	Não	Rejeita-se H_0
	SVM	Wilcoxon	0.0050	Sim	Não é possível rejeitar H_0
	KNN	t-Student	0.0189	Sim	Não é possível rejeitar H_0

Tabela 15 – Comparação de desempenho do classificador SGD em relação aos demais sobre o córpus TweetSentBR.

Fonte: Elaborada pelo autor.

Os experimentos dos classificadores clássicos foram feitos sobre os dois corpora utilizados nesta pesquisa. O córpus 2000-tweets-BR contendo alto grau de desbalanceamento entre as classes positivo, negativo, neutro e "ambos". Todas as 61 mensagens da classe "ambos" foram removidas do experimento. E o córpus TweetSentBR contendo 15 mil tweets e baixo grau de desbalanceamento dividido nas classes positivo, negativo e neutro. Em seguida, os classificadores foram treinados por meio da técnica de validação cruzada onde $k = 10$ que resultou nas acurácias médias para comparação de desempenho entre os classificadores. Os resultados dos testes apontam para boa performance dos algoritmos Regressão Logística e Stochastic Gradient Descent sobre corpora em português do Brasil. Embora os resultados tenham chegado a 71% de acurácia média, ainda está abaixo dos resultados do estado arte sobre textos em português.

A próxima seção se dedica aos experimentos com a técnica *ensemble* composto pelo classificadores clássicos sobre os mesmos corpora a fim de analisar o desempenho.

5.4 EXPERIMENTOS COM *ENSEMBLE*

Nesta fase são apresentados os experimentos com os classificadores compondo a técnica de *Ensemble*. Aqui os classificadores clássicos de Aprendizagem de Máquina são combinados e comparados. Nesta seção o classificador Random Forest também será abordado por se tratar de um *Ensemble* de vários classificadores do tipo Decision Tree, Árvore de Decisão, com diferentes configurações.

5.4.1 Random Forest

Nesta subseção se analisa o algoritmo Random Forest cuja aplicação em atividades de AS não é constante. Esse algoritmo é um *ensemble* de Decision Trees com diferentes configurações conforme descrito na seção 4 sobre os dois corpora utilizados nesta pesquisa.

Os primeiros resultados foram obtidos através do treinamento e teste desse classificador sobre o córpus 2000-tweets-BR (TELES; SANTOS; SOUZA, 2016), e em seguida os mesmos testes foram aplicados sobre o córpus TweetSentBR (BRUM, 2018).

Nº de Estimadores	**Acurácia**	**Desvio Padrão**	**F-Measure**	**Tempo de Execução em segundos**
3	**49%**	0,05	**26%**	55
30	47%	0,04	25%	17
60	45%	0,05	24%	24
100	46%	0,05	25%	34
200	47%	0,05	25%	82

Tabela 16 – Desempenho do classificador Random Forest com tamanho máximo definido sobre o corpus 2000-tweets-BR.

Fonte: Elaborada pelo autor.

O Random Forest não demonstrou bom desempenho em nenhuma das configurações aplicadas sobre o córpus 2000-tweets-BR conforme se pode verificar na tabela 16. Já na tabela 17 é possível perceber a evolução em acurácia média bem como em F-Measure em relação às configurações apresentadas na tabela 16.

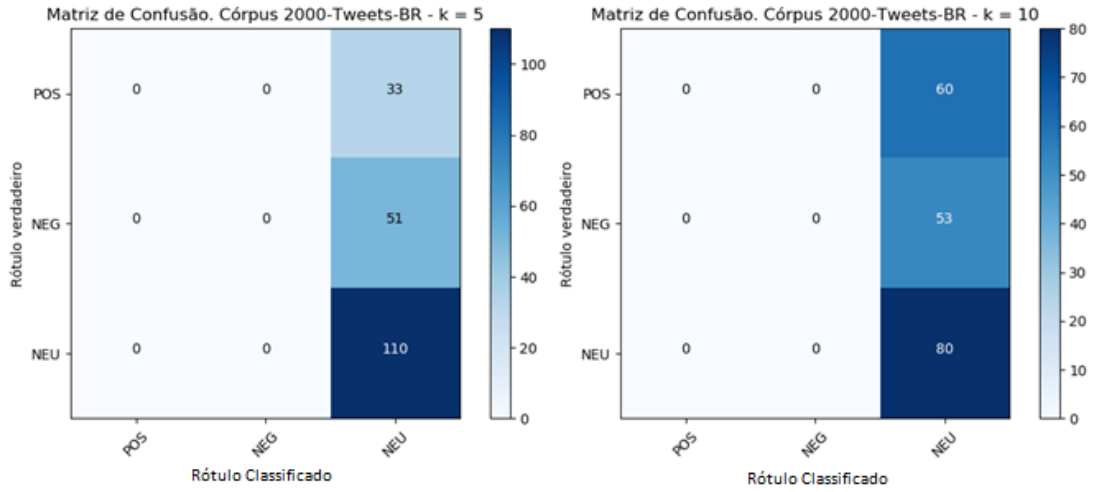


Figura 23 – Matriz de confusão do Random Forest sobre o córpus 2000-tweets-BR com tamanho limitado para árvores.

Fonte: Elaborada pelo autor.

O resultado apresentado na tabela 17 apresenta semelhança entre os classificadores de 30 a 200 estimadores em acurácia e F-Measures. O número de estimadores representa a quantidade de Árvores de Decisão utilizada pelo algoritmo. Embora os resultados sejam semelhantes, o tempo de execução pode ser utilizado para inferir que a configuração com 30 estimadores e sem limite de tamanho para a árvore é uma configuração interessante.

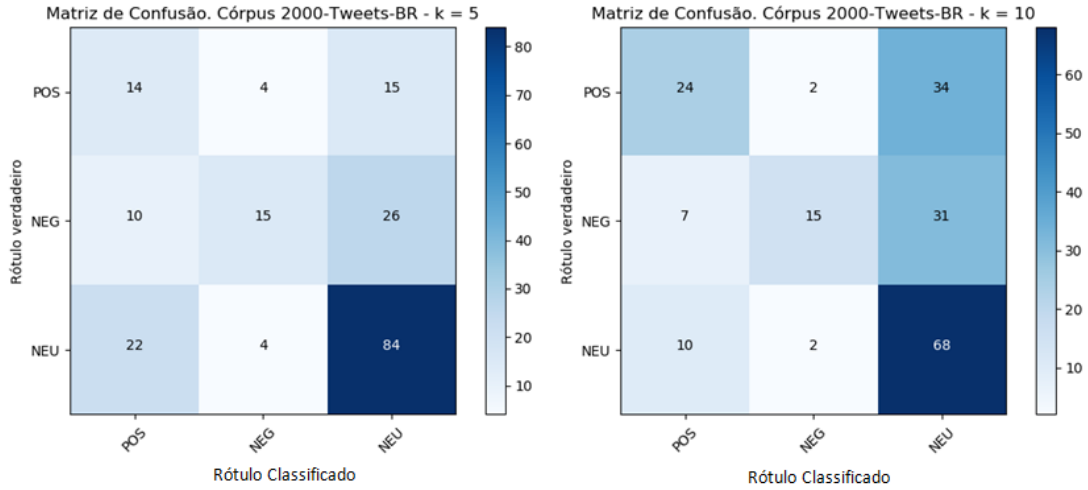


Figura 24 – Matriz de confusão do Random Forest sobre o córpus 2000-tweets-BR sem tamanho limitado para árvores

Fonte: Elaborada pelo autor.

Nº de Estimadores	**Acurácia**	**Desvio Padrão**	**F-Measure**	**Tempo de Execução em Segundos**
3	62%	0,04	41%	20
30	**66**%	0,05	**47**%	**70**
60	66%	0,03	47%	168
100	66%	0,03	47%	228
200	66%	0,02	47%	488

Tabela 17 – Desempenho do classificador Random Forest sem tamanho máximo definido sobre o corpus 2000-tweets-BR.

Fonte: Elaborada pelo autor.

A matriz de confusão na Figura 23 mostra o desempenho do Random Forest com o tamanho máximo de expansão das Árvores de Decisão limitado a cinco. Essa configuração não obteve bons resultados, e isso se torna mais explícito ao observar a figura. Nota-se total descalibramento do algoritmo prevendo todas as classes como neutras.

Na Figura 24 constam os resultados do desempenho do Random Forest, desta vez sem limite de expansão para as Árvores de Decisão. Nessa imagem é possível verificar a evolução do algoritmo na atividade de classificação sobre o mesmo córpus.

Ambos resultados exibem o estado do algoritmo durante o processo de treinamento e teste sob o processo de treinamento utilizando a Validação Cruzada, onde $k = 10$. No exemplo, as matrizes foram criadas quando $k = 5$ e $k = 10$.

A seguir as próximas tabelas mostram os resultados do Random Forest sobre o córpus TweetSentBR. As configurações aplicadas seguem o mesmo formato de variar os estimado-

res, treinamento e teste com e sem tamanho máximo de expansão das Árvores de Decisão.

Nº de Estimadores	**Acurácia**	**Desvio Padrão**	**F-Measure**	**Tempo de Execução em Segundos**
3	42%	0,30	25%	0.268
30	**45%**	0,27	**26%**	0.818
60	44%	0,31	25%	20
100	45%	0,32	26%	28
200	45%	0,32	26%	55

Tabela 18 – Desempenho do classificador Random Forest com limite definido para tamanho das árvores sobre o córpus TweetSentBR

Fonte: Elaborada pelo autor.

Os testes feitos sobre o córpus TweetSentBR não obtiveram bons resultados na configuração com tamanho máximo definido para as Árvores de Decisão. Contudo, verificou-se que a configuração com 30 estimadores é mais performática do que qualquer outra configuração aplicada neste algoritmo. Esse resultado se repete nas tabelas 18 e 19 onde a medida F-Measure aponta boa precisão e o tempo de excução indica a viabilidade computacional.

No detalhe, fica claro que o Random Forest é menos performático ao se definir uma baixa margem para criação das Árvores de Decisão mesmo com córpus com poucos registros como o 2000-tweets-BR, que nesta pesquisa 1939 das 2000 mensagens são utilizadas (ver capítulo 4 e seção 4.3.1).

Nº de Estimadores	**Acurácia**	**Desvio Padrão**	**F-Measure**	**Tempo de Execução em Segundos**
3	64%	0,19	42%	21
30	**69%**	0,30	**46%**	77
60	68%	0,17	46%	174
100	69%	0,15	46%	236
200	69%	0,16	46%	496

Tabela 19 – Desempenho do classificador Random Forest sem limite definido para tamanho das árvores sobre córpus TweetSentBR.

Fonte: Elaborada pelo autor.

Os resultados deste passo de treinamento e teste do Random Forest, sem limite de expansão das Árvores de Decisão, sobre o córpus TweetSentBR apresentados nas tabelas 18 e 19 podem ser melhor visualizados nas matrizes de confusão nas figuras 25 e 26.

Os testes indicam que entre as configurações desse classificador, aquela que obteve acurácia e F-Measure médias altas e menor tempo de execução em segundos conta com

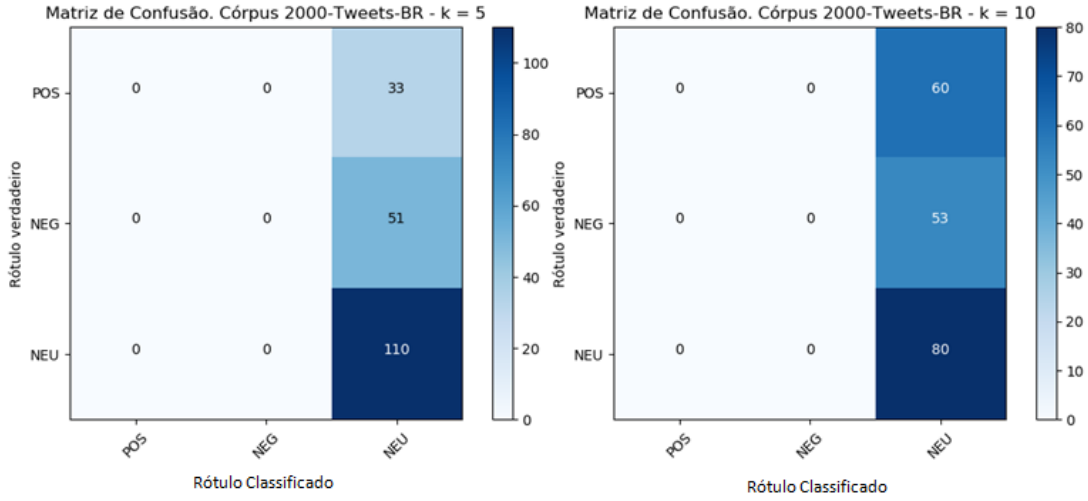


Figura 25 – Matriz de confusão do Random Forest sobre o córpus TweetSentBR com tamanho limitado para árvores.

Fonte: Elaborada pelo autor.

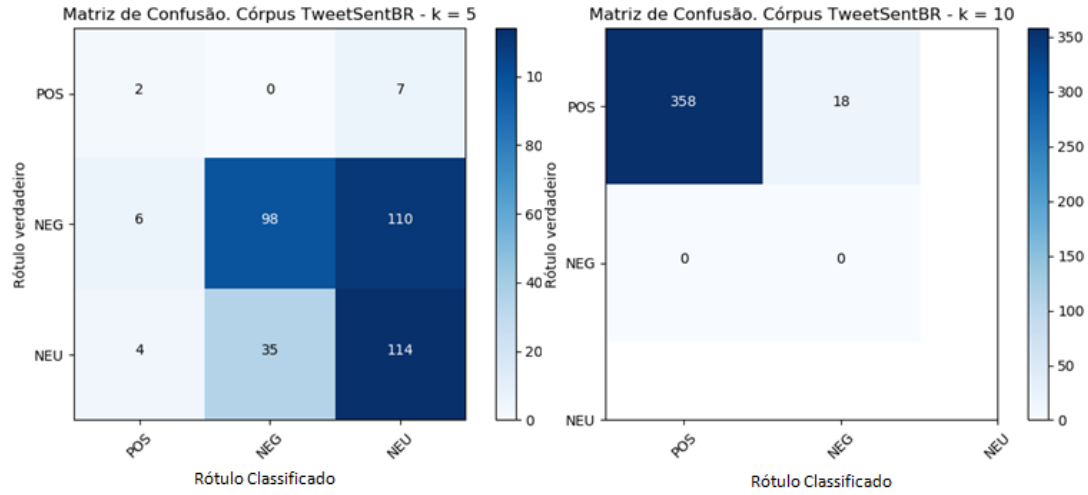


Figura 26 – Matriz de confusão do Random Forest sobre o córpus TweetSentBR sem tamanho limitado para árvores.

Fonte: Elaborada pelo autor.

trinta estimadores. Todos os testes foram aplicados em dois corpora com três classes onde normalmente o desempenho dos classificadores é menor.

5.4.2 Experimentos com *ensemble* de classificadores clássicos

O objetivo de combinar classificadores com diferentes configurações para atividades de Análise de Sentimentos é obter bons resultados na classificação de textos curtos. As combinações que formam um sistema de múltiplos classificadores foram feitas de duas formas: a primeira, a partir da seleção dos menores Brier Score médios. Segundo GUL et al., quanto menor o resultado em Brier Score obtido por um classificador, mais recomendado é aquele algoritmo para compor um *ensemble*. Por isso, nesta etapa, seleciona-se aqueles algoritmos que obtém os menores Brier Scores para formação de um dos *ensembles*. A segunda forma de combinação foi feita por meio da união de todos os classificadores combinando-os entre si. Nesta fase não se considerou o algoritmo Random Forest por ele ser um *ensemble* composto por algoritmos de Árvores de Decisão, Decision Trees, com diferentes configurações. Os comitês construídos para esta seção são baseados nos testes dos classificadores executados de forma isolada a fim de obter parâmetros para testes de comparação dos seus desempenhos. A heterogeneidade e a diversidade esperadas para sistemas de múltiplos classificadores, neste trabalho, são obtidas por meio dos diferentes algoritmos implementados pelos classificadores para realização das atividades de classificação, e também pelas suas diferentes configurações de parâmetros, e por meio da técnica de validação cruzada onde se separa os conjuntos de treinamento e testes conforme recomendado por POLIKAR.

Os resultados foram submetidos aos testes de normalidade e de hipótese onde foram aplicados o teste de Shapiro-Wilk para verificar a normalidade da distribuição, o teste de t-Student para testes de hipóteses sobre distribuições normais, e o teste de Wilcoxon para testes de hipóteses para distribuições que não pertencem a uma distribuição normal. Ambos os testes foram aplicados sobre o conjunto de testes onde os resultados são baseados no valor de significância $\alpha = 0.05$ que obedece à regra de 95% de confiança para o resultado obtido.

5.4.2.1 Experimentos sobre o Córpus 2000-tweets-BR

Nesta etapa todos os classificadores foram combinados entre si e testados sobre o córpus 2000-tweets-BR. O *ensemble* composto por K-Nearest Neighbor, Regressão Logística e Support Vector Machine está baseado na métrica Brier Score onde sugere a combinação dos classificadores com o menor resultado dessa métrica.

Na tabela 20 os resultados mostram que todos os *ensemble* tiveram desempenho baixo em todas as médias. Ressalta-se também o baixo desempenho do *ensemble* baseado no Brier Score cujo resultado em acurácia e F-Measure médias é baixo. Para este caso, observa-se um F-Measure distante da acurácia, o que indica baixa precisão. O próximo passo é a análise estatística desses resultados a iniciar pelo teste de normalidade de Shapiro-Wilk para definir o tipo de teste de hipótese que será aplicado. O teste de Shapiro-Wilk é um teste que verifica se as amostras provém de uma distribuição normal.

Ensemble	**Acurácia**	**F-Measure**	**Brier Score**	**Tempo de Execução em segundos**
KNN + Multilayer Perceptron + SVM	42%	29%	Não	1.250
Regressão Logística + Multilayer Perceptron + SVM	51%	33%	Não	1.360
Gaussian NB + Multilayer Perceptron + SVM	64%	49%	Não	144
Multilayer Perceptron + Regressão Logística + Gaussian NB	69%	52%	Não	163
SGD+ KNN + Gaussian NB	64%	49%	Não	164
Regressão Logística+ SGD+Gaussian NB	53%	43%	Não	861
Regressão Logística +Gaussian NB+KNN	64%	49%	Não	233
KNN + Regressão Logística + SVM	54%	39%	Sim	1.224

Tabela 20 – Desempenho dos *ensembles* sobre o córpus 2000-tweets-BR.

Fonte: Elaborada pelo autor.

Em caso, dessa hipótese ser positiva, aplica-se, então, o teste t-Student; caso contrário será aplicado o teste de Wilcoxon. Dessa forma, será possível afirmar se há diferença significativa entre os resultados.

O teste de Shapiro-Wilk foi aplicado sobre os resultados provenientes da tabela 20 a fim de identificar o tipo de teste a ser aplicado. Para isso, as hipótese nula H_0 foi definida sob a premissa de que os dados provém de uma distribuição normal. Os resultados presentes na tabela 21 demonstram que as acurácias médias formam uma distribuição normal, isto é, não rejeitam a hipótese nula. Assim, para estes resultados, é possível aplicar o teste t-Student. Para comparar os resultados selecionou-se o *ensemble* com o melhor desempenho considerando a acurácia média, F-Measure e, por fim, o tempo médio de execução em segundos. De acordo com esses parâmetros, o *ensemble* composto por Multilayer Perceptron, Regressão Logística e GaussianNB é a combinação que será comparada às demais sobre os mesmos dados.

Os testes de hipóteses presentes na tabela 22 foram aplicados sobre as acurácias médias de todos os classificadores indicam que há diferença significativa entre o *ensemble* composto por Multilayer Perceptron, Regressão Logística e GaussianNB em relação à maioria dos *ensembles* listados na tabela. A combinação composta por KNN, Regressão Logística e SVM cuja métrica base é o Brier Score, é a exceção. Esse comitê rejeitou a hipótese nula, indicando assim, que não há diferença significativa entre este *ensemble* e o *ensemble* de maior desempenho em acurácia e F-Measure médias.

Ensemble	**p-value**	**Desvio Padrão**	**p-value $<\alpha$**	**Resultado**
KNN + Multilayer Perceptron + SVM	0,881	0,04	Não	Não é possível rejeitar H_0
Regressão Logística + Multilayer Perceptron + SVM	0,256	0,05	Não	Não é possível rejeitar H_0
Gaussian NB + Multilayer Perceptron + SVM	0,471	0,02	Não	Não é possível rejeitar H_0
Multilayer Perceptron + Regressão Logística + Gaussian NB	0,796	0,04	Não	Não é possível rejeitar H_0
SGD+ KNN + Gaussian NB	0,670	0,02	Não	Não é possível rejeitar H_0
Regressão Logística+ SGD+Gaussian NB	0,246	0,03	Não	Não é possível rejeitar H_0
Regressão Logística +Gaussian NB+KNN	0.670	0,02	Não	Não é possível rejeitar H_0
KNN + Regressão Logística + SVM	0.647	0,04	Não	Não é possível rejeitar H_0

Tabela 21 – Teste de normalidade de Shapiro-Wilk sobre os *ensembles* treinados sobre o córpus 2000-tweets-BR

Fonte: Elaborada pelo autor.

5.4.2.2 Experimentos sobre o Córpus TweetSentBR

Embora os experimentos sobre o córpus TweetSentBR tenham sido feitos sob as mesmas condições definidas no início desta seção, existe a exceção sobre o *ensemble* composto por Multilayer Perceptron, Stochastic Gradient Descent e Regressão Logística. Este comitê não pôde ser composto pelos mesmos algoritmos que compuseram o sétimo comitê baseado sobre KNN, Regressão Logística e SVM nos testes presentes na tabela 20. A métrica utilizada para selecionar os classificadores para compor este *ensemble* indicou a combinação Multilayer Perceptron, SGD e Regressão Logística por obterem menores resultados de Brier Score alinhados com boas acurácias e F-Measure médias.

Analisando a tabela 9 pode-se levantar a questão do porquê não selecionar o algoritmo SVM para compor o *ensemble* baseado no Brier Score. Embora o SVM tenha obtido um Brier Score muito baixo, o que de fato seria bom, os seus demais resultados no mesmo teste foram baixos. Então, a seleção do Multilayer Perceptron foi decidida por meio da análise de todos os desempenhos o que possibilitou o descarte do SVM para a seleção

Ensembles		**Tipo de Teste**	**p-value**	**p-value $<\alpha$**	**Resultado**
Multilayer Perceptron+ Regressão Logística e GaussianNB	KNN + Multilayer Perceptron + SVM	t-Student	0.001	Sim	Não é possível rejeitar H_0
	Regressão Logística + Multilayer Perceptron + SVM	t-Student	0.014	Sim	Não é possível rejeitar H_0
	Gaussian NB + Multilayer Perceptron + SVM	t-Student	0.002	Sim	Não é possível rejeitar H_0
	SGD+ KNN + Gaussian NB	t-Student	0.0004	Sim	Não é possível rejeitar H_0
	Regressão Logística+ SGD + Gaussian NB	t-Student	1.082017e-05	Sim	Não é possível rejeitar H_0
	Regressão Logística + Gaussian NB+ KNN	t-Student	0.0004	Sim	Não é possível rejeitar H_0
	KNN + Regressão Logística + SVM	t-Student	0.07	Não	Rejeita-se H_0

Tabela 22 – Comparação entre os *ensembles* sobre o córpus 2000-tweets-BR

Fonte: Elaborada pelo autor.

baseada no Brier Score.

Na tabela 23 é possível verificar que os desempenhos foram melhores sobre o córpus TweetSentBR devido ao baixo grau de desbalanceamento do córpus. Os melhores resultados foram os *ensembles* compostos pelos classificadores Multilayer Perceptron, Regressão Logística e GaussianNB com o melhor resultado, e o proposto pela métrica Brier Score contendo os classificadores Multilayer Perceptron, SGD e Regressão Logística. O próximo passo será análise estatística a fim de verificar se há diferença significativa entre os classificadores em relação ao *ensemble* com melhor resultado.

Ensemble	**Acurácia**	**F-Measure**	**Brier Score**	**Tempo de Execução em segundos**
KNN + Multilayer Perceptron + SVM	39%	27%	Não	1.474
Regressão Logística + Multilayer Perceptron + SVM	49%	31%	Não	1.583
Gaussian NB + Multilayer Perceptron + SVM	66%	48%	Não	163
Multilayer Perceptron + Regressão Logística + Gaussian NB	71%	50%	Não	184
SGD+ KNN + Gaussian NB	67%	48%	Não	182
Regressão Logística+ SGD+Gaussian NB	55%	42%	Não	928
Regressão Logística +Gaussian NB+KNN	67%	48%	Não	260
Multilayer Perceptron + SGD + Regressão Logística	71%	48%	Sim	1.257

Tabela 23 – Desempenho do Ensemble sobre o córpus TweetSentBR.

Fonte: Elaborada pelo autor.

Na tabela 24 é possível verificar os testes de normalidade aplicados sobre os resultados obtidos na tabela 23, onde os resultados de três *ensembles* não são provenientes de uma distribuição normal e rejeitam a hipótese nula, da existência da normalidade dos dados. Dessa forma, esses comitês serão submetidos ao teste de Wilcoxon enquanto os demais serão submetidos ao teste t-Student. O desvio padrão, quando alto, indica que os dados estão mais afastados da média e espalhados em diversos valores. O *ensemble* com os algoritmos Regressão Logística, Multilayer Perceptron e SVM obteve o maior desvio padrão entre todos os demais comitês. Todos os resultados serão comparados com o *ensemble* que obteve maior resultado considerando acurácia e F-Measure médias, e o tempo médio de execução. Assim, o comitê composto por Multilayer Perceptron, Regressão Logística e GaussianNB será comparado com demais.

Ensemble	**p-value**	**Desvio Padrão**	**p-value $<\alpha$**	**Resultado**
KNN + Multilayer Perceptron + SVM	0,048	0,24	Sim	Rejeita-se H_0
Regressão Logística + Multilayer Perceptron + SVM	0,199	0,33	Não	Não é possível rejeitar H_0
Gaussian NB + Multilayer Perceptron + SVM	0,093	0,15	Não	Não é possível rejeitar H_0
Multilayer Perceptron + Regressão Logística + Gaussian NB	0,315	0,16	Não	Não é possível rejeitar H_0
SGD+ KNN + Gaussian NB	0,046	0,16	Sim	Rejeita-se H_0
Regressão Logística+ SGD+Gaussian NB	0,965	0,19	Não	Não é possível rejeitar H_0
Regressão Logística +Gaussian NB+KNN	0,046	0,16	Sim	Rejeita-se H_0
Multilayer Perceptron + SGD + Regressão Logística	0,201	0,14	Não	Não é possível rejeitar H_0

Tabela 24 – Teste de normalidade de Shapiro-Wilk sobre os *ensembles* treinados no córpus TweetSentBR.

Fonte: Elaborada pelo autor.

Nos testes de hipóteses da tabela 25 se buscou responder a mesma questão sobre os desempenhos dos classificadores. A hipótese nula indica que há diferença significativa entre o *ensemble* formado pelos classificadores Multilayer Perceptron, Regressão Lógisca e GaussianNB em relação às outras combinações. Os testes de t-Student e de Wilcoxon aplicados sobre os dados gerados pelos algoritmos indicam, que existe diferença estatística entre o melhor *ensemble* e a maioria dos comitês, inclusive, em relação ao comitê baseado sobre a métrica Brier Score.

Ensemble		**Tipo de Teste**	**p-value**	**p-value $<\alpha$**	**Resultado**
Multilayer Perceptron + Regressão Logística + Gaussian NB	KNN + Multilayer Perceptron + SVM	Wilcoxon	0,005	Sim	Não é possível rejeitar H_0
	Regressão Logística + Multilayer Perceptron + SVM	t-Student	0,02	Sim	Não é possível rejeitar H_0
	Gaussian NB + Multilayer Perceptron + SVM	t-Student	0,57	Não	Rejeita-se H_0
	SGD+ KNN + Gaussian NB	Wilcoxon	0,38	Não	Rejeita-se H_0
	Regressão Logística+ SGD + Gaussian NB	t-Student	0,08	Não	Rejeita-se H_0
	Regressão Logística + Gaussian NB+ KNN	Wilcoxon	0,38	Não	Rejeita-se H_0
	Multilayer Perceptron + SGD + Regressão Logística	t-Student	0,92	Não	Rejeita-se H_0

Tabela 25 – Teste de hipótese sobre os ensembles treinados no córpus TweetSentBR

Fonte: Elaborada pelo autor.

5.5 CONSIDERAÇÃO DO CAPÍTULO

Os resultados obtidos nos experimentos podem ser analisados juntamente com os resultados obtidos pelos demais autores que aplicaram Análise de Sentimentos (AS) sobre problemas do tipo categórico conforme apresentado no capítulo 3.

Neste trabalho, os resultados médios de acurácia e F-Measure são inferiores aos reportados em trabalhos semelhantes da literatura: TELES; SANTOS; SOUZA atingiu 82,73% em acurácia na classificação de tweets utilizando a técnica de smoothing, para balanceamento de classes, com classificadores solos enquanto neste trabalho foi atingido 69% de acurácia média sobre o mesmo córpus, contudo, sem aplicar a técnica de balanceamento de classes. Outro trabalho, cujas atividades são semelhantes à esta pesquisa, foi o desenvolvido por DOSCIATTI; FERREIRA; PARAISO cujo resultado obtido em acurácia foi de 60,3% utilizando SVM sobre um córpus anotado manualmente. Neste trabalho, a aplicação desse mesmo classificador sobre um córpus desbalanceado atingiu 28%. AVANÇO; BRUM; NUNES relatou 95% de acurácia média nas atividades de classificação sobre o córpus TweetSentBR expandido. Aqui, sobre o mesmo córpus sem expansão, os resultados foram de 71% de acurácia média aplicando *ensemble* dos classificadores Multilayer Perceptron, Regressão Logística e GaussianNB. BRUM no seu trabalho aplicando *ensemble* sobre o córpus TweetSentBR expandido obteve F-Measure média de 62,14%. Neste trabalho chegou-se aos 50% dessa mesma métrica.

As diferenças de configurações de experimentos, como a correção do desbalanceamento, serão realizadas em trabalhos futuros para uma melhor comparação com os trabalhos da literatura.

6 CONCLUSÃO

Este trabalho foi desenvolvido sobre a área de Análise de Sentimentos aplicada à atividade de classificação de textos curtos em português do Brasil. O problema abordado nesta pesquisa foi a aplicação da técnica de *ensemble* sobre o problema de classificação de múltiplas classes por meio da abordagem de Aprendizagem de Máquina. Para atingir o resultado, os classificadores clássicos de AM como Multilayer Perceptron, Stochastic Gradient Descent, K-Nearest Neighbor, Regressão Logística, Gaussian Naïve Bayes, Support Vector Machine e Random Forest foram analisados e utilizados como estimadores, exceto o Random Forest, para compor a técnica de *ensemble*. Todos os algoritmos foram combinados entre si, para que, por meio do sistema de voto majoritário, fosse possível obter bons resultados na atividade de classificação.

Os fundamentos desta pesquisa, bem como os trabalhos sobre os quais ela se baseia, são, em sua maioria, elaborados sobre atividades de classificação de textos em português nas suas duas variantes, e também sobre os tipos de problemas binário e categórico. Dessa forma, foi possível reunir os trabalhos elaborados no idioma português e também as suas fontes de dados de textos.

Para elaboração dos experimentos, dois córpus sobre contextos diferentes foram utilizados para treinar todos os classificadores sob as mesmas condições. O córpus 2000-tweets-BR, com alto grau de desbalanceamento entre as classes, e TweetSentBR, que contém boa distribuição entre classes, são corpora em português do Brasil compostos por três classes sendo positivo, negativo e neutro.

A primeira fase dos experimentos se concentrou na seleção do método para extração de características dos textos. Nessa etapa, os algoritmos Continuous Bag-Of-Words, Skip-gram e Term Frequency-Inverse Document Frequency foram utilizados para extrair características do córpus 2000-tweets-BR. Em seguida, três classificadores primários (SVM, SGD e Multilayer Perceptron) foram utilizados para uma atividade de classificação cuja média das acurácias foi analisada. Os resultados obtidos indicaram a seleção do Term Frequency-Inverse Document Frequency como método adequado para extrair características dos textos nesta pesquisa. Esse resultado responde à questão levantada neste trabalho quanto ao algoritmo ideal para extração de características de textos em português do Brasil.

Na segunda fase a maioria dos classificadores utilizados nesta pesquisa, exceto o Random Forest, foi treinada e testada sobre os corpora, onde os classificadores Regressão Logística e SGD obtiveram os melhores resultados em acurácia, F-Measure, Brier Score, e o menor tempo de execução médios sobre ambos corpora. Os testes de hipótese foram aplicados sobre os dados a fim de descobrir se há diferença significativa entre os melhores modelos e os demais. Dessa forma, o Regressão Logística e o SGD foram comparados

com todos os outros classificadores clássicos. Os resultados, na maioria das comparações, indicam superioridade do Regressão Logística e do SGD em relação aos classificadores concorrentes. Os experimentos sobre os classificadores em testes isolados indicaram, então, que os algoritmos Regressão Logística e SGD são algoritmos com boa performance para atividades de classificação de textos curtos sobre problemas de múltiplas classes.

Conhecendo os desempenhos de todos os algoritmos, inicia-se a terceira fase que consiste nos testes com *ensemble* de classificadores clássicos combinados entre si. Nesta fase, a combinação Multilayer Perceptron, Regressão Logística, e GaussianNB obteve melhor resultado sobre ambos corpora. Já as combinações selecionadas por meio da análise do Brier Score obtiveram resultados diferentes. O *ensemble* composto por KNN, Regressão Logística, e SVM sobre o córpus 2000-tweets-BR obteve baixo desempenho e um alto custo computacional em tempo médio de execução quando comparado com os demais *ensembles*. Por outro lado, a combinação Multilayer Perceptron, SGD, e Regressão Logística obteve bom resultado sobre o córpus TweetSentBR, porém o alto tempo de execução caracteriza alto custo computacional.

Na categoria de *ensemble* também se explorou o algoritmo Random Forest a fim de compará-lo com os resultados dos *ensembles*. Para este experimento, verificou-se que a melhor configuração do Random Forest conta com trinta estimadores, isto é, trinta Árvores de Decisão. Porém, nenhuma configuração foi superior ao *ensemble* formado pelos classificadores clássicos. Por meio das matrizes de confusão se pôde constatar que esse tipo de algoritmo mostrou-se impreciso no que diz respeito à classificação de textos curtos em português do Brasil sobre problemas de múltiplas classes.

Por fim, neste trabalho foi possível considerar o *ensemble* composto por Multilayer Perceptron, Regressão Logística e GaussianNB como um comitê de classificadores adequado para atividades de classificação de textos curtos em português do Brasil sobre problemas de múltiplas classes, em que os treinamentos e testes se baseiam sobre o vetores de características criados por meio do TF-IDF. Por outro lado, os modelos *singles*, obtiveram desempenho melhor que a melhor combinação de classificadores. Pode-se considerar que a pouca quantidade de registros nas bases e também os seus desbalanceamentos influenciaram para o baixo resultado do *ensemble* em comparação com o melhor modelo isolado, o modelo de Regressão Logística. Os resultados obtidos sobre essas atividades são baseados sobre acurácia, F-Measure, Brier Score e menor tempo de execução médios.

6.1 TRABALHOS FUTUROS

As atividades Análise de Sentimentos (AS) sobre textos em português do Brasil são desafiadores e a cada trabalho elaborado sobre esse tema surgem novas possibilidades de pesquisas. Neste trabalho se pôde verificar algumas possibilidades importantes para contribuir com a comunidade.

- Ao lidar com a escassez de bases de dados de textos curtos em português do Brasil, o problema pode ser mitigado com a técnica de composição de documentos utilizada por (ZAMPIERI; BECKER, 2013) a fim de aumentar o número de exemplos de mensagens pertencentes à mesma classe. Aplica-se essa técnica para criar aleatoriamente novas mensagens originadas de mensagens atuais.

- Com o aumento dos dados pode ser possível chegar grandes quantidades de registros, e assim, possibilitar a aplicação de aprendizagem profunda (Deep Learning) por meio do treinamento utilizando grande quantidade de mensagens.

- Os problemas de desbalanceamento das bases de dados podem ser investigados de forma mais efetiva e testados com a aplicação da técnica de smoothing bastante utilizada por CASTRO; SOUZA; OLIVEIRA.

- Também é possível, a partir dos resultados deste trabalho, iniciar uma pesquisa referente ao impacto nos resultados sobre classificação de textos curtos em português do Brasil utilizando bases de dados desbalanceadas.

- Neste trabalho aplicou-se *ensemble* com três classificadores distintos, contudo se pode aplicar testes com mais classificadores de modo que resultem em uma combinação ímpar para evitar empate. Espera-se que o aumento dos estimadores auxilie no aumento de performance em acurácia e F-Measure médias no resultado final.

- O problema de geração de diversidade para sistemas de múltiplos classificadores também pode ser um tema a seguir a partir deste trabalho. Nessa linha se pode investigar, por exemplo, quais são os impactos ao não aplicar as técnicas de reamostragem (resempling) para geração de diversidade como bootstraping ou bagging conforme aponta POLIKAR.

- A aplicação de algoritmos genéticos para geração de diversidade também é uma linha a se pesquisar a partir deste trabalho. Os algoritmos genéticos podem ser utilizados na criação de novos classificadores com diferentes características para compor um *ensemble*.

REFERÊNCIAS

ALLAHYARI, M.; POURIYEH, S.; ASSEFI, M.; SAFAEI, S.; TRIPPE, E. D.; GUTIERREZ, J. B.; KOCHUT, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

ALMASHRAEE, M.; DÍAZ, D. M.; PASCHKE, A. Emotion level sentiment analysis: The affective opinion evaluation. In: *EMSA-RMed@ ESWC*. [S.l.: s.n.], 2016.

ALVES, A. L. F.; BAPTISTA, C. D. S.; FIRMINO, A. A.; OLIVEIRA, M. G. d.; PAIVA, A. C. d. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: a case study with the 2013 fifa confederations cup. In: ACM. *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*. [S.l.], 2014. p. 123–130.

AVANÇO, L.; BRUM, H.; NUNES, M. Improving opinion classifiers by combining different methods and resources. *XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, p. 25–36, 2016.

AVANÇO, L. V. *Sobre normalização e classificação de polaridade de textos opinativos na web*. Dissertação (Mestrado) — Universidade de São Paulo, 2015.

AVANÇO, L. V.; NUNES, M. d. G. V. Lexicon-based sentiment analysis for reviews of products in brazilian portuguese. In: IEEE. *2014 Brazilian Conference on Intelligent Systems*. [S.l.], 2014. p. 277–281.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 5, p. 135–146, 2017.

BRUM, H.; NUNES, M. das G. V. Building a Sentiment Corpus of Tweets in Brazilian Portuguese. In: CHAIR), N. C. C.; CHOUKRI, K.; CIERI, C.; DECLERCK, T.; GOGGI, S.; HASIDA, K.; ISAHARA, H.; MAEGAARD, B.; MARIANI, J.; MAZO, H.; MORENO, A.; ODIJK, J.; PIPERIDIS, S.; TOKUNAGA, T. (Ed.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. ISBN 979-10-95546-00-9.

BRUM, H. B. *Expansão de recursos para análise de sentimentos usando aprendizado semi-supervisionado*. Dissertação (Mestrado) — Universidade de São Paulo, 2018.

CASTRO, D.; SOUZA, E.; OLIVEIRA, A. L. de. Discriminating between brazilian and european portuguese national varieties on twitter texts. In: IEEE. *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2016. p. 265–270.

CASTRO, D. W.; SOUZA, E.; VITÓRIO, D.; SANTOS, D.; OLIVEIRA, A. L. Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties. *Applied Soft Computing*, Elsevier, v. 61, p. 1160–1172, 2017.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

COUCH, S.; KAZAN, Z.; SHI, K.; BRAY, A.; GROCE, A. A differentially private wilcoxon signed-rank test. *arXiv preprint arXiv:1809.01635*, 2018.

DOSCIATTI, M. M.; FERREIRA, L. P. C.; PARAISO, E. C. Anotando um corpus de notícias para a análise de sentimentos: um relato de experiência (annotating a corpus of news for sentiment analysis: An experience report). In: *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2015. p. 121–130.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012.

EKMAN, P. An argument for basic emotions. *Cognition & emotion*, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992.

FILHO, P. P. B.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2013.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, American Psychological Association, v. 76, n. 5, p. 378, 1971.

FRANÇA, T. C. de; OLIVEIRA, J. Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013. In: *Proceedings of the III Brazilian Workshop on Social Network Analysis and Mining (BRASNAN)*. [S.l.: s.n.], 2014. p. 128–139.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. *Revista Brasileira de Linguística Aplicada*, SciELO Brasil, v. 13, n. 4, 2013.

FREITAS, C.; MOTTA, E.; MILIDIÚ, R.; CÉSAR, J. Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. *XI Encontro de Linguıstica de Corpus, s/p*, 2012.

FREITAS, C.; MOTTA, E.; MILIDIÚ, R.; CESAR, J. Sparkle vampire lol! annotating opinions in a book review corpus. In: CAMBRIDGE SCHOLARS PUBLISHING CAMBRIDGEˆ EUK UK. *11th Corpus Linguistics Conference*. [S.l.], 2013. p. 128–146.

GOETHE, J. W. von. *Wilhelm Meisters Lehrjahre*. [S.l.]: G. Grote'sche, 1873. v. 1.

GOKHALE, R.; FASLI, M. Matrix factorization for co-training algorithm to classify human rights abuses. In: IEEE. *2018 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2018. p. 2170–2179.

GUIMARÃES, L. M. S.; MEIRELES, M. R. G.; ALMEIDA, P. E. M. d. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. *Perspectivas em Ciência da Informação*, SciELO Brasil, v. 24, n. 1, p. 169–190, 2019.

GUL, A.; PERPEROGLOU, A.; KHAN, Z.; MAHMOUD, O.; MIFTAHUDDIN, M.; ADLER, W.; LAUSEN, B. Ensemble of a subset of knn classifiers. *Advances in data analysis and classification*, Springer, v. 12, n. 4, p. 827–840, 2018.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011.

HARTMANN, N.; AVANÇO, L.; FILHO, P. P. B.; DURAN, M. S.; NUNES, M. D. G. V.; PARDO, T. A. S.; ALUÍSIO, S. M. et al. A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In: *LREC*. [S.l.: s.n.], 2014. p. 3865–3871.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

JÚNIOR C. M., M. H. B. T. S. F. S. N.; BARBOSA, L. Paramopama: a brazilian-portuguese corpus for named entity recognition. *12th National Meeting on Artificial and Computational Intelligence (ENIAC)*, 2015.

JÚNIOR L. BARBOSA, T. M. C. M. Uma arquitetura híbrida lstm-cnn para reconhecimento de entidades nomeadas em textos naturais em língua portuguesa. *XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 2016.

KOEHRSEN, W. *Random Forest Simple Explanation*. 2017.

KRIPENDORFF, K. Reliability in content analysis: Some common misconceptions. *Human Communications Research*, v. 30, p. 411–433, 2004.

KRIPPENDORFF, K. *Content analysis; an introduction to its methodology*. [S.l.], 1980.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *International conference on machine learning*. [S.l.: s.n.], 2014. p. 1188–1196.

LIU, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LOPES, L. A.; MACHADO, V. P.; RABÊLO, R. A.; FERNANDES, R. A.; LIMA, B. V. Automatic labelling of clusters of discrete and continuous data with supervised machine learning. *Knowledge-Based Systems*, Elsevier, v. 106, p. 231–241, 2016.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, Elsevier, v. 5, n. 4, p. 1093–1113, 2014.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

MILLER, T.; PEDELL, S.; LOPEZ-LORCA, A. A.; MENDOZA, A.; STERLING, L.; KEIRNAN, A. Emotion-led modelling for people-oriented requirements engineering: The case study of emergency systems. *Journal of Systems and Software*, Elsevier, v. 105, p. 54–71, 2015.

MOHAMMAD, S. M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In: *Emotion measurement*. [S.l.]: Elsevier, 2016. p. 201–237.

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: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2018. p. 324–334.

MORAES ANDRÉ L. L. SANTOS, M. S. R. R. M. M. F. R. M. S. M. W. Classificação de sentimentos em nível de sentença: uma abordagem de múltiplas camadas para em lingua portuguesa. *XIII Encontro Nacional de Inteligência Artificial e Computacional*, 2016.

MORAES, S. M.; MANSSOUR, I. H.; SILVEIRA, M. S. 7x1-pt: um corpus extraído do twitter para análise de sentimentos em língua portuguesa (7x1-pt: a corpus extracted from twitter for sentiment analysis in portuguese language). In: *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2015. p. 21–25.

MORETTIN, L. G. *Estatística básica: probabilidade e inferência: volume único*. [S.l.]: Pearson Prentice Hall, 2010.

NASUKAWA, T.; YI, J. Sentiment analysis: Capturing favorability using natural language processing. In: ACM. *Proceedings of the 2nd international conference on Knowledge capture*. [S.l.], 2003. p. 70–77.

NEY, H. *Natural Language Processing*. 2019. Disponível em <http://www-i6.informatik.rwth-aachen.de/>.

NICULAE, V.; ZAMPIERI, M.; DINU, L.; CIOBANU, A. M. Temporal text ranking and automatic dating of texts. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. [S.l.: s.n.], 2014. p. 17–21.

OLIVEIRA, H. G.; SANTOS, A. P.; GOMES, P. Assigning polarity automatically to the synsets of a wordnet-like resource. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK. *3rd Symposium on Languages, Applications and Technologies*. [S.l.], 2014.

ORTONY, A.; CLORE, G.; COLLINS, A. The cognitive structure of emotions. cam (bridge university press. *New York*, 1988.

PAGANO, A. S.; PARAISO, E. C. Estudo exploratório de categorias gramaticais com potencial de indicadores para a análise de sentimentos. In: SBC. *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. [S.l.], 2017. p. 17–21.

PANG, B.; LEE, L. et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc., v. 2, n. 1–2, p. 1–135, 2008.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*. [s.n.], 2014. p. 1532–1543. Disponível em: <http://www.aclweb.org/anthology/D14-1162>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. *GloVe: Global Vectors for Word Representation*. 2019. Disponível em <https://nlp.stanford.edu/projects/glove/>.

POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, IEEE, v. 6, n. 3, p. 21–45, 2006.

RAMOS, I.; BERRY, D. M.; CARVALHO, J. Á. Requirements engineering for organizational transformation. *Information and Software Technology*, Elsevier, v. 47, n. 7, p. 479–495, 2005.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: PISCATAWAY, NJ. *Proceedings of the first instructional conference on machine learning*. [S.l.], 2003. v. 242, p. 133–142.

REGALADO, R. V.; AGARAP, A. F.; BALIBER, R. I.; YAMBAO, A.; CHENG, C. Use of word and character n-grams for low-resourced local languages. In: IEEE. *2018 International Conference on Asian Language Processing (IALP)*. [S.l.], 2018. p. 250–254.

ROCHA, P. A.; SANTOS, D. Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*, ICMC/USP, 2000.

ROSA, H.; CARVALHO, J. P.; BATISTA, F. Detecting a tweet's topic within a large number of portuguese twitter trends. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK. *3rd Symposium on Languages, Applications and Technologies*. [S.l.], 2014.

SAIF, H.; FERNÁNDEZ, M.; HE, Y.; ALANI, H. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

SARDINHA, T. B. *Lingüística de corpus*. [S.l.]: Editora Manole Ltda, 2004.

SHERKAT, M.; MENDOZA, A.; MILLER, T.; BURROWS, R. Emotional attachment framework for people-oriented software. *arXiv preprint arXiv:1803.08171*, 2018.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2012. p. 218–228.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information processing & management*, Elsevier, v. 45, n. 4, p. 427–437, 2009.

SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHISHMAN, R.; ALVES, I. M. Construction of a portuguese opinion lexicon from multiple resources. In: *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2011.

STATISTA. *Number of social network users in Brazil from 2017 to 2023 (in millions)*. 2019. Disponível em <https://www.statista.com/statistics/244936/number-of-facebook-users-in-brazil>.

SURYANSH. *Gradient Descent: All You Need to Know*. 2018. Disponível em <https://hackernoon.com/gradient-descent-aynk-7cbe95a778da>.

TELES, V.; SANTOS, D.; SOUZA, E. Uma análise comparativa de técnicas supervisionadas para mineraçao de opiniao de consumidores brasileiros no twitter. *Proceedings of the XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2016)*, p. 217–228, 2016.

THAKKAR, H.; PATEL, D. Approaches for sentiment analysis on twitter: A state-of-art study. *arXiv preprint arXiv:1512.01043*, 2015.

VITÓRIO, D.; SOUZA, E.; TELES, I.; OLIVEIRA, A. L. Investigating opinion mining through language varieties: a case study of brazilian and european portuguese tweets. In: SBC. *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. [S.l.], 2017. p. 43–52.

WIEBE, J.; WILSON, T.; CARDIE, C. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, Springer, v. 39, n. 2-3, p. 165–210, 2005.

WOLFGRUBER, M. *Sentiment Analyse mit lokalen Grammatiken: Wissensbasierter Ansatz zur Extraktion von Sentiments in Hotelbewertungen*. [S.l.]: Wissenschaftliche Schriften/Universitätsbibliothek LMU München, 2015.

ZACCARA, R. C. C. *Anotação e classificação automática de entidades nomeadas em notícias esportivas em Português Brasileiro*. Dissertação (Mestrado) — Universidade de São Paulo, 2012.

ZAMPIERI, M.; BECKER, M. Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, v. 5, 2013.

ZAMPIERI, M.; MALMASI, S.; DRAS, M. Modeling language change in historical corpora: the case of portuguese. *arXiv preprint arXiv:1610.00030*, 2016.