



UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ
CAMPUS LUIZ MENEGHEL - CENTRO DE CIÊNCIAS TECNOLÓGICAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ERIKSON JÚLIO DE AGUIAR

**ANÁLISE DE SENTIMENTO EM REDES SOCIAIS
UTILIZANDO COMBINAÇÃO DE CLASSIFICADORES**

BANDEIRANTES-PR

2017

ERIKSON JÚLIO DE AGUIAR

**ANÁLISE DE SENTIMENTO EM REDES SOCIAIS
UTILIZANDO COMBINAÇÃO DE CLASSIFICADORES**

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Norte do Paraná para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. André Luis A. Menolli

BANDEIRANTES-PR

2017

ERIKSON JÚLIO DE AGUIAR

**ANÁLISE DE SENTIMENTO EM REDES SOCIAIS
UTILIZANDO COMBINAÇÃO DE CLASSIFICADORES**

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Norte do Paraná para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Prof. Dr. André Luis A. Menolli
Universidade Estadual do Norte do Paraná
Orientador

Prof. Dr. Bruno Squizato Faiçal
Universidade Estadual do Norte do Paraná

Prof. Me. Glauco Carlos Silva
Universidade Estadual do Norte do Paraná

Dedico este trabalho aos meus pais e a memória de Edna Aguiar, Emerson Amorelli e Maria de Lourdes, pessoas muito importantes, que me viram entrar na graduação, mas não puderam me ver concluir.

AGRADECIMENTOS

Sou grato aos meus pais que me apoiaram todos os dias durante esses 4 anos de curso, sempre sendo atenciosos, pacientes, companheiros e batalhadores, fazendo tudo sempre pensando no melhor para mim. Agradeço ao meu orientador André Menolli, por ter me auxiliando durante esses anos de projetos, bem como, em meu trabalho de conclusão de curso, estando disponível, sendo paciente e me passando vários ensinamentos preciosos. Também gostaria de registrar meu reconhecimento ao professor Bruno Faiçal, que me ajudou em muito no meu trabalho de conclusão de curso, principalmente na parte que envolvia aprendizado de máquina, mas não só neste projeto, como também me tirou muitas dúvidas durante o curso, me deixando ainda mais curioso sobre os temas relacionados a este trabalho, me instigando querer saber mais e futuramente fazer um mestrado e doutorado com ênfase em aprendizado de máquina. Gostaria de deixar meus agradecimentos a todos os amigos que fiz e estiveram sofrendo juntos na graduação, fazendo trabalhos, compartilhando conhecimento, "zuando" e é claro apoiando um ao outro em vários momentos.

"Toda a nossa ciência, comparada com a realidade, é primitiva e infantil – e, no entanto, é a coisa mais preciosa que temos". (Albert Einstein)

AGUIAR, E. J.. **Análise de Sentimento em Redes Sociais Utilizando Combinação de classificadores**. 74 p. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual do Norte do Paraná, Bandeirantes-PR, 2017.

RESUMO

As redes sociais são um dos maiores meios de comunicação atualmente, dentro delas são gerados todo o tipo de informação que podem estar relacionadas com pessoas, acontecimentos, lugares e diversos outros fatores. Os usuários dentro das redes sociais expressam suas opiniões de forma livre, deixam sua personalidade e preferências expostas para o mundo. Logo, a análise de sentimentos em redes sociais vem se tornando mais frequente, visto que esse tipo de informação pode ser importante para uma empresa descobrir as preferências de seus clientes em relação aos seus produtos e serviços. O trabalho propõe um método para estimar sentimentos em redes sociais para a língua portuguesa, tendo como foco o Twitter; o método utiliza de uma abordagem baseada em algoritmos de aprendizado de máquina que é denominado comitê, no qual combina a predição de um conjunto de seis algoritmos e define o valor predito como o mais votado entre eles, considerando que o voto entre eles tem peso. Para executar esse processo, são realizados alguns testes com uma base de dados de *tweets* em português já rotulada com as classes: negativo (-1), neutro (0) e positivo (1). Para avaliar o desempenho das técnicas, foi utilizado as seguintes métricas de desempenho: acurácia, precisão, *recall*, f1-score e erro, a base de teste foi classificada usando os algoritmos e analisados os resultados de acordo com as métricas propostas isoladamente. Além disso, também foram analisados os serviços de análise de sentimentos disponíveis no mercado, IBM Watson e Microsoft *Text Analytics*. O método proposto obteve uma acurácia de aproximadamente 86 % sendo superior aos outros nesse aspecto. O próximo passo foi realizar a análise estatística utilizando de algumas técnicas, com o propósito de verificar se o método proposto tem diferença estatística para as outras abordagens apresentadas, assim concluiu-se que o método proposto tem diferença apenas para os métodos: árvore de decisão, IBM Watson e Microsoft *Text Analytics*, portanto, para as outras técnicas são equivalentes estatisticamente; os resultados desses testes foram cruciais para determinar se o método proposto tem diferenças significativas para os demais.

Palavras-chave: Aprendizado de Máquina, Análise de Dados, Classificação de Sentimentos, Redes Sociais

AGUIAR, E. J.. **Sentiment Analysis in Social Networks Using Classifiers Ensembles**. 74 p. Final Project (Bachelor of Science in Computer Science) – State University Northern of Parana , Bandeirantes–PR, 2017.

ABSTRACT

Social networks are one of the largest means of communication currently, within which are generated all kinds of information that may be related to people, events, places and various other factors. Users within social networks express their opinions freely, leaving their personality and preferences exposed to the world. Therefore, sentiment analysis in social networks is becoming more frequent, since this type of information may be important for a company in discovering the preferences of its customers in relation to their products and services. The study proposes an approach to estimate sentiment in social networks for the portuguese language, focusing on Twitter; the method uses a machine learning algorithm approach that is called a committee, in which it combines the prediction of a set of six algorithms and defines the predicted value as the most voted among the algorithms, considering that the votes of the algorithms have weight. To perform this process, some tests are performed with a database of Portuguese tweets already labeled with the classes: negative (-1), neutral (0) and positive (1). To evaluate the performance of the techniques, the following performance metrics were used: accuracy, precision, recall, f1-score and error, the test base was classified using the algorithms and the results were analyzed according to the metrics proposed individually. In addition, sentiment analysis services available on the market were also tested, IBM Watson and Microsoft Text Analytics. The proposed method obtained an accuracy of approximately 86 % being superior to others in this respect. The next step was to perform the statistical analysis using some techniques, in order to verify if the proposed method has statistical difference for the other approaches presented, so it was concluded that the method has difference only for the following techniques: decision tree, IBM Watson and Microsoft Text Analytics, therefore demonstrating being statistically equivalent; the results of these tests were crucial to determine the significant differences of the method proposed for other techniques.

Keywords: Machine Learning, Data Analytics, Sentiment Classification, Social Networks

LISTA DE ILUSTRAÇÕES

Figura 1 – Hierarquia do Aprendizado Indutivo. Adaptado de (REZENDE, 2003).	23
Figura 2 – Esquema Genérico de Uma Árvore de Decisão. Adaptado de (FACELI et al., 2011).	27
Figura 3 – Matriz de Confusão. Adaptado de (GAMA, 2010).	30
Figura 4 – Matriz de Confusão Para Três Classes.	31
Figura 5 – Exemplo de <i>tweet</i> publicado pelo usuário @CavalinhodoFla_	34
Figura 6 – Técnicas de Análise de Sentimento. Adaptado de (MEDHAT; HASSAN; KORASHY, 2014).	36
Figura 7 – Estrutura da Pesquisa	41
Figura 8 – Método Proposto	47
Figura 9 – Boxplot das Acurácias dos Algoritmos Utilizados	52
Figura 10 – Gráfico ROC	52
Figura 11 – Matriz de Confusão do Comitê	53
Figura 12 – Estruturação dos Testes	53
Figura 13 – Matriz de Confusão do TSviz	56
Figura 14 – Matriz de Confusão do Watson	58
Figura 15 – Matriz de Confusão do <i>Text Analytics</i>	59
Figura 16 – Matriz de Confusão do Naive Bayes	71
Figura 17 – Matriz de Confusão do SVM	71
Figura 18 – Matriz de Confusão da Árvore de Decisão	72
Figura 19 – Matriz de Confusão da Random Forest	72
Figura 20 – Matriz de Confusão da Regressão Logística	73

LISTA DE TABELAS

Tabela 1 – Resultados das Métricas de Desempenho do Algoritmos Utilizados . . .	51
Tabela 2 – Tabela Com os P-valores Gerados a Partir do Teste <i>Post-hoc</i> de Nemenyi Para o Método Proposto	54
Tabela 3 – Tempo de Execução dos Algoritmos Utilizados	55
Tabela 4 – Métricas de Desempenho do TSviz	56
Tabela 5 – Tabela Com os P-valores Gerados a Partir do Teste <i>Post-hoc</i> de Nemenyi Para o TSviz	57
Tabela 6 – Métricas de Desempenho do Watson e <i>Text Analytics</i>	58
Tabela 7 – Tabela Com os P-valores Gerados a Partir do Teste <i>Post-hoc</i> de Nemenyi Para o Watson	59
Tabela 8 – Tabela Com os P-valores Gerados a Partir do Teste <i>Post-hoc</i> de Nemenyi Para o Microsoft <i>Text Analytics</i>	59

LISTA DE ABREVIATURAS E SIGLAS

OVR	Um Contra Todos
SVM	<i>Support Vector Machine</i>
ROC	<i>Receiver Operating Characteristic</i>
TFP	Taxa de Falsos Positivos
TVP	Taxa de Verdadeiros Positivos
AUC	<i>Area Under Curve</i>
PLN	Processamento de Linguagem Natural
NRC	National Research Council
NLTK	<i>Natural Language Toolkit</i>
PANAS-x	<i>Positive Affect Negative Affect Scale</i>
API	<i>Application Programming Interface</i>
REST	<i>Representational State Transfer</i>
SKLEARN	<i>Scikit-Learn</i>
CSV	<i>Comma-separated values</i>

SUMÁRIO

1	INTRODUÇÃO	21
1.1	Objetivos	21
1.2	Justificativa	22
2	FUNDAMENTAÇÃO TEÓRICA	23
2.1	Aprendizado de Máquina	23
2.1.1	Classificação Multiclasses	24
2.1.2	Técnicas de Classificação	24
2.1.3	Métricas de Desempenho para Classificadores	29
2.2	Processamento de Linguagem Natural	33
2.3	Rede Social Twitter	33
2.4	Análise de Sentimento em Comentários do Twitter	34
3	TRABALHOS RELACIONADOS	37
4	ESTRUTURAÇÃO DA PESQUISA	41
4.1	Planejamento Inicial	41
4.2	Fase Exploratória	41
4.3	Desenvolvimento	43
4.4	Avaliação dos Classificadores	45
5	MÉTODO PROPOSTO	47
5.1	Pré Processamento	47
5.2	Comitê	48
5.3	Ajuste dos Pesos	48
5.4	Avaliação do Desempenho	49
5.4.1	Sentimento	49
6	RESULTADOS	51
6.1	Análise da Abordagem Proposta	51
6.2	Análise de Técnicas Semelhantes	55
7	CONCLUSÃO	61
	REFERÊNCIAS	63

APÊNDICES	69
APÊNDICE A – MATRIZ DE CONFUSÃO DOS MÉTODOS UTILIZADOS	71

1 INTRODUÇÃO

Atualmente, as pessoas utilizam as redes sociais de várias formas, tanto para se manter informados sobre as novidades que ocorrem em seu meio social, quanto para conversar com seus amigos e colegas de trabalho. Assim as redes sociais podem ser vistas como um complemento da nossa interação social convencional, o adicional é que utilizamos de recursos tecnológicos para a troca de informação, como os computadores e *smartphones*. Algumas das redes sociais atuais são: (i) Facebook, (ii) LinkedIn, (iii) Twitter, (iv) Instagram, (v) Snapchat. Estas tem como finalidade desde de enviar mensagens simples, fotos, vídeos curtos ou longos até publicações de outro usuários.

As redes sociais mais populares e que tem uma diversidade maior de conteúdos que podem ser publicados são o Facebook e o Twitter. Todo os dias, transitam em torno de um bilhão de mensagens no Twitter, publicadas por mais de 100 milhões de usuários. Simultaneamente os 1,23 bilhão de usuários do Facebook curtiram publicações seis bilhões de vezes (ZAGO; SILVA, 2014). Desse modo os seus usuários compartilham de todo o tipo de conteúdo, bem como imagens, textos, curtidas e outros.

Dentro das redes sociais os usuários formam um perfil próprio, a partir de suas opiniões expressas dentro da rede que na grande maioria é por forma de texto. Logo, podemos extrair várias informações úteis desse texto, assim surgem abordagens como Gonçalves et al. (2012), que tem a finalidade de analisar essas opiniões e gerar conhecimento que serão pertinentes em alguma área.

A análise de sentimento ou análise de opinião segundo Rosa (2015) tem o objetivo de determinar a polaridade do sentimento do usuário sobre um assunto o classificando em positivo, negativo ou neutro. Contudo, pode variar em outras aplicações podendo ter mais ou menos classes.

Nos últimos anos a análise de sentimento tem sido tema de diversas pesquisas tanto no ambiente acadêmico como em (DING; LIU; ZHANG, 2009; HOGENBOOM et al., 2013), bem como no mercado, como afirma Rosa (2015). A análise de sentimentos pode ser de grande valia para as empresas que querem saber o que os clientes pensam sobre seus produtos. Assim as mesmas podem identificar uma inconformidade com algum produto e tomar as medidas necessárias antecipadamente antes que se torne um caos e ocorra grande perda financeira.

1.1 Objetivos

A partir dessas perspectivas, este trabalho tem o objetivo desenvolver uma abordagem para estimar o sentimento nos comentários do Twitter na língua portuguesa.

Para a realização dessa tarefa será necessário os seguintes objetivos específicos:

- Selecionar técnicas de aprendizado de máquina que são mais utilizadas segundo (SILVA, 2016a; GRANDIN; ADAN, 2016; AUGUSTYNIAK et al., 2014) para estimar o sentimento expresso em textos;
- Implementar os algoritmo de aprendizado de máquina na linguagem Python de forma individual;
- Combinar as técnicas de classificação utilizadas, a fim de obter o melhor resultado da predição dos classificadores, que será decido por votação;
- Comparar os algoritmos individualmente com o resultado da combinação dos mesmos;
- Avaliar o desempenho dos classificadores;
- Analisar os resultados obtidos.

1.2 Justificativa

Com o intensivo uso das redes social por todo o tipo de pessoa, pode-se redescobrir o quanto é importante a análise de dados nesse meio, como afirma França et al. (2014), assim começou a aumentar o números de pesquisas sobre o assunto, pois nesses conjuntos de dados podem ter informação útil que contribua com o nosso meio social. Contudo, o número de trabalhos sobre o tema ainda é bem limitado, como declara Silva (2016b) os trabalhos sobre a área de análise de sentimento, mineração de opinião e outros temas relacionados, começaram a surgir com mais frequência após os anos 2000. Devido a complexidade de se trabalhar com texto, principalmente de redes sociais que é desestruturado e o processo para fazer o tratamento do mesmo exige um custo de tempo alto.

Afinal, os trabalhos sobre o assunto comumente abordam problemas para classes binárias, alguns podem até englobar uma terceira classe, entretanto não ultrapassam essa fronteira. Em relação aos trabalho correlatos, Silva (2016b) assegura que temos uma gama pequena de pesquisas que envolvem o sentimento classificado com polaridade, isto é, o sentimento não é somente positivo e negativo, mas tem uma força do quanto ele é positivo ou negativo. Além disso, outro ponto válido a se considerar é o pequeno número de explorações na literatura que envolvem a língua portuguesa de Portugal, tal como a do Brasil, portanto necessita-se de trabalhos que levantem essas propostas. Pois a opinião gerada a partir do comentário de uma pessoa no Twitter, pode fornecer informação que contribua para geração de outros tipos de trabalhos que apliquem a análise de sentimento a outras problemáticas.

2 FUNDAMENTAÇÃO TEÓRICA

Para a o desenvolvimento deste trabalho, foi fundamental o estudo de vários conceitos relacionados a redes sociais, processamento de linguagem natural, métodos de análise de sentimento em redes sociais e aprendizado de máquina. Esta seção tem o objetivo de apresentar alguns assuntos que auxiliarão no entendimento dos conceitos nas próximas seções do trabalho.

2.1 Aprendizado de Máquina

O Aprendizado de máquina é uma subárea de inteligência artificial, cujo tem como propósito desenvolver métodos computacionais baseado em aprendizado, isto é, criar técnicas que façam com que as máquinas aprendam determinados conceitos. Para realizar essa tarefas esses métodos se baseiam em experiências anteriores, no qual são treinados com um conjunto de dados, ou melhor, aprendem algum conceito, para que melhorem seu desempenho na execução dessa tarefa (REZENDE, 2003; MITCHELL, 1997).

O aprendizado indutivo, é basicamente o aprendizado de máquina, que é executado a partir da observação de exemplos anteriores, assim em Rezende (2003) apresenta a hierarquia do aprendizado indutivo como mostra a Figura 1. As principais abordagens utilizadas segundo Castanheira (2008) são:

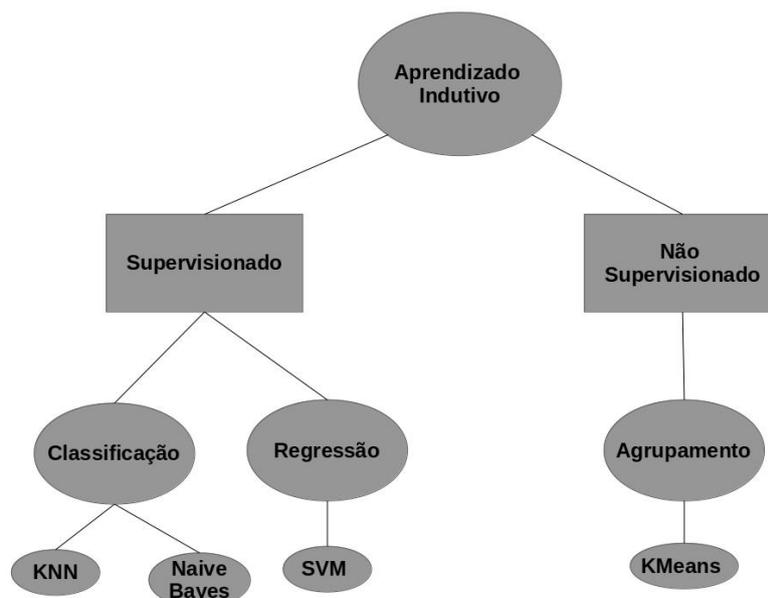


Figura 1 – Hierarquia do Aprendizado Indutivo. Adaptado de (REZENDE, 2003).

- **Classificação:** Define o conjunto de dados em algumas classes, cujo os atributos inseridos nelas recebem seu rótulo. Quando se insere novos atributos no contexto é realizado um processo que verifica a correlação dele com cada classe, para dessa forma inserir na classe de maior correlação com o mesmo;
- **Regressão:** É similar a classificação, contudo trabalha com um conjunto contínuo de valores.

Aprendizado Não Supervisionado

- **Agrupamento:** O agrupamento faz a junção de uma população desuniforme em grupos uniformes, que são formados a partir de características semelhantes dos indivíduos da população, contrário a classificação que tem classes pré definidas, que os indivíduos são inseridos.

2.1.1 Classificação Multiclasses

O aprendizado de máquina abstrai todo o tipo de dado, isto é, o transforma para seguir um padrão adequado para os algoritmos, assim os dados podem ser reconhecidos. Quando se trabalha com classificação temos uma base de treinamento rotulada com as classes que o preditor irá reconhecer, contudo nem sempre as classes que devem ser preditas são binárias. Logo, algumas vezes é mais interessante utilizar de classes binárias, pelo fato de alguns algoritmos serem preferencialmente binários, porém também é válido pela redução da complexidade computacional, métodos com mais classes são mais custosos (FÜRNKRANZ, 2002; FACELI et al., 2011).

Portanto para resolver esse problema podemos utilizar de algumas técnicas, como o método um contra todos, do inglês *one-against-all* (OVR). Essa técnica é representada por uma matriz de $n \times n$, no qual temos k classes, que são criados k preditores binários, cujo cada um tem a tarefa de identificar uma classe i das demais (LORENA; CARVALHO, 2008).

2.1.2 Técnicas de Classificação

Este trabalho terá como foco o aprendizado supervisionado, mais especificadamente as abordagens de classificação. É constituído de diversos algoritmos, bem como: (i) classificador Naive Bayes, (ii) Árvore de Decisão, (iii) Random Forest, (iv) Support Vector Machine (SVM) e (v) Regressão Logística, que são descritos a seguir:

Classificador Naive Bayes

O Naive Bayes é um classificador probabilístico, baseado no teorema de Bayes, que é representado pela Equação 2.1 (NORVIG; RUSSELL, 2014). Esse classificador manipula

uma fórmula matemática herdada de Bayes, que tem como saída o valor v_{nb} , cujo seu resultado é o rótulo de classe predita (MITCHELL, 1997). A fórmula do classificador Baysiano é definido pela Equação 2.2.

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i|\mathbf{d}) \quad (2.1)$$

$$v_{nb} = \underset{v_j \in V}{\operatorname{argmax}} (P(v_j) \prod_i P(v_i|v_j)) \quad (2.2)$$

Esse método tem variantes, como o Multinomial e o Bernoulli, no caso o Multinomial, é bastante utilizado para a classificação de textos, no qual a técnica é descrita como a contagem da frequência do vetor de palavras. Esse vetor é definido como $\theta_y = (\theta_{y1}, \theta_{y2}, \dots, \theta_{y,})$, cada valor θ_{yi} é a probabilidade da palavra ocorrer dentro do texto. A frequência relativa é descrita pelo parâmetro a seguir (SKLEARN, 2017) :

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (2.3)$$

Onde:

- T é o conjunto de treinamento;
- i é a frequência da característica;
- y é a classe;
- α é a suavização;
- $N_{yi} = \sum_{x \in T} x_i$;
- $N_y = \sum_{i=1}^{|T|} N_{yi}$.

Máquina de Vetor de Suporte

O algoritmo Máquina de Vetor de Suporte (SVM) é baseado na teoria do aprendizado estatístico. Os SVMs usam um mapeamento não linear para converter a base utilizada para o treinamento em uma outra com dimensão superior. Esse algoritmo busca o caso ótimo para a divisão do limite de decisão. Pode ser comparado as Redes Neurais Artificiais, tem um ótimo desempenho na predição, assim os SVMs vem recebendo grande atenção na comunidade científica (HAN, 2005; FACELI et al., 2011).

O aprendizado estatístico é utilizado nos SVMs segundo Santos (2002), devido a necessidade de verificar matematicamente a precisão das técnicas em classificar valores que não estão na base de treinamento. Portanto essa teoria surge para suprir a necessidade

de descrever os fatores que devem ser supervisionados para proporcionar um desempenho superior. Para que um classificador faça uma predição mais precisa mesmo que um valor não seja conhecido nos exemplos de treinamento, o mesmo deve trabalhar com medidas a fim de reduzir a perda, no qual o valor desejado da perda é definido pela fórmula do risco funcional (SANTOS, 2002) :

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y) \quad (2.4)$$

Os SVMs são classificadores binários, contudo para realizar a transformação de um classificador multiclases para binários, podemos utilizamos da política um contra todos, bem como foi feito neste trabalho (FACELI et al., 2011).

Os problemas linearmente separáveis, isto é quando se é capaz de dividir os dados conforme suas classes por um hiperplano, que é definido pela Equação 2.5, são resolvidos pelos SVMs lineares. Com a finalidade de dividir o conjunto de dados, são definidas duas linhas, que são as fronteiras entre as classes, em que é formada por objetos do conjunto de dados que estão sobre essas linhas, mais conhecidos como vetor de suporte (Turchetti Maia, 2007; SOARES, 2008; OLIVEIRA, 2012). Como também, um detalhe que deve ser considerado sobre essa técnica, segundo Oliveira (2012) um SVM está no melhor caso quando a distância entre as duas bordas é máxima.

$$f(x) = \mathbf{w} \cdot x + b = 0 \quad (2.5)$$

O Kernel dos SVMs por sua vez, são empregados por ter um método de cálculo simples e ser capaz de representar espaços obscuros. Para realizar sua computação utiliza-se do produto escalar de dois vetores que estão no espaço vetorial. Os valores do Kernel podem ser lineares, *Radial-Basis Function*, Sigmodal e Gaussiano na maioria dos casos (LORENA; CARVALHO, 2008).

Árvore de Decisão

É um algoritmo de aprendizado de máquina, que tem o objetivo receber valores de entrada que podem ser contínuos ou discretos e fornecer como resposta a decisão tomada pela técnica, em que pode se obter como resultado somente dois valores possíveis: negativo ou positivo (NORVIG; RUSSELL, 2014). A Figura 2 apresenta a forma de uma árvore de decisão visualmente, no qual \mathbf{x} e \mathbf{y} são os atributos e \mathbf{A} , \mathbf{B} , \mathbf{C} e \mathbf{D} são os objetivo.

A Árvore de Decisão um dos algoritmos de inferência mais utilizados, por ser simples e movido por expressões de "se-então", vindas da lógica proposicional. O mesmo se baseia na estratégia dividir para conquistar, isto é, um problema complexo é dividido em vários sub-problemas mais simples e é resolvido cada pedaço da árvore, ou melhor cada problema é solucionado de forma recursiva (MITCHELL, 1997; SILVA, 2005).

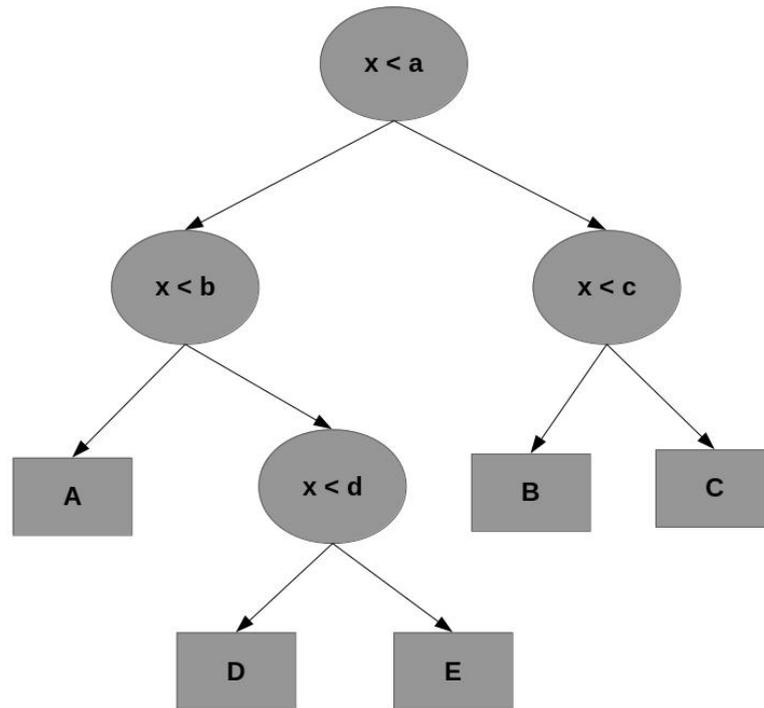


Figura 2 – Esquema Genérico de Uma Árvore de Decisão. Adaptado de (FACELI et al., 2011).

Para particionar corretamente uma árvore de decisão é interessante utilizar de alguma política, em que temos algumas estabelecida na literatura. Bem como, o índice Gini, que tem como objetivo mensurar a pureza da informação contida nos nós, isso irá definir quantas classes vão para cada nó. No caso de uma maior impureza o número de classes inseridas naquele nó tem que aumentar. A fórmula que auxilia nesse processo de identificação é dada por (SILVA, 2005):

$$IG = 1 - \sum_{i=1}^c p_i^2 \quad (2.6)$$

- **IG** : É o valor do índice de Gini;
- p_i : Frequência da classes em cada nó;
- **c** : é o número de classes.

As árvores de decisão são utilizadas em muitas aplicações, bem como na área médica, e que pode ser utilizada para fazer o diagnóstico de doenças, onde temos como entrada as várias do paciente, ou seja os sintomas e tem como resposta um valor positivo ou negativo para determinada doença que está no nó final da árvore (MITCHELL, 1997).

Random Forest

A Random Forest é um algoritmo que gera árvores aleatórias desenvolvido por (BREIMAN, 2001), no qual utiliza de uma ideia semelhante a da árvore de decisão, contudo de forma mais otimizada. Esse algoritmo consiste basicamente de um comitê de classificadores de árvore de decisão, isto é, tem suas predições combinadas e decide por uma votação uniforme entre os classificadores, o rótulo mais votado que é selecionado como valor predito, isto é chamado de floresta aleatória (GÓMEZ, 2012).

A definição matemática segundo Breiman (2001), consiste fundamentalmente de uma coleção de k árvores aleatórias, no qual temos θ que é um conjunto de números aleatórios que está entre 1 e K . As florestas aleatórias são definidas pela seguinte forma:

$$(h(\mathbf{x}, \theta_k), k = 1, \dots) \quad (2.7)$$

Onde θ_k , são vetores aleatórios que são distribuídos de forma independente para que que direcione os classificadores a votar na classe mais popular \mathbf{x} .

A margem mede a medida em que o número médio de votos em X , Y para a classe certa excede o voto médio para qualquer outra classe. Quanto maior a margem, mais confiança na classificação.

Para que o algoritmo tenha uma maior precisão em acertar a classe mais popular é calcula a margem. É definida pelo número médio de votos nos vetores \mathbf{X} e \mathbf{Y} para qualquer outra classe. A Equação 2.8 expressão a fórmula da margem (BREIMAN, 2001).

$$mar(\mathbf{X}, \mathbf{Y}) = \bar{I} - Max(\bar{J}) \quad (2.8)$$

- \bar{I} : É a média de $h_k(\mathbf{X}) = \mathbf{Y}$;
- \bar{J} : É a média de $h_k(\mathbf{X}) = j$. Onde $j \neq \mathbf{Y}$.

O algoritmo de floresta aleatórias utiliza da técnica de amostragem arbitrária com reposição, Boosting, para escolher suas amostras treino de forma eventual. Esse algoritmo tem essencialmente dois parâmetros de entrada, o K que é o número de árvores e n o número de atributos que devem ser considerados (FACELI et al., 2011).

Regressão Logística

O modelo de Regressão Logística é utilizada na análise multivariada de dados, mas também é uma técnica de análise de regressão, que utiliza da máxima verossimilhança para fazer predições, cujo tem como resultado uma variável categórica. A título de exemplo, seria a ausência ou presença de uma doença, no caso de um problema binário, entretanto

também pode abordar problemas com mais de uma categoria (DAYTON, 1992). Assim para casos binários a fórmula do modelo logístico é definida como (BISSACOT, 2015):

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (2.9)$$

Esse pode ser um modelo de regressão linear generalizado, em que é constituído por algumas componentes: (i) a aleatória combinada com variáveis preditoras, (ii) a sistêmica, que relaciona as variáveis preditoras com a variável resposta e (iii) a função de ligação, chamada função *logit* 2.10 que correlaciona os valores desejados com a variáveis preditoras (RAPOSO, 2014; KRZANOWSKI, 1998).

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x \quad (2.10)$$

Os modelos apresentados anteriormente são modelos univariados, para modelos multivariados, ou seja que trabalham com mais de duas classes de resposta na variável discriminante, segundo Bissacot (2015) a função *logit* para esse modelo é:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (2.11)$$

Assim, o modelo logístico é descrito como:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (2.12)$$

Um ponto importante a se destacar nesse modelo é o parâmetro *odds ratio* (razão de possibilidades), no qual é definido como o quanto é provável um dado de entrada pertencer a uma classe de saída. Como exemplo, para um modelo de gripe, que tem como saída positivo, negativa e neutra. Tendo como entrada sintomas como: cefaleia, tosse, narinas congestionadas, que resulta em uma possibilidade = 4 para positivo. Assim a partir da análise, vemos que tem uma probabilidade alta do indivíduo estar com gripe (BISSACOT, 2015).

2.1.3 Métricas de Desempenho para Classificadores

As métricas de desempenho são válidas para verificar o comportamento do preditor, assim serve como base para avaliar um classificador (FACELI et al., 2011). Os classificadores não conseguem acertar todos os casos por ter uma grande variabilidade casos seguindo outro padrão, desse modo temos a matriz de confusão que indica os acertos e erros de um classificador, a Figura 3 mostra uma matriz de confusão para duas classes.

Os valores contidos na matriz de confusão são descritos como (GAMA, 2010):

		Classe Predita	
		P	N
Classe Verdadeira	P	VP	FN
	N	FP	VN

Figura 3 – Matriz de Confusão. Adaptado de (GAMA, 2010).

- **VP:** São os verdadeiros positivos, o total de valores classificados corretamente que pertencem a classe positiva;
- **VN:** São os verdadeiros negativos, o número total de valores classificados corretamente que pertencem a classe negativa;
- **FP:** São os verdadeiros negativos, o total de valores classificados incorretamente que pertencem a classe positiva;
- **FN:** São os verdadeiros negativos, o total de valores classificados incorretamente que pertencem a classe negativa;
- **n:** São todos os valores classificados, onde $n = VP + VN + FP + FN$.

Com a matriz de um preditor formada, podemos extrair algumas medidas de desempenho, como Monard e Baranauskas (2003) apresenta:

- **Acurácia:** É a taxa de acerto do classificador e é definido pela Equação 2.13.

$$ac = \frac{FP + FN}{n} \quad (2.13)$$

- **Precisão:** É dimensão de exemplos positivos classificados corretamente dentro todos os positivos, é definido pela Equação 2.14.

$$prec = \frac{VP}{VP + FP} \quad (2.14)$$

- **Sensibilidade:** É a taxa de acerto ocorrido nos indivíduos da classe positiva, é definido pela equação 2.15.

$$sens = \frac{VP}{VP + FN} \quad (2.15)$$

- **Especificidade** É a taxa de acerto ocorrido nos indivíduos da classe negativa, é definido pela Equação 2.16. É complementar a taxa de falsos positivos.

$$esp = \frac{VN}{VN + FP} = 1 - TFP \quad (2.16)$$

- **F1 Score:** É a média harmônica entre a sensibilidade e especificidade, é definido pela Equação 2.17.

$$F1 = \frac{2 * prec * sens}{prec + sens} \quad (2.17)$$

Já para multiclassés utilizamos de algumas modificações nas fórmulas e na matriz de confusão, como mostra a Figura 4.

		Classe Predita		
		-1	0	1
Classe Verdadeira	-1	a	b	c
	0	d	e	f
	1	g	h	i

Figura 4 – Matriz de Confusão Para Três Classes.

Para fazer o cálculo acurácia, precisão e sensibilidade, vemos em [Lunardi, Viterbo e Bernardini \(2016\)](#) as fórmulas que são apresentadas abaixo:

- A **acurácia** para multiclassés, é calculada a partir da soma dos valores da diagonal principal matriz de confusão dividido pela soma de todos os valores da matriz.

$$ac = \frac{a + e + i}{(a + b + c + d + e + f + g + h + i)} \quad (2.18)$$

- A **precisão** é calculada por classe, no qual é definida pelo número de acertos da classe dividido pelo número de predições positivas, que é a coluna da matriz. Por exemplo para a classe -1, temos a seguinte equação:

$$prec = \frac{a}{(a + d + g)} \quad (2.19)$$

- A **sensibilidade** também é calculada por classe, e é definida pelo número de acertos da classe dividido pelo número de exemplos positivos, que é a linha da matriz. Por exemplo para a classe 1, temos a seguinte equação:

$$sens = \frac{i}{(g + h + i)} \quad (2.20)$$

Uma abordagem válida para avaliar o desempenho de classificadores é a análise ROC (*Receiver Operating Characteristic*), ou seja é um método para avaliar, organizar e selecionar técnicas de predição. Análise ROC quando surgiu não tinha o objetivo de avaliar algoritmos de aprendizado de máquina, foi feito basicamente para modelos de diagnóstico médico, no qual foi aplicado a economia, psicologia e outros, até chegar ao aprendizado de máquina. É bastante utilizado em aprendizado de máquina quando temos classes desbalanceadas, em que o número de amostras de cada classe tem uma variabilidades grande (PRATI; BATISTA; MONARD, 2008).

Para realizar a análise recorre-se ao gráfico ROC que é uma forma visual de avaliar um classificador, esse é um modelo bidimensional baseado em probabilidade. Cujo o eixo das abscissas indica a taxa de falsos positivos (TFP) e o eixo das ordenadas a taxa de verdadeiros positivos (TVP). Anteriormente foi apresentado o conceito de matriz de confusão como mostra a Figura 3, assim podemos fazer o paralelo com os valores do TFP e TVP (PRATI, 2006; FACELI et al., 2011). No qual podem ser descritos como:

$$TFP = 1 - \frac{VN}{VN + FP} = 1 - esp \quad (2.21)$$

$$TVP = \frac{VP}{VP + FN} = sens \quad (2.22)$$

O gráfico ROC produz valores entre 0 e 1, que formará uma curva baseado nas predições do classificador. Um fator importante para mesurar esse classificador é a área abaixo da curva (AUC), em inglês *Area Under Curve*. Cujos valores variam entre 0 e 1, no qual quando mais próximo do valor 1, melhor será o classificador. Nesse gráfico é definida uma linha limite, que indicará qual o valor mínimo de AUC deve ser considerado

para que o mesmo ainda seja viável para a aplicação (PRATI; BATISTA; MONARD, 2008; FACELI et al., 2011).

2.2 Processamento de Linguagem Natural

O processamento de linguagem natural vem sendo utilizado em diversas aplicações no mercado dos aplicativos, bem como em bots do Facebook, na assistente pessoal do google entre outras aplicações. Portanto a PLN (processamento de linguagem natural) é a utilização de sistemas computacionais para a processamento e interpretação de sentenças em linguagem natural, diferente das linguagens entendidas pelo computador, como C, C++, Java, Python e mais uma infinidade de linguagens (PEREIRA, 2014). Assim podemos utilizar do PNL para processar sentenças com a finalidade de obter um texto limpo para a máquina manipular de forma mais fácil, bem como foi realizado nesse trabalho na etapa de normalização do texto.

Para processar uma sentença textual é necessário seguir alguns estágios, que resultará no texto limpo para utilização. Esses passos são: (i) Extração dos *tokens* e (ii) Remoção das *Stop words*, pode ser incluídos mais fases, como o stemming e a substituição de sinônimos. Mas afinal cada uma dessas fases segundo Rodrigues (2016) realiza as seguintes tarefas:

- **Extração dos *Tokens*:** Os *tokens* são partes significativas extraídas a partir do fluxo de entrada do texto. Nessa fase podemos aplicar expressões regulares, que são estruturas padronizadas que representam um conjunto de caracteres, que serão inseridos no conjunto de *tokens*. Assim podemos remover informação adicional como (. , ; ! ? []), porém isso depende do contexto;
- **Remoção das *Stop Words*:** Nesse passo são removidas as palavras sem informação adicional significativa, como artigos, preposições e conjunções. Tendo como exemplo, as palavras, "para", "de", "como", "as", "e" e mais uma infinidade de sentenças.

2.3 Rede Social Twitter

Desde da pré história os homens se organizam em sociedade, compartilhando suas tarefas como caça, cultivo de alimentos, criação do gado e várias outras tarefas, que contribuem para o bom funcionamento do meio social. Nos dias de hoje a nossa sociedade também coopera entre si, com a produção, venda e compra de produtos, é um ciclo que move o contexto social. Pelo grande número de pessoas no mundo vemos um novo modelo social, formando redes de pessoas que tem um mesmo interesse, as chamadas redes sociais. Logo as redes sociais são definidas em Marteleto (2001) como um conjunto de indivíduos

independentes, unidos por ideias e recursos que estão em torno de valores que os mesmos partilham.

Como o surgimento de novas tecnologias, como a grande rede mundial de computadores, as redes sociais tem uma nova face, no qual as pessoas utilizando um computador pessoal, o *Smartphone* ou qualquer outro tipo de dispositivo com acesso a internet, para compartilhar suas ideias. Os softwares de computadores que vem fazendo o papel de facilitar a comunicação entre as pessoas, o Facebook é um deles, mas também temos Twitter, Instagram, Snapchat, Google+, Myspace entre outros softwares que são caracterizados como do tipo rede social virtual.

O twitter é umas das redes sociais mais utilizadas atualmente, e [Recuero e Zago \(2016\)](#) o define como uma ferramenta de micro-mensagens que surgiu em 2006 com a finalidade de facilitar a interação entre os diversos tipos de usuário. Nele os usuário em um campo de texto respondem a seguinte pergunta: "O que está acontecendo ?", tem no máximo 140 caracteres para respondê-la, essa sentença escrita quando enviada é chamada de "*tweet*", como vemos na figura 5.



Figura 5 – Exemplo de *tweet* publicado pelo usuário @Cavaliinho do Fla_

Dentro do Twitter é possível seguir pessoas e ser seguido por elas, como exemplo, você pode seguir seu jogador favorito e acompanhar o que ele está "*twittando*", ou melhor, o que ele publica em seu perfil na rede social. Se o usuário desejar se comunicar diretamente com um outro usuário deve-se utilizar o "@" *nickname* do destinatário, em seguida a frase que deseja enviar ([RECUERO; ZAGO, 2016](#)).

O Twitter classifica os termos que estão sendo "*twittados*" no momento, todos os que tem "#" seguido de um termo, como "#motogpnoportv" é um assunto na rede, dessa forma também podemos buscar qualquer tipo de assunto utilizando o carácter "#" ([RECUERO; ZAGO, 2016](#)). O temas comentados pelos usuários da rede são diversificados, bem como sobre corrupção, a partida de futebol do domingo, notícias do mundo, saúde pública e assim por diante.

2.4 Análise de Sentimento em Comentários do Twitter

Com o fato de grande parte da sociedade estar expressando opiniões na internet e em redes sociais virtuais atualmente, a capacidade de gerar informação só tende a au-

mentar. Assim são gerados muitos documentos dentro da web, como os que expressam todo o tipo de opinião ou sentimento das pessoas. Existem diversos textos com conteúdo sobre opinião política, discussão sobre preço de produtos, divergência entre pessoas, descrevendo acontecimentos do mundo e outros. As redes sociais como Twitter e Facebook são alguns dos locais onde estão localizados geralmente esse tipo de documento (GOMES, 2013).

Nessa perspectiva surgem alguns trabalhos como é o caso deste, que objetivam analisar e classificar o sentimento contido em textos, isto é, fazer a análise de sentimento em textos. Que pode ser definida segundo Silva (2016a) como uma área interdisciplinar que faz a intersecção entre a mineração de dados, linguística, aprendizado de máquina, processamento de linguagem natural e análise textual. Cujo a finalidade é investigar a emoção expressa dentro de um pedaço de texto, como um *tweet* por exemplo. Na análise dos resultados podem conter sentimentos positivos ou negativos, porém também pode-se considerar o sentimento nulo, no qual o texto não expressa nenhuma emoção.

O Twitter é um excelente local para a extração de dados com a finalidade de analisar sentimento, pois os usuários dentro dessa rede social compartilham todo o tipo de informação pessoal, como o que estão pensando ou fazendo durante o dia. Todavia, na análise de textos do Twitter temos algumas dificuldades, abordadas em Silva (2016a), que são: (i) O texto muito curto, (ii) Variação da ortografia de usuário para usuário, (iii) Dados ruidosos devido a variação ortográfica, (iv) Detectar negação nas sentenças, (v) *Strop Words* seguindo uma linguagem informal, (vi) Entender símbolos especiais contidos na rede, (vii) Enorme variação dos tópicos, (viii) Alta quantidade de dados postados diariamente, (ix) Contém multilinguagens, (x) A *tokenização* sem padrão.

Contudo, na literatura como vemos em Medhat, Hassan e Korashy (2014), cita duas abordagens principais para análise de sentimento : (i) a léxica e (ii) por aprendizado de máquina, como mostra a Figura 6. Essas abordagens podem ser descritas como:

- **Abordagem por aprendizado de máquina:** Esse método utiliza de técnicas de aprendizado de máquina, sendo elas: supervisionada, que necessita de um conjunto de treinamento, a mais comum é a classificação; e também temos técnicas não supervisionadas, que não necessita do conjunto de treinamento, a mais comum é o agrupamento, que cria junções segundo características semelhantes do texto. Dentre as técnicas de aprendizado de máquina temos: o classificador Neive Bayes, Árvore de Decisão, Redes Neurais e outros (CHAMANSINGH; HOSEIN, 2016; Asiaee T. et al., 2012).
- **Abordagem Léxica:** Essa abordagem trabalha com lista de palavras, que são divididas no fase de pré processamento, a partir dessas listas utilizam de dois principais métodos: (i) baseado em dicionário e (ii) baseado em corpus. O método baseado em

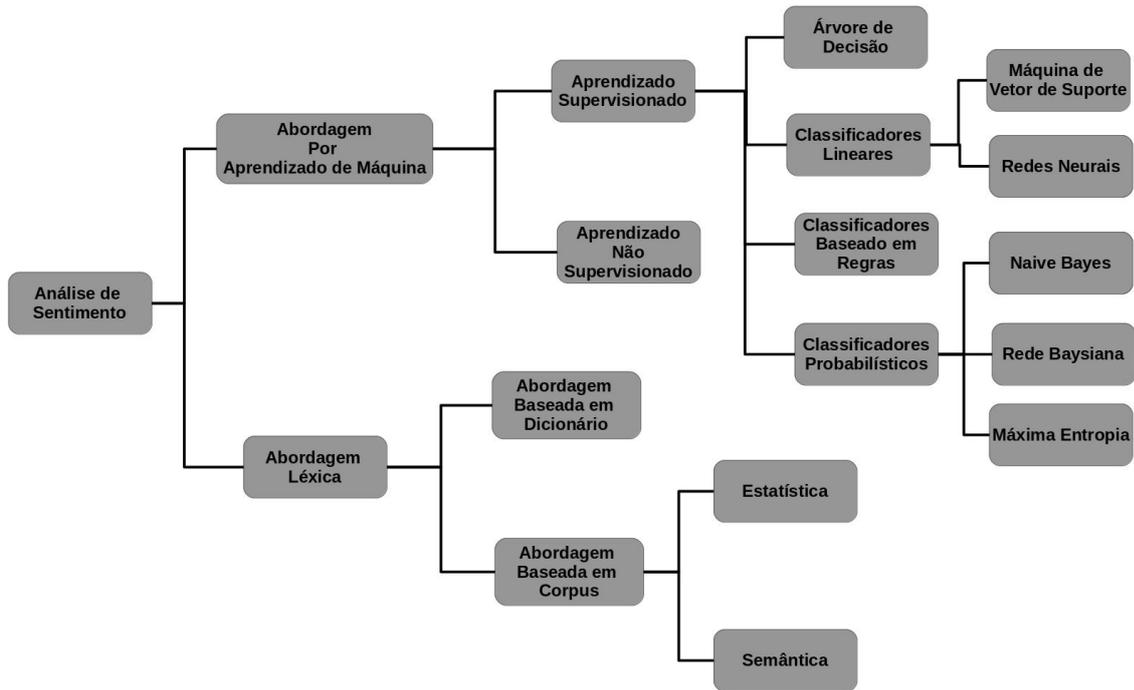


Figura 6 – Técnicas de Análise de Sentimento. Adaptado de (MEDHAT; HASSAN; KORASHY, 2014).

dicionários, utiliza de palavras geralmente separadas manualmente, que são as palavras de opinião, no qual são rotulas com uma classe, valor e outras características; com essas palavras é criada um dicionário onde elas são ligadas a seus sinônimos e antônimos, para ter um dicionário mais amplo. O método baseado em corpus, cria uma lista de palavras chamadas sementes, no qual tem como base para identificação da opinião os padrões semânticos das palavras, esse método encontra opinião em um contexto específico da palavras, bem como sendo um verbo, adjetivo e outras classes de palavras. A título de exemplo, podemos considerar uma ferramentas que utilizam esse tipo de abordagem, conhecida como WordNet (MEDHAT; HASSAN; KORASHY, 2014).

3 TRABALHOS RELACIONADOS

Na literatura existem vários trabalhos envolvendo análise de sentimento em texto, alguns utilizam as abordagens que empregam métodos léxicos, outros o aprendizado de máquina. Portanto, este capítulo tem o objetivo de apresentar alguns artigos relacionados com o assunto. Primeiramente serão discutidos os trabalhos envolvendo análise de sentimento em texto de modo geral. Em seguida são abordados temas relacionados mais especificadamente com este trabalho, tratando da análise de sentimento nos comentários do Twitter.

O trabalho de Santos, Becker e Moreira (2014), aborda a análise de sentimento em textos multilíngues (não tem uma linguagem específica), com a finalidade de avaliar produtos. O mesmo propõe a utilização de 8 emoções básicas: (i) alegria, (ii) tristeza, (iii) surpresa, (iv) medo, (v) raiva, (vi) repugnância, (vii) confiança e (viii) expectativa, que estão disponíveis no dicionário criado pelo *National Research Council* (NRC). A grande maioria dos dicionários linguísticos que esse trabalho pretende utilizar estão em inglês, assim os autores tem o desafio de traduzir as palavras e avaliar o quanto o mesmo é preciso. Por ser uma abordagem léxica o mesmo necessita de um corpus, no qual o utilizado foi o de revisões feitas em livros na loja virtual da Amazon, para se formar um outro conjunto de palavras para testes. Na execução dos experimentos os autores empregaram duas bibliotecas do Python: a *goslate* para a tradução dos textos utilizando o Google Tradutor e o *Natural Language Toolkit* (NLTK) para o processamento de linguagem natural. No final do trabalho os dados de treinamentos correspondente as revisões de livros obtiveram uma maior precisão do que os dados dos dicionários padrões.

Gonçalves et al. (2012) debate uma abordagem para a análise de sentimento em textos do Twitter, utilizando um método léxico, vindo do PANAS-x (Positive Affect Negative Affect Scale) chamando PANAS-t. O trabalho tem a finalidade de verificar o humor dos usuários do Twitter relacionados à assuntos, como política, desastres no mundo, saúde, como também eventos esportivos. Para a realização dessa tarefa foram coletados cerca de 1,8 bilhões de *tweets* de 2006, data da criação da rede social, até agosto de 2009. Como primeiro passo os autores aplicaram a limpeza do texto, que consistiu em aplicar a radicalização, remover as stopwords, caracteres indesejados e URLs. Em seguida aplicou-se a *tokenização* do texto utilizando o espaço como separador dos *tokens*. Os autores citam alguns sentimentos que podem ser classificados pelas técnicas, como : (i) medo, (ii) tristeza, (iii) culpa, (iv) honestidade, (v) timidez, (vi) fadiga, (vii) surpresa, (viii) jovialidade, (ix) auto-confiança, (x) atenção e (xi) serenidade. Para avaliar a precisão do PANAS-t foram selecionados alguns eventos que tiveram destaque mundialmente, como H1N1, a queda do avião da AirFrance, bem como, o Obama. Com isso, os testes feitos com a técnica visando

o H1N1 tiveram como resultado os sentimentos de atenção e medo, já para o avião da AirFrance que caiu em 2009 foram: medo e tristeza e assim por diante. O Twitter como qualquer outra rede social, tal como o Facebook, é um local onde as pessoas expõem suas opiniões sobre determinado assunto. Portanto, os dados contidos nelas são de grande auxílio para a pesquisa científica.

Sobre a classificação de sentimento utilizando abordagens de aprendizado de máquina vale a pena citar trabalhos como (AL-AMRANI; LAZAAR; ELKADIRI, 2017; PANG; LEE; VAITHYANATHAN, 2002; XIA; ZONG; LI, 2011). Nesse mesmo contexto temos Al-Amrani, Lazaar e Elkadiri (2017), que utiliza dos respectivos algoritmos para classificação do sentimento nos textos : (i) PART, (ii) Máquina de Vetor de Suporte (SVM), (iii) Árvore de Decisão, (iv) Naive Bayes e (v) Regressão Logística. Assim, o mesmo tem o objetivo de comparar esses algoritmos, mostrando a eficiência de cada um sobre um conjunto de dados de 5574 SMSs rotulados com duas classes: (i) positiva e (ii) negativa. Porém, não utilizam a base total para a execução do trabalho, só foram utilizados 200 itens de dado, no qual são 167 positivos e 33 negativos. Como resultado final dos experimentos os autores chegaram conclusão que o algoritmo da regressão logística foi o melhor entre os apresentados.

A classificação de sentimento é um conceito que vem ganhando espaço no meio científico, logo, surgem trabalhos que têm como propósito classificar sentimento em redes sociais como o twitter (MUKHERJEE et al., 2012; WANG et al., 2017; BRAVO-MARQUEZ; FRANK; PFAHRINGER, 2015). Bouazizi e Ohtsuki (2016) denotam a classificação de sentimento em *tweets*, focando em rotular três classes: (i) positiva, (ii) negativa e (ii) neutra. O trabalho obtém um conjunto de dados de 40000 *tweets*, que foram divididos em conjunto de treinamento e teste para fornecer ao algoritmo de aprendizado de máquina. Na execução dos experimento os autores ressaltam a utilização do algoritmo *Random Forest* para a classificação. O processo de avaliação do algoritmo proposto no trabalho, é feito com base em algumas métricas, como: Taxa de verdadeiro positivo, Taxa de falso positivo, Precisão, Sensibilidade e *F-Measure*, no qual obtiveram na grande maioria das métricas um valor de 81%.

Explorando mais especificadamente sobre a classificação de sentimento para a língua portuguesa, temos alguns trabalhos como em (ROSA; RODRIGUEZ; BRESSAN, 2013; AVANÇO; NUNES, 2014; SOUZA et al., 2016a). Em Grandin e Adan (2016) por exemplo, um dos trabalhos que aborda o idioma português, tem o objetivo de criar um sistema que analisa o sentimento dos *tweets* que contenha qualquer tópico de interesse. Para criar esse sistema os autores optaram por utilizar a tecnologia Ruby on Rails e JavaScript para implementar um sistema web, no qual o mesmo deve ter boa usabilidade e deve classificar os sentimentos com precisão. O sistema Piegas, como foi chamado pelos autores, utiliza do algoritmo Naive Bayes para a classificação em uma base com 19000 *tweets*.

O princípio chave para a classificação de sentimento, foi a extração de *tweets* contendo *emoticons* que foram classificados como positivos e negativos, e assim puderam obter um resultado satisfatório para o classificador.

Silva (2016a) apresenta em seu trabalho, um método que utiliza a técnica de combinação de classificadores para a análise de sentimento em texto das redes sociais. O autor foca no Twitter, utilizando de algumas bases de tweets já rotuladas em inglês, dos quais foram: (i) Sanders, Stanford Twitter Corpus, Obama-McCain Debate e *Health Care Reform*. Os algoritmos empregados na combinação de classificadores foram: Naive Bayes Multinomial, SVM, Random Forest e Regressão Logística, por serem bem citados na literatura para esse tipo de aplicação. Por fim foi feita a análise de desempenho da técnica proposta, que por sua vez mostrou alguns resultados interessantes, com valor máximo de acurácia sendo 84.89 % para a base Sanders.

Em contraste com os trabalhos apresentados anteriormente que na maioria são para a língua inglesa, o presente trabalho tem foco na análise de sentimentos para a língua portuguesa em texto da rede social Twitter. Para realizar esse processo será utilizado da abordagem de combinação de classificadores, empregando os algoritmos de aprendizado de máquina mais populares para rotular texto. Assim pretende-se obter um método de análise de sentimentos com uma confiança maior em suas predições.

4 ESTRUTURAÇÃO DA PESQUISA

Esta seção compreende apresentar os materiais e métodos utilizados durante a vigência desta pesquisa e também descrever todos os passos que foram seguidos para obter sucesso na execução dos objetivos propostos. A Figura 7 mostra a estrutura sistemática da pesquisa.

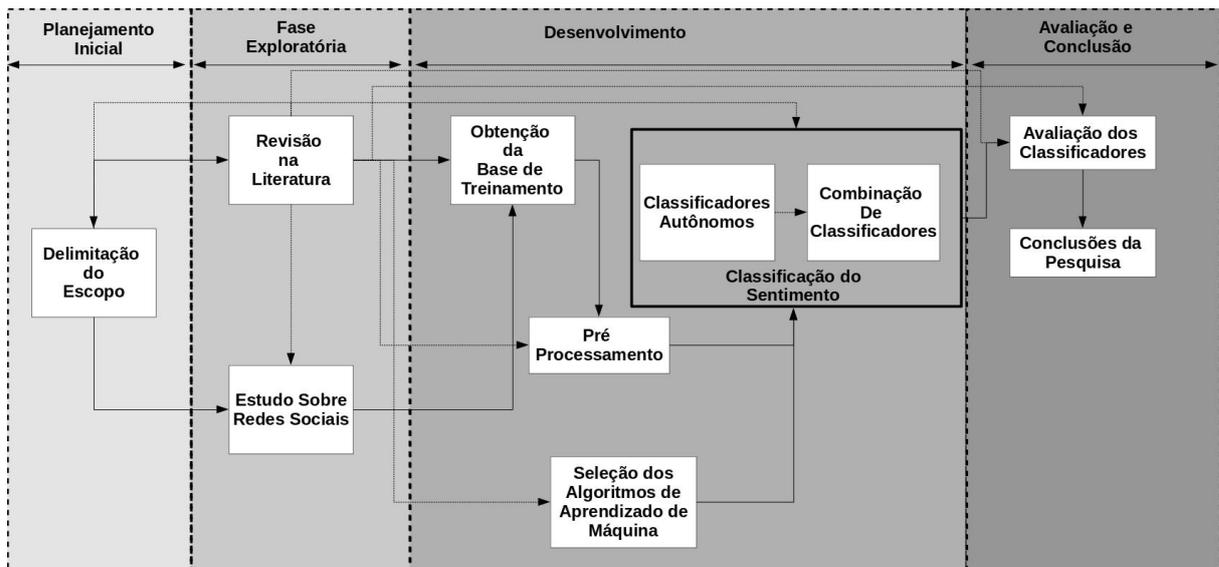


Figura 7 – Estrutura da Pesquisa

4.1 Planejamento Inicial

Esta é fase que vem logo após a escolha do tema do projeto, no qual foi definido algumas indagações introdutórias, como : **Qual a rede social mais adequada para o tema análise de sentimento ? - Quais os algoritmos mais utilizados na literatura para classificar texto ? - Quais das duas abordagens para classificação de sentimento serão mais adequadas para este escopo, a léxica ou por aprendizado de máquina ?**.

Portanto, as questões levantadas no decorrer do trabalho foram importantes para determinar o escopo da pesquisa. Assim o trabalho não divergiu para pontos irrelevantes, ou seja, que poderiam consumir um tempo da pesquisa em assuntos que por fim não seriam utilizados.

4.2 Fase Exploratória

Revisão na Literatura

A revisão bibliográfica tem o objetivo de proporcionar ao pesquisador o conhecimento mais detalhado do campo do trabalho, bem como apresentar as pesquisas correlatas ao tema que estão sendo feitas, como foi visto na seção anterior. Também possibilita o conhecimento sobre alguns assuntos que não fazem parte direta da área de enfoque do pesquisador, principalmente da computação, que é aplicada em outras áreas. Esses conceitos foram apresentados na seção 2.

Estudo nas Redes Sociais

Atualmente temos redes sociais com diferentes características, mas também algumas semelhanças. A partir disso foi necessário um estudo em cima dessas redes sociais, com a finalidade de avaliar e explorar qual delas tem um maior custo benefício para a aplicação desta pesquisa.

O estudo foi focada em duas delas: (i) Facebook e (ii) Twitter, por ter uma característica serem bem populares e os usuários publicarem muitas mensagens de texto. Os requisitos analisados nas mesmas foram:

1. Primeiramente foi analisado **quais informações as mesmas dariam acesso**. Nesse requisito o Twitter se saiu melhor, já que disponibiliza os *tweets* de todos os usuários, dados do perfil, região onde foi *twittado* um assunto e outros. Já o Facebook é bem limitado, pois concede acesso somente a comentários de pessoas ou páginas que tem relação direta com a conta vinculada.
2. Logo após, foi explorado **qual delas é mais fácil de obter conteúdo**. O Twitter conta com um conjunto de *Application Programming Interface* (APIs), como a *Representational State Transfer* (REST) API que faz a extração dos *tweets* a partir de *hashtags* definidas. Também temos a *Streaming* API, que extrai dados em tempo real, diferente da REST API captura dados históricos. O Facebook conta com a *Graph* API, que disponibiliza capturar comentários a partir de termos adicionados na busca. Nesse requisito o Facebook se saiu melhor pois, fornece dados históricos de longa data. Logo, o Twitter tem um problema na REST API, segundo [Twitter \(2017\)](#) só é possível capturar dados de até uma semana atrás.

Afinal, na problemática deste trabalho a variável que tem maior peso é a grande quantidade de *tweets*. Com isso se conclui que a melhor rede social para o problema é o Twitter, por facilitar o acesso aos dados com suas APIs e grande parte do seus comentários serem texto puro. Um outro aspecto que deve-se levar em conta para a escolha do twitter, é que temos disponível na web algumas bases de *tweets* que podem ser utilizadas como treinamento dos algoritmos.

4.3 Desenvolvimento

Obtenção da Base de Treinamento

Com o objetivo de ensinar os algoritmos de aprendizado de máquina a reconhecer o sentimento contido dentro dos textos, foi necessário obter bases de *tweets* já rotuladas para o treinamento deles. Assim, foi realizada uma busca sistemática na literatura e em sites específicos, com a finalidade de encontrar conjuntos de dados abertos para o acesso e utilização neste trabalho.

Primeiramente a busca foi realizada visando conjuntos de dados para português do Brasil, logo foi encontrada uma base publicada pelo grupo de pesquisa MiningBR ¹, que rotulou essa base manualmente, identificando o sentimento no texto e o classificando apenas como negativo (-1), Neutro (0) ou Positivo (1). O grupo utilizou esses dados para as elaboração de alguns trabalhos, bem como (SOUZA et al., 2016c; SOUZA et al., 2016d; SOUZA et al., 2016b). No site do grupo pode-se encontrar a base no divida em dois grupos, contudo neste trabalho foi feita a junção do mesmo. A fim de expandir a base e deixá-la com número maior de registros, assim tudo foi gravado em um arquivo do tipo Comma-separated values (CSV).

Seleção dos Algoritmos de Aprendizado de Máquina

Na literatura relacionada ao aprendizado de máquina temos vários algoritmos de classificação, por essa vasta gama de algoritmos surge a dúvida de qual utilizar em cada aplicação. Assim esta seção teve o propósito de realizar uma exploração na literatura para encontrar quais os métodos mais utilizados para classificação de texto, a fim de determinar quais deles seriam utilizados neste trabalho. Portanto a partir dos trabalhos de Silva (2016a), Grandin e Adan (2016) e Augustyniak et al. (2014), concluiu-se que os classificadores utilizados são:

- **Naive Bayes:** Esse algoritmo foi selecionado neste trabalho por ser bem popular para a problemática abordada, ou seja, para a classificação de texto. Os autores que trabalharam com o mesmo, como em Souza et al. (2016a) conseguiram resultados interessantes. Assim aplicou-se o algoritmo Naive Bayes Multinomial neste trabalho, que tem o objetivo de prever a classe que um dado pertence, a partir da probabilidade vinda da fórmula.
- **Máquina de Vetor de Suporte:** Optou-se da utilização do SVM neste trabalho, por ser segundo Faceli et al. (2011) um dos melhores classificadores que temos na literatura, várias pesquisas vem utilizando do mesmo para a classificação de textos como em (Asiaee T. et al., 2012). Uma observação que deve ser levada em conta

¹ <<https://sites.google.com/site/miningbrgroup/home/publications>>

é que o algoritmo SVM é binário, assim é definido a estratégia um contra todos (OVR) para tratar essa questão.

- **Árvore de Decisão:** O algoritmo foi empregado nesse trabalho por ser bem popular para classificação e também é válido para abordagens que exploram textos, como [Rihab, Elkhelifi e FaBouchlahem Rihabiz \(2016\)](#) utiliza em seu trabalho.
- **Random Forest:** Este trabalho utilizou desse algoritmo com a finalidade de comparar um algoritmo mais otimizado de árvores de decisão, que conceituados na literatura. A abordagem utilizando Random Forest também é válida para texto como [Bouazizi e Ohtsuki \(2016\)](#) faz em seu trabalho.
- **Regressão Logística:** Esse método foi utilizado neste trabalho, pois segundo [Silva \(2016a\)](#) é um dos mais utilizados na literatura para classificação de texto, é utilizado em vários trabalhos como em ([MORENO, 2015](#); [MITTAL](#); [GOEL, 2012](#); [SILVA](#); [HRUSCHKA](#); [HRUSCHKA, 2014](#)). Este algoritmo também utiliza da estratégia OVR para trabalhar com multiclases.

Em algoritmos de aprendizado de máquina geralmente os valores dos parâmetros tem que ser especificado pelo usuário. Isso é muito ruim para predição, pois a escolha se baseia em testes empíricos em até mesmo na aleatoriedade, assim ocorrendo um erro generalizado maior e uma predição incorreta. Temos pesquisas na literatura que trabalham somente com objetivo de desenvolver algoritmos otimizados para buscar os melhores parâmetros, isto é, os que proporcionem um erro generalizado menor ([ROSSI, 2009](#)).

Logo, neste trabalho foi empregado uma dessas técnicas otimizadas para a busca de melhores parâmetros para os algoritmos apresentados anteriormente. Foi selecionado o *Grid Search*, que faz a busca exaustiva em cima de alguns parâmetros pré estabelecidos como espaço de busca para o mesmo ([FARIAS, 2016b](#)). O SKLEARN tem uma função que dá suporte ao *Grid Search*, que tem como entrada o algoritmo e o espaço de busca. Essa técnica foi selecionada por ser conhecida na literatura, contudo tem a desvantagem de ser custosa. Porém para a essa aplicação o tempo de processamento não foi o empecilho, assim teve alguns efeitos interessantes para alguns algoritmos.

Classificação de Sentimento

Para alcançar os objetivos propostos, foi necessário a implementação dos algoritmos passo a passo, até atingir o modelo completo, ou seja, que ainda está em desenvolvimentos. Para realizar o desenvolvimento do modelo foi utilizado a linguagem Python na versão 3.5, também foi empregado um *framework* de aprendizado de máquina, conhecido como Skit-learn (SKLEARN), que possui diversas implementações de métodos de aprendizado de máquina, tal como a normalização e transformação do texto. A Seção 5 apresenta de forma mais detalhada o funcionamento do método.

Classificadores Autônomos

Para atingir o objetivo deste trabalho foi proposto um método para análise de sentimento, que se baseia nas principais abordagens apresentadas para classificação de texto na literatura e tem como propósito classificar o sentimento para a língua portuguesa.

Inicialmente foi definido a utilização dos algoritmos de aprendizado de máquina de forma isolada, assim foi possível testar cada um deles com a base de dados e obter seus resultados separadamente. Esse processo é importante para determinar o desempenho individual de cada algoritmo, com a finalidade que em um próximo passo seja realizada a comparação de cada um deles com a junção desses classificadores. Assim poderá se obter os resultados para esse pesquisa.

Combinação de Classificadores

Essa é a fase que irá demonstrar o procedimento proposto e definir suas configurações, para que o comitê (combinação de classificadores), obtenha uma predição mais confiável.

Com os algoritmos de aprendizado de máquina selecionados com base na literatura, como foi apresentado anteriormente, foi possível desenvolver um método que combine os resultados dos algoritmos, utilizando de alguma estratégias para definição da predição final, afim de melhorar o desempenho do mesmo. Existem várias estratégias para definir o valor predito pela combinação dos classificadores como [Faceli et al. \(2011\)](#) cita em seu livro. Contudo neste trabalho optou-se em se utilizar o regime da votação, no qual a classe mais votada dentre os algoritmos é a que dará o rótulo aos exemplos.

Para realizar a análise de sentimento primeiramente buscou-se na literatura os métodos utilizados para realizar esse processo. Assim optou-se pela análise de sentimento utilizando técnicas de aprendizado de máquina, no qual muitas delas trabalham com sentimento negativo, positivo e outras até incluem o neutro. Outro ponto a se considerar é o a utilização base para a língua portuguesa, que conta com poucos trabalhos, se comparados com os que temos na língua inglesa, por exemplo na base de *papers* da ACM library, contém aproximadamente 114.052 trabalhos relacionados ao termo *sentiment analysis*.

4.4 Avaliação dos Classificadores

Esta é a última etapa da estrutura da pesquisa, no qual foi dividida em duas subseções que descrevem os métodos de avaliação do modelo proposto e a conclusão do trabalho que define o que foi alcançado de resultados com essa pesquisa.

Quando se trabalha com aprendizado de máquina o domínio que temos acesso é somente aquele incluído nos casos de testes, portanto é necessário verificar o desempenho dos modelos utilizados, assim como verificar a sua precisão com valores inesperados. Essa

avaliação pode apresentar a taxa de erro do método, sua precisão e várias outras métricas, assim é possível concluir se o método é válido para se utilizar (FACELI et al., 2011).

Assim, foi realizado uma pesquisa com a finalidade de explorar na literatura as medidas comumente utilizados para avaliação de modelos de preditivos em aprendizado de máquina. Logo, várias métricas foram encontradas como Santos (2013) apresenta em seu trabalho. Entretanto, após algumas análises, foi decidido utilizar as seguintes métricas de avaliação: (i) para amostragem, foi empregada a validação cruzada (*cross-validation*), (ii) para medidas de desempenho foram: a acurácia, o erro, precisão, sensibilidade, especificidade e o *f1 score*, (iii) para a avaliar os classificadores, foi utilizada da curva ROC.

E por fim é realizada a análise estatísticas do método proposto e dos algoritmos isolados, considerando três técnicas estatísticas: (i) Shapiro Wilk, (ii) método de Friedman e (iii) *Post-hoc* Friedman-Nemenyi. O primeiro teste utilizado foi o de Shapiro Wilk que indica a pressuposições de normalidade de uma amostra, assim a partir do p-valor podemos analisar se a hipótese é aceita ou recusada; se o p-valor for maior que 0.05 (5%) aprovada, caso contrário recusada (PEREIRA et al., 2010). Logo, se for aceito utilizaremos o teste T e se for recusado o teste de Friedman. O método de Friedman Friedman (1937) é basicamente um teste não paramétrico que também fornece um p-valor, no qual indica se os grupos de amostras tem diferença, isto é, se o p-valor for maior que 0.05 sem diferença e se for menor com diferença. Após utilizar o método de Friedman se houver diferença, temos que verificar as diferenças significativas pelo teste *post-hoc* de Friedman-Nemenyi (BENAVOLI; CORANI; MANGILI, 2016).

5 MÉTODO PROPOSTO

Este trabalho propõe um método para análise de sentimento em comentários do twitter, utilizando da combinação de algoritmos de aprendizado de máquina. A estrutura desse método pode ser observado na Figura 8.

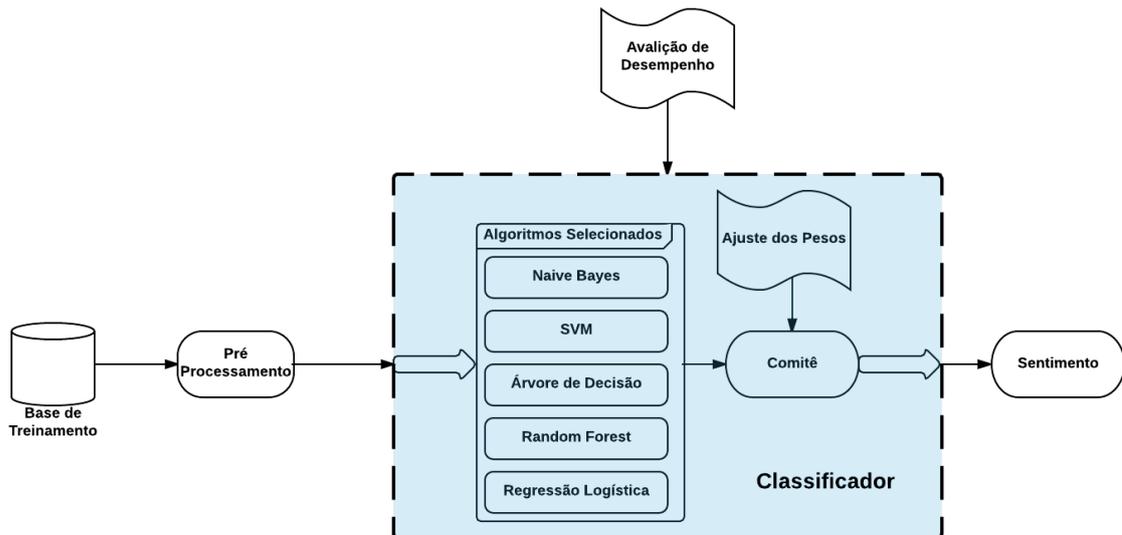


Figura 8 – Método Proposto

5.1 Pré Processamento

Normalização

Esta fase foi desenvolvida com uso de técnicas de PLN, com a finalidade de deixar o texto mais limpo para as próximas etapas. Para isso empregamos a biblioteca NLTK¹ do Python com a versão 3.5. Ela conta com ferramentas que realizam todas as fases do processamento de linguagem natural, mas também de algoritmos para a classificação do texto. Os recursos utilizados foram:

1. **Limpeza do texto com expressão regular:** Nesse passo foi retirado caracteres indesejados do texto, como *links*, que não são significativos nas etapas futuras. Foi utilizado as seguintes expressões regulares: $http |S+$ que localiza os links e substitui por vazio, $@ |S+$ que encontra os texto equivalente aos usuários e substitui por vazio e $[|S]+$ localiza os restante como alguns caracteres da codificação ASCII.

¹ <<http://www.nltk.org/>>

2. **Extração dos *tokens*:** Os *tokens* foram separados utilizando uma função no NLTK. Ele toma como base os espaços, assim a cada espaço ele inicia um novo *token* e o guarda na lista;
3. **Remoção das *stop words*:** Foi aplicada a função "*corpus.stopwords.words*" do NLTK e escolhida a língua portuguesa, desse modo foi retirada as *stop words* de cada *tweet*.

Transformação do Texto

Os algoritmos de aprendizado de máquina fornecidos pela biblioteca utilizada não entendem o texto puro, assim é necessário realizar a conversão dos valores textuais para um tipo numérico. Contudo, para efetuar essa tarefa foi empregada a técnica *Bag of words*, ou saco de palavras em português, a partir da biblioteca Scikit-Learn (SKLEARN) do Python, que conta com várias ferramentas do tipo. O processo de extração foi feito a partir de toda a base de *tweets* pré processada chama-se a biblioteca do SKLEARN, que transforma a coleção de textos em um vetor numérico a partir da ocorrência das palavras dentro do corpo textual (SILVA, 2016b).

5.2 Comitê

O comitê utiliza da abordagem de combinação de classificadores, mais especificamente o método da votação que tem como finalidade maximizar a precisão da predição, ou seja, esse método tem como propósito ser melhor do que uma escolha aleatória de algoritmo (FACELI et al., 2011). A técnica de combinação de classificadores é utilizado na literatura de duas formas, concedendo pesos maiores para alguns algoritmos ou deixando com os mesmos pesos para que todos tenham o poder do voto semelhante (BERNARDINI, 2002).

Portanto, este trabalho tem como próximo passo a utilização do comitê de classificadores de aprendizado de máquina. Os classificadores utilizados são: (i) Naive Bayes, (ii) Máquina de Vetor de Suporte, (iii) Árvore de Decisão, (iv) *Random Forest*, (v) Regressão Logística. Optou-se inicialmente testar os mesmo sem ponderar os algoritmos, porém será realizado um estudo para definir a melhor forma de ponderar os mesmos. Assim a predição do algoritmo com maior peso terá uma pontuação mais alta em seu voto.

5.3 Ajuste dos Pesos

Os algoritmos que são utilizados no comitê tem valores diferentes de acurácia, assim é necessário analisar a força do voto de cada um, para que os melhores tenham uma força maior, a fim de conseguir uma melhor precisão no comitê. Logo, para definir os pesos, foi criado um regime, no qual é governado pela seguinte equação:

$$peso_i = \frac{ac_i}{\sum_{j=0}^n ac_j} \quad (5.1)$$

Onde:

- ac é a acurácia do algoritmo;
- i e j são os índices dos algoritmos.

5.4 Avaliação do Desempenho

Validação Cruzada

Com a finalidade de dividir a base total, em duas, uma de treinamento e outra de teste, é comumente utilizado na literatura validação cruzada com o K-Fold ([BESPALOV et al., 2011](#)). Este método decompõe a base em treino e teste aleatoriamente, com a possibilidade de sobreposição, refazendo esse processo K vezes. Neste trabalho utilizamos o $K = 10$, gerando 10 grupos de teste, que foram executados de uma forma lógica e sequencial a partir do a implementação utilizada que vem agregada a biblioteca SKLEARN. Assim são iniciadas as predições seguindo esse processo de divisão, no qual é interessante inserir algumas métricas para verificar o desempenho do algoritmo, como esses valores gerados. Para cada valor de K, é realizado o cálculo da média aritmética para calcular o desempenho final do algoritmo ([FACELI et al., 2011](#)).

Métricas de Desempenho

As métricas utilizadas para mensurar o desempenho do algoritmos foram basicamente: (i) acurácia, (ii) precisão, (iii) sensibilidade (recall), (iv) *F1 score* e o (v) erro. [Faceli et al. \(2011\)](#) cita as principais métricas em seu livro, mas somente para 2 classes, contudo nesta problemática foram utilizadas três classes. Dessa forma, muda em partes a matriz de confusão e alguns fragmentos dos cálculos, como em [Nascimento e Cruz \(2013\)](#) que trabalha com muticlasses. A matriz de confusão, para três classes como foi apresentado na Figura 4 da Seção 2.

5.4.1 Sentimento

Em análise de sentimentos temos algumas possibilidades de polaridade a serem apresentadas, na literatura algumas autores apresentam as classes somente como positivo e negativo ([FARIAS, 2016a](#)), outros demonstram as classes positivo negativo e neutro como em [Silva \(2016b\)](#). Esse são as formas mais primitivas de classificar sentimento em texto, contudo esses método podem evoluir ainda mais, classificando emoções, como medo, culpa, tristeza e outras ([GONÇALVES et al., 2012](#)).

Assim, o método proposto tem como objetivo classificar o sentimento multiclass, no qual tem como resultado três rótulos: positivo(1), neutro(0) e negativo(-1). Pois a base de treinamento utilizada neste trabalho aborda essas três classes, a partir dessa perspectiva optou-se por utilizar os mesmos rótulos no método proposto também.

6 RESULTADOS

Esta seção tem como propósito apresentar os experimentos realizados, bem como os resultados obtidos no decorrer deste trabalho.

6.1 Análise da Abordagem Proposta

Como vimos anteriormente algumas métricas foram propostas, a fim de avaliar o desempenho dos algoritmos utilizados, tal como a abordagem proposta utilizando comitê. A Tabela 1 apresenta os resultados das métricas utilizadas para os algoritmos isolados e também para o comitê. Esses valores foram obtidos a partir da execução do *script* que contém a implementação dos algoritmos de classificação juntamente com a utilização do *Grid Search*.

Tabela 1 – Resultados das Métricas de Desempenho dos Algoritmos Utilizados

Algoritmos	Acurácia	Precisão	Recall	F1 Score	Erro
Naive Bayes	0.8107	0.8145	0.8107	0.8126	0.2729
SVM	0.8412	0.8467	0.8412	0.8440	0.2127
Árvore de Decisão	0.8007	0.8274	0.8007	0.8158	0.2837
Random Forest	0.8568	0.8623	0.8568	0.8599	0.2067
Reg. Logística	0.8429	0.8450	0.8429	0.8429	0.2063
Comitê	0.8685	0.8694	0.8685	0.8689	0.1809

A tabela acima apresenta algumas pontuações dos algoritmos, isto é, as métricas de desempenho que foram descritas na Seção 2, como pode ser observado o que tem as melhores pontuações é o comitê, com uma taxa de acerto de 86 %, assim a partir da tabela podemos concluir que ele foi o melhor para a base de dados utilizada no trabalho, que contém 2516 registros para a língua portuguesa. Graficamente podemos visualizar a diferença das acurácias dos algoritmos pelo *boxplot* na Figura 9.

No entanto, o gráfico *Receiver Operating Characteristic* (ROC), como mostra a Figura 10, apresenta um comportamento diferente do que diz as acurácias, logo o classificador que gera melhores modelos de classificação para o *dataset* utilizando neste trabalho foi o Naive Bayes. Cujos tem um valor de *Area Under Curve* (AUC) igual a 0.73, no qual é melhor que o comitê que está praticamente no meio do ranking de melhores classificadores, com um valor de AUC igual a 0.68.

Todos os valores de métricas são obtidos a partir das matrizes de confusão geradas pelos algoritmos, a Figura 11 apresenta a matriz de confusão do algoritmo que obteve a melhor acurácia, dos outros algoritmos também foram geradas e podem ser observadas no Apêndice A.

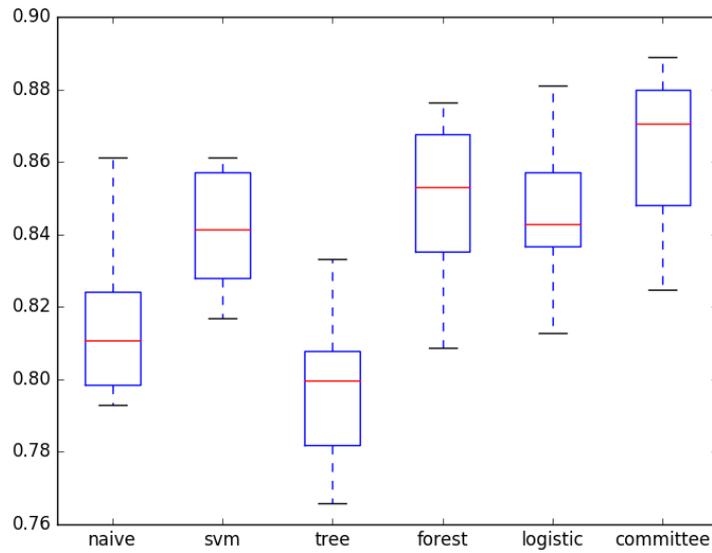


Figura 9 – Boxplot das Acurácias dos Algoritmos Utilizados

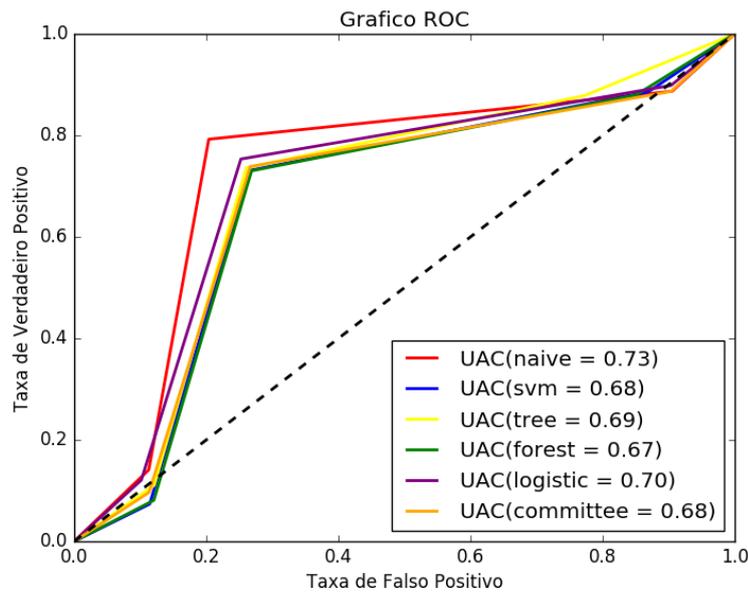


Figura 10 – Gráfico ROC

Análise Estatística

Com o propósito de validar os experimentos estatisticamente, foi realizado nessa etapa a análise dos resultados utilizando alguns testes, como a Figura 12 apresenta, seguindo as ideias de (MOZETIC; GRĀAR; SMAILOVIĆ, 2016; DEMSAR, 2006; MALHOTRA, 2015).

Primeiramente foi aplicado o teste de Shapiro Wilk, que segundo Torman, Coster e Riboldi (2012) tem o propósito de verificar se uma amostra segue a distribuição normal, no qual uma forma de verificar sua veracidade, é por meio do p-valor. Essa fase será

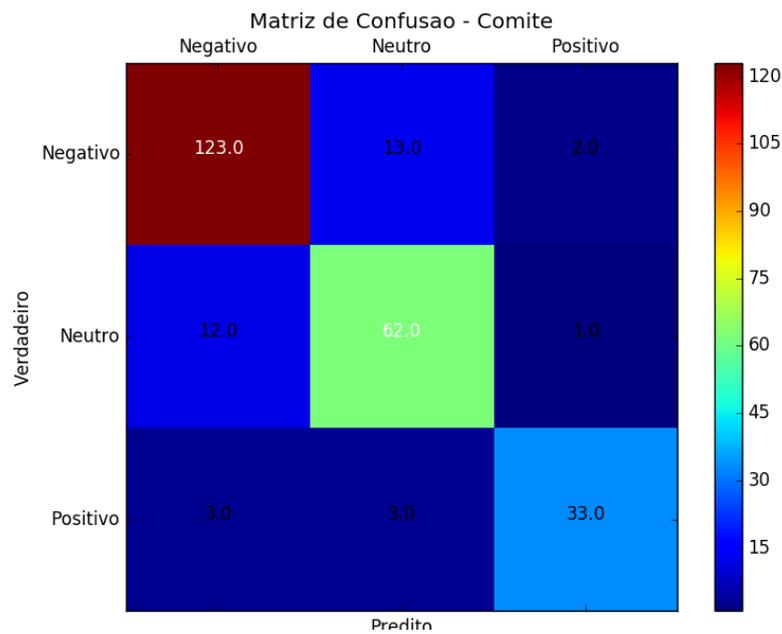


Figura 11 – Matriz de Confusão do Comitê

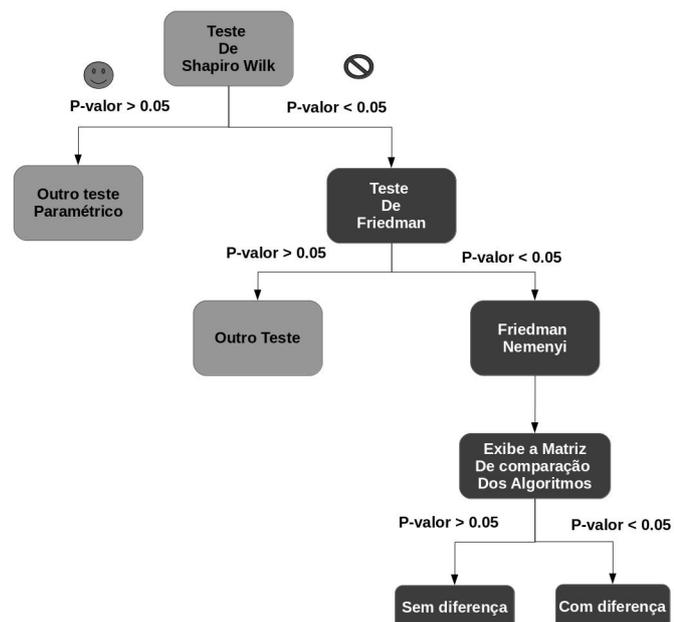


Figura 12 – Estruturação dos Testes

importante para avançar para o próximo passo, assim fazem a análise a verificação de duas hipóteses (LOPES; BRANCO; SOARES, 2013):

- H_0 (Hipótese nula): A amostra segue a distribuição normal. Se o p-valor ≥ 0.05 ;
- H_1 : A amostra não segue a distribuição normal. Se o p-valor ≤ 0.05 .

Após verificar a normalidade das amostras pelo teste de Shapiro Wilk, a hipótese H_0 foi negada, aderindo a hipótese H_1 . Assim segundo [Malhotra \(2015\)](#) é necessário utilizar um teste não paramétrico, no qual o mais adequado para essa perspectiva foi o teste de Friedman, pois o mesmo trabalha com mais de dois conjuntos de exemplos e neste trabalho é abordado seis conjunto de exemplos.

O teste de Friedman ([FRIEDMAN, 1937](#)) é um tipo de teste de hipótese não paramétrico adequado para mais de dois grupos de amostras, no qual os grupos devem ser ordenados seguindo um *ranking* de desempenho, por exemplo, o algoritmo com melhor acurácia é o número 1 do ranking, o segundo o número 2 e assim por diante. As hipóteses desse teste são ([DEMSAR, 2006](#)):

- H_0 (Hipótese nula): Todos os algoritmos são equivalentes. Se o p-valor ≥ 0.05 ;
- H_1 : Os algoritmos tem diferença. Se o p-valor ≤ 0.05 .

Executando o teste de Friedman ordenando o algoritmos da seguinte forma: comitê, Random Forest, Regressão Logística, SVM, Naive Bayes e Árvore de Decisão, a partir dos valores de acurácia. Portanto obtivemos o p-valor igual a 0.015, em que nega a hipótese nula H_0 , logo segundo ([BENAVOLI; CORANI; MANGILI, 2016](#)) temos que usar um teste *post-hoc* para verificar quais as diferença significativas entre os algoritmos.

O teste *post-hoc* Nemenyi é o mais adequado pós teste de Friedman ([DEMSAR, 2006](#)). É um teste pareado para amostras independentes, no qual é realizado o teste de todos os algoritmos entre si, a fim de verificar se os mesmo tem diferença, em que se o p-valor ≥ 0.05 não tem diferença e se p-valor ≤ 0.05 tem diferença. A Tabela 2 apresenta os p-valores atingidos com esse teste.

Tabela 2 – Tabela Com os P-valores Gerados a Partir do Teste *Post-hoc* de Nemenyi Para o Método Proposto

	Comitê	R. Forest	R. Logística	SVM	N. Bayes
R. Forest	0.314				
R. Logística	0.783	0.012			
SVM	0.329	1	0.011		
N. Bayes	0.896	0.026	0.999	0.024	
A. Decisão	0.00013	0.14	1.4e-7	0.151	5.7e-7

Com os p-valores obtidos pode-se realizar a análise da hipótese deste trabalho, que consiste em o comitê tem diferença para todos os outros algoritmos. Assim iremos analisar somente a primeira coluna da Tabela 2. No entanto, a hipótese definida para esse trabalho não é uma verdade, pois a maioria dos algoritmos não teve diferença para o comitê, somente a árvore de decisão que teve diferença. Logo, isso quer dizer que os algoritmos são estatisticamente equivalentes para realizar predições para esta perspectiva.

Em suma podemos concluir que para conjuntos de dados pequenos como o utilizado nesse trabalho que tem 2516 registros, o comitê indicado e como vimos na Figura 9 obtém uma maior acurácia. No entanto se o conjunto de dados crescer significativamente, pode-se utilizar o algoritmo Naive Bayes, pois como vemos na Figura 10 o Naive Bayes gera melhores modelos de classificação para essa perspectiva, bem como tem um custo em relação ao tempo de execução menor comparado com os outros algoritmos que não tem diferença para o comitê, como apresenta a Tabela 3.

Tabela 3 – Tempo de Execução dos Algoritmos Utilizados

	N. Bayes	SVM	A. Decisão	R. Forest	R. Logística	Comitê
Tempo (s)	0.20	5.62	3.28	1.78	0.72	11.18

6.2 Análise de Técnicas Semelhantes

Essa seção tem como finalidade realizar um contraste entre o método proposto e uma abordagem recente relatada na literatura, bem como duas ferramentas utilizadas para classificar emoção em texto. Primeiramente será exibido o TSviz um método de análise de sentimentos que utiliza da abordagem léxica, logo após será apontadas duas ferramentas das quais são: (i) o IBM Watson¹ que tem um módulo de processamento de linguagem natural e (ii) o módulo de análise de sentimentos dos serviços cognitivos da Microsoft,².

TSviz

O TSViz (RIOS et al., 2017) foi apresentado no último Brazilian Conference on Intelligent Systems (BRACIS, 2017) essa abordagem também tem um *dashboard*³ localizado em uma página da web, no qual podemos fazer o uso para visualizar o sentimento dos usuários do Twitter a partir de determinadas *hashtags* no decorrer do tempo.

O Tsviz funciona utilizando de um Corpus, por ser uma abordagem léxica, o Corpus empregado é chamado Reli, em que contém várias palavras extraídas de resenhas de livros clássicos em português. Para calcular o sentimento primeiramente cada palavras da frase é transformada em um vetor de características utilizando o conceito de Word2vec⁴, ou seja é transformado em um vetor numérico. Assim podemos calcular a similaridades das palavras com as palavras do Corpus para verificar a qual classe ela pertence, isto é, se a palavra mais similar estiver na classes positiva então essa palavra recebe o rótulo positivo dentro da frase, logo é feito isso para todas as palavras até terminar a frase. Após isso

¹ <<https://www.ibm.com/watson/services/natural-language-understanding/>>

² <<https://azure.microsoft.com/pt-br/services/cognitive-services/text-analytics/>>

³ <<https://dashboard.tsviz.com.br/dashboard/producer/15>>

⁴ <<https://code.google.com/archive/p/word2vec/>>

utilizamos as Equações 6.1 e 6.2 que soma quantas palavras de cada classe tem dentro da frase (F).

$$P = \sum F_{positive}(w) \quad (6.1)$$

$$N = \sum F_{negative}(w) \quad (6.2)$$

As Equações 6.3 e 6.4 realizaram uma normalização dos valores, obtendo NP e NN, no qual serão subtraídos NP - NN que resultará no valor do sentimento, que poderá ser -1,0 ou 1.

$$NP = \frac{NP}{N + P} \quad (6.3)$$

$$NN = \frac{N}{N + P} \quad (6.4)$$

O desempenho obtido por essa técnica, utilizando de métricas de desempenho que são apresentados na Tabela 4. Cujo tem uma taxa de acerto igual a 43 %, se saindo pior que a abordagem proposta neste trabalho. Para obter as métricas foi utiliza da matriz de confusão gerada pelo método, como mostra a Figura 13.

Tabela 4 – Métricas de Desempenho do TSviz

	Acurácia	Precisão	Recall	F1 Score	Erro
TSviz	0.43	0.49	0.43	0.46	0.71

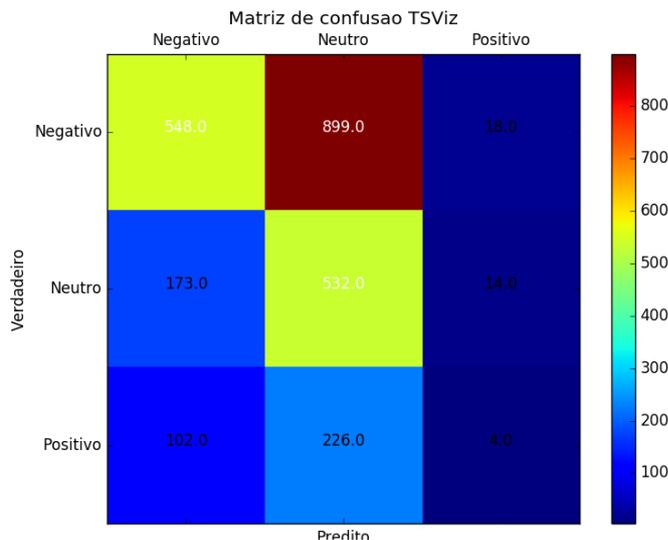


Figura 13 – Matriz de Confusão do TSviz

Como foi feito no método proposto para provar estatisticamente se o mesmo tem ou não diferença entre si, neste tópico é exibido o mesmo processo da Figura 12 para o TSviz. Primeiramente é empregado o teste de Shapiro Wilk, em que obtivemos um p-valor que recusou a hipótese nula H_0 , então segue-se para o teste de Friedman, no qual teve um p-valor igual 0.018, logo teremos que utilizar o *post-hoc* Nemenyi para verificar as diferenças significativas entres os métodos. Pelo teste *post-hoc* Nemenyi alcançamos a Tabela 5, no qual pode-se concluir que o TSviz tem diferença estatística para o Comitê. Portanto podemos concluir que o Comitê por ter uma maior taxa de acerto, pode ser mais adequado nas condições que foi utilizado, para classificar sentimentos em comentários do Twitter do que o TSviz.

Tabela 5 – Tabela Com os P-valores Gerados a Partir do Teste *Post-hoc* de Nemenyi Para o TSviz

	Comitê	R. Forest	R. Logística	SVM	N. Bayes	A. Decisão
R. Forest	0.379					
R. Logística	0.832	0.012				
SVM	0.363	1	0.011			
N. Bayes	0.944	0.032	0.999	0.027		
A. Decisão	0.001	0.156	7.4e-8	0.165	4.4e-7	
TSviz	7.1e-14	5.9e-14	2e-16	6e-14	2e-16	1.5e-8

IBM Watson e Microsoft *Text Analytics*

O IBM Watson e o Microsoft *Text Analytics* são métodos mais voltados a aplicações comerciais, pois pertencem a empresas privadas, contudo ambas disponibilizam um tempo de teste, em que seus usuários tem liberdade para utilizar da ferramenta, porém com algumas limitações. Esse métodos são um boa aquisição é para auxiliar no processo estratégico da empresas.

O IBM Watson é composto por várias APIs integradas, bem como a *natural language understanding* (NLU), que contém uma *feature* para análise de sentimentos, no qual o usuário passa um texto, o mesmo identifica a linguagem que o texto pertence e o sentimento é analisado pela API na linguagem identificada, atribuindo os rótulos positivo, negativo ou neutro. Em relação ao desempenho podemos ver na Tabela 6 que a acurácia do Watson foi de 42 %, sendo pior que o nosso método proposto. Vale a pena ressaltar que todos esses valores foram calculados a partir da matriz de confusão do método, como vemos na Figura 14.

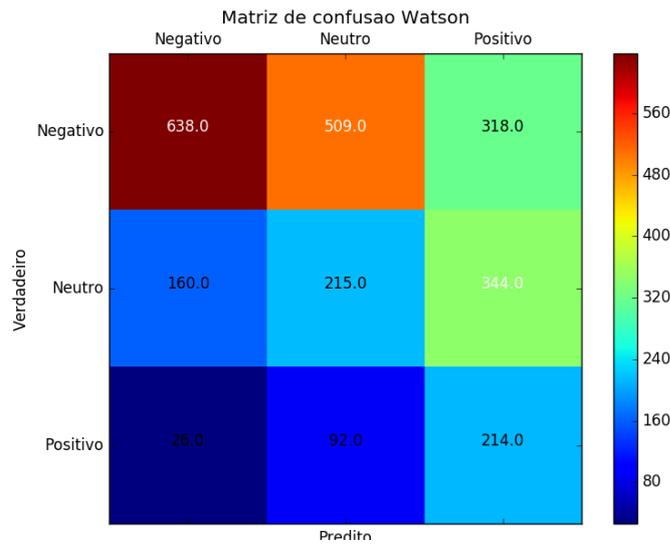


Figura 14 – Matriz de Confusão do Watson

Os serviços cognitivos da Microsoft disponibiliza diversas ferramentas, uma delas é o *Text Analytics* que contém o módulo para análise de sentimento, no qual o usuário passa o texto por uma API REST e tem como retorno um Json, contendo a resposta do sentimento no texto em forma numérica decimal que está no intervalo entre 0 e 1. Logo, para tratar isso realizamos uma conversão desses valores seguindo a lógica:

- Se valor ≥ 0.6 , recebe o rótulo 1 (positivo);
- Se valor > 0.4 e valor < 0.6 , recebe o rótulo 0 (neutro);
- Se valor ≤ 0.4 , recebe o rótulo -1 (negativo).

No que se diz respeito ao desempenho, como pode-se observar na Tabela 6, o Microsoft *Text Analytics* teve uma acurácia de 58 %, portanto obteve uma taxa de acerto menor que o método proposto neste trabalho. Vale destacar, que os valores foram calculados com base na matriz de confusão representada na Figura 15

Tabela 6 – Métricas de Desempenho do Watson e *Text Analytics*

	Acurácia	Precisão	Recall	F1 Score	Erro
Watson	0.42	0.55	0.42	0.48	0.98
T. Analytics	0.58	0.65	0.58	0.62	0.63

Análise Estatística

Como foi feito com os algoritmos anteriormente também será executado para essas duas ferramentas, a fim de verificar se os métodos tem diferença estatística. Inicialmente aplicamos o teste de Shapiro Wilk, no qual foi recusado, negando a hipótese nula, logo após foi aplicado o teste de Friedman que obteve o p-valor igual a 0.0008. Portanto devemos

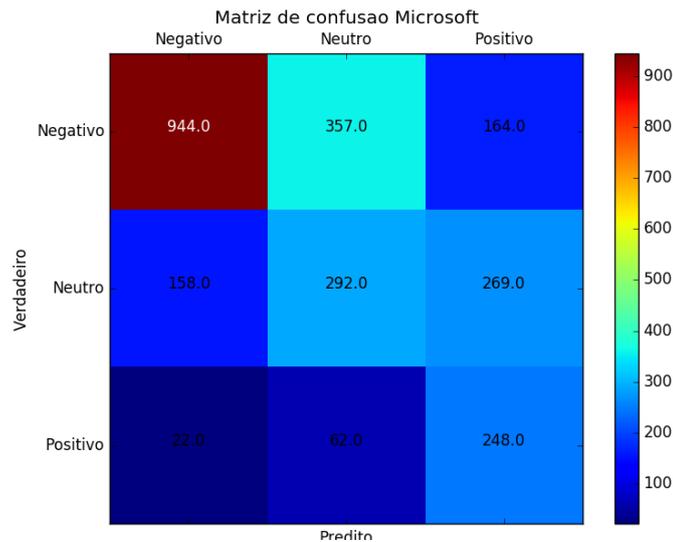


Figura 15 – Matriz de Confusão do *Text Analytics*

recorrer a mais um teste, o *post-hoc* Nemenyi, em que tem como resultado a Tabela 7 para o Watson e a Tabela 8 para o *Text Analytics*, como podemos observar nas tabelas ambos tiveram diferença em relação ao Comitê. Assim podemos concluir que o Comitê é mais adequado para classificar sentimento em comentários do Twitter nas condições que foram utilizados, por ter uma maior taxa de acerto maior do que as ferramentas Watson NLU e *Text Analytics*.

Tabela 7 – Tabela Com os P-valores Gerados a Partir do Teste *Post-hoc* de Nemenyi Para o Watson

	Comitê	R. Forest	R. Logística	SVM	N. Bayes	A. Decisão
R. Forest	0.522					
R. Logística	0.908	0.028				
SVM	0.502	1	0.026			
N. Bayes	0.985	0.084	0.999	0.026		
A. Decisão	0.0004	0.254	6.4e-7	0.269	5.1e-16	
Watson	2e-16	2e-16	2e-16	2e-16	2e-16	2e-16

Tabela 8 – Tabela Com os P-valores Gerados a Partir do Teste *Post-hoc* de Nemenyi Para o Microsoft *Text Analytics*

	Comitê	R. Forest	R. Logística	SVM	N. Bayes	A. Decisão
R. Forest	0.522					
R. Logística	0.908	0.028				
SVM	0.502	1	0.026			
N. Bayes	0.985	0.084	0.999	0.026		
A. Decisão	0.0004	0.254	6.4e-7	0.269	5.1e-16	
T. Analytics	2.3e-14	8.9e-16	2e-16	9.2e-14	2e-16	2e-10

7 CONCLUSÃO

A análise de sentimentos em redes sociais é um tema que necessita de métodos que obtenham maior precisão em suas predições, principalmente para a língua portuguesa que conta com poucos trabalhos se comparado com a gama em inglês. Este trabalho apresentou uma técnica de análise de sentimentos em comentários do Twitter em português utilizando da abordagem por aprendizado de máquina e comparou o seu desempenho com os demais métodos que são utilizados na literatura e no mercado.

Os teste realizados iniciaram pela avaliação de desempenho das técnicas, que mostraram que o método proposto utilizando o comitê de algoritmos de aprendizado de máquina tem maior taxa de acerto, precisão e um menor erro em relação aos algoritmos utilizados de forma isolada e aos demais métodos apresentados neste trabalho. uma vantagem neste aspecto é que literatura conta com poucos trabalhos que tem uma acurácia acima de 80 % para um conjuntos de dados de *tweets* na língua portuguesa.

Após mensurar os algoritmos em relação ao seus desempenho foi realizado alguns testes estatístico a fim de validar se o método proposto tem diferença para os algoritmos isolados, bem como para o TSviz, Watson e *Text Analytics*. Aplicando o primeiro teste chegamos a conclusão que o conjunto de amostras não segue uma distribuição normal, logo foi aplicado outro teste para verificar se o comitê tem diferença para os demais. Portanto pode-se concluir que o mesmo não tem diferença estatística para os algoritmos Naive Bayes, Random Forest, SVM e Regressão Logística, no entanto tem diferença para a Árvore de Decisão, Watson e Microsoft *Text Analytics*.

Portanto, chegamos a conclusão que o comitê é uma proposta interessante para ser utilizada em um o conjunto de dados que contém poucos registros, como foi feito neste trabalho. Porém o algoritmo Naive Bayes pode ser indicando para bases bem maiores, pois segundo o gráfico ROC do mesmo ele apresenta os melhores modelos de classificação para essa perspectiva e também tem um tempo de execução menor que os outros algoritmos utilizados.

Como trabalhos futuros pode-se considerar algumas questões em aberto neste trabalho como: a utilização de uma base de dados maior para o treinamento dos algoritmos, pois com esse tipo de base o desempenho pode ser maior; outro ponto é a utilização de base de dados em outras linguagens, para comparar o desempenho com a base em português e também realizar a classificação de sentimentos multilinguagens, isto é, classificar tweets em outros idiomas como inglês e espanhol; atualmente foi apresentado neste trabalho a classificação multiclases sendo positivo, negativo e neutro, futuramente pretende-se estender esse conceito, incluindo a polaridade do sentimento, ou seja, se o mesmo foi forte

positivo, forte negativo, fraco positivo, fraco negativo e neutro; e uma proposta interessante que se visa incluir, que vem sendo comentada na comunidade acadêmica é o conceito de aprendizado profundo, no qual o mesmo pode ser utilizado para classificar texto, com o auxílio das redes neurais.

REFERÊNCIAS

- AL-AMRANI, Y.; LAZAAR, M.; ELKADIRI, K. E. Sentiment Analysis using supervised classification algorithms. In: *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications - BDCA '17*. New York, New York, USA: ACM Press, 2017. p. 1–8. ISBN 9781450348522.
- Asiaee T., A. et al. If you are happy and you know it... tweet. In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. New York, New York, USA: ACM Press, 2012. p. 1602.
- AUGUSTYNIAK, L. et al. Simpler is better Lexicon-based ensemble sentiment classification beats supervised methods. In: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. [S.l.]: IEEE, 2014. p. 924–929.
- AVANÇO, L. V.; NUNES, M. d. G. V. Lexicon-based sentiment analysis for reviews of products in brazilian portuguese. In: *2014 Brazilian Conference on Intelligent Systems*. [S.l.: s.n.], 2014. p. 277–281.
- BENAVOLI, A.; CORANI, G.; MANGILI, F. Should we really use post-hoc tests based on mean-ranks. *Journal of Machine Learning Research*, v. 17, n. 5, p. 1–10, 2016.
- BERNARDINI, F. C. *Combinação de classificadores simbólicos para melhorar o poder preditivo e descritivo de Ensembles*. São Carlos: Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, 2002.
- BESPALOV, D. et al. Sentiment classification based on supervised latent n-gram analysis. In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, 2011. p. 375.
- BISSACOT, A. C. G. *Estudo Comparativo Entre Regressão Logística Binária e Redes Neurais Artificiais Na Avaliação Dos Resultados Clássicos Hosmer, Lemeshow e Sturdivant*. Dissertação (Mestrado) — Universidade Federal de Itajubá, 2015.
- BOUAZIZI, M.; OHTSUKI, T. Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets. In: *2016 IEEE Global Communications Conference (GLOBECOM)*. [S.l.]: IEEE, 2016. p. 1–6. ISBN 978-1-5090-1328-9.
- BRAVO-MARQUEZ, F.; FRANK, E.; PFAHRINGER, B. From unlabelled tweets to twitter-specific opinion words. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2015. (SIGIR '15), p. 743–746. ISBN 978-1-4503-3621-5.
- BREIMAN, L. Random Forests. *Machine Learning*, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001.
- CASTANHEIRA, L. G. Aplicação de técnicas de mineração de dados em problemas de classificação de padrões. *Belo Horizonte: UFMG*, 2008.

- CHAMANSINGH, N.; HOSEIN, P. Efficient sentiment classification of Twitter feeds. In: *2016 IEEE International Conference on Knowledge Engineering and Applications (IC-KEA)*. [S.l.]: IEEE, 2016. p. 78–82.
- DAYTON, C. M. Logistic regression analysis. *Stat*, p. 474–574, 1992.
- DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, 2006. ISSN 1532-4435.
- DING, X.; LIU, B.; ZHANG, L. Entity discovery and assignment for opinion mining applications. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2009. (KDD '09), p. 1125–1134. ISBN 978-1-60558-495-9. Disponível em: <http://doi.acm.org/10.1145/1557019.1557141>.
- FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. *Livros Técnicos e Científicos*, Grupo Gen - LTC, 2011.
- FARIAS, E. d. S. *Relevância da Tradução de Textos de Português para Inglês no processo de Classificação Binária de Sentimento de postagens rápidas em Redes Sociais Online*. Dissertação (Mestrado) — Universidade Federal de Campina Grande, 2016.
- FARIAS, V. A. E. d. *Uma Abordagem para a Modelagem de Desempenho e de Elasticidade para Bancos de Dados em Nuvem*. 2016.
- FRANÇA, T. C. et al. Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. *XXIX Simpósio Brasileiro de Banco de Dados-SBBD*, v. 14, 2014.
- FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937.
- FÜRNKRANZ, J. Round robin classification. *J. Mach. Learn. Res.*, JMLR.org, v. 2, p. 721–747, mar. 2002. ISSN 1532-4435.
- GAMA, J. *Knowledge discovery from data streams*. [S.l.]: CRC Press, 2010.
- GOMES, H. J. C. *Text Mining: análise de sentimentos na classificação de notícias*. Tese (Doutorado), 2013.
- GÓMEZ, S. N. Dissertação de Mestrado, *Random Forests Estocástico*. 2012.
- GONÇALVES, P. et al. Panas-t: Uma escala psicométrica para medição de sentimentos no twitter. 2012.
- GRANDIN, P.; ADAN, J. M. Piegas: A systems for sentiment analysis of tweets in portuguese. *IEEE Latin America Transactions*, v. 14, n. 7, p. 3467–3473, July 2016. ISSN 1548-0992.
- HAN, J. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 1558609016.

- HOGENBOOM, A. et al. Exploiting emoticons in sentiment analysis. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2013. (SAC '13), p. 703–710. ISBN 978-1-4503-1656-9. Disponível em: <<http://doi.acm.org/10.1145/2480362.2480498>>.
- KRZANOWSKI, W. J. An introduction to statistical modelling. Arnold, a member of the Hodder Headline Group., 1998.
- LOPES, M. M.; BRANCO, V. T. F. C.; SOARES, J. B. Utilização dos testes estatísticos de kolmogorov-smirnov e shapiro-wilk para verificação da normalidade para materiais de pavimentação. *Transportes*, v. 21, n. 1, p. 59–66, 2013.
- LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Estratégias para a Combinação de Classificadores Binários em Soluções Multiclasses. *Revista de Informática Teórica e Aplicada*, v. 15, n. 2, p. 65–86, dec 2008.
- LUNARDI, A. de C.; VITERBO, J.; BERNARDINI, F. C. Análise de sentimentos utilizando técnicas de classificação multiclasse. *Tópicos em Sistemas de Informação: Minicursos SBSI 2016*, 2016.
- MALHOTRA, R. *Empirical Research in Software Engineering: Concepts, Analysis, and Applications*. [S.l.]: Chapman & Hall/CRC, 2015. ISBN 1498719724, 9781498719728.
- MARTELETO, R. M. Análise de redes sociais: aplicação nos estudos de transferência da informação. *Ciência da informação*, SciELO Brasil, v. 30, n. 1, p. 71–81, 2001.
- MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, v. 5, n. 4, p. 1093–1113, dec 2014.
- MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.
- MITTAL, A.; GOEL, A. *Stock Prediction Using Twitter Sentiment Analysis*. 2012.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. In: *Sistemas inteligentes: fundamentos e aplicações*. [S.l.: s.n.], 2003. p. 89–114.
- MORENO, Á. C. *Análise de Sentimentos na Classificação de Comentários Online Aplicando Técnicas de Text Mining*. Dissertação (Mestrado) — Instituto Universitário de Lisboa, 2015.
- MOZETIC, I.; GRČAR, M.; SMAILOVIĆ, J. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, Public Library of Science, v. 11, n. 5, p. 1–26, 05 2016.
- MUKHERJEE, S. et al. Twisent: A multistage system for analyzing sentiment in twitter. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2012. (CIKM '12), p. 2531–2534. ISBN 978-1-4503-1156-4.
- NASCIMENTO, L. C.; CRUZ, C. B. M. *Mineração de dados e adaptação de modelos de classificação de cobertura e uso da terra para imagem Worldview 2*. Foz Iguaçu: Simpósio Brasileiro de Sensoriamento Remoto, 2013.

- NORVIG, P.; RUSSELL, S. *Inteligência Artificial, 3a Edição*. [S.l.]: Elsevier Brasil, 2014.
- OLIVEIRA, R. B. *Método de detecção e classificação de lesões de pele em imagens digitais a partir do modelo Chan-Vese e máquina de vetor de suporte*. 2012.
- PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (EMNLP '02), p. 79–86.
- PEREIRA, J. M. et al. Predição da neutralização do efeito coagulante da peçonha de *Bothrops pauloensis* pelo extrato aquoso de *Hedychium coronarium* (zingiberaceae) através de modelos de regressão = prediction of neutralization of clotting effect of venom *Bothrops pauloensis*... *Bioscience Journal*, v. 26, n. 6, 2010.
- PEREIRA, M. J. S. *Processamento de linguagem natural para produtos de seguros*. 2014.
- PRATI, R.; BATISTA, A.; MONARD, M. *Curvas ROC para avaliação de classificadores*. 2008.
- PRATI, R. C. *Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos*. Tese (Doutorado) — Universidade de São Paulo, São Carlos, jul 2006.
- RAPOSO, L. M. *Aplicação de Regressão Logística e Redes Neurais Probabilísticas Na Classificação de Resistência Aos Antirretrovirais Lopinavir e Nelfinavir*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2014.
- RECUERO, R.; ZAGO, G. Em busca das "redes que importam": redes sociais e capital social no twitter. *LÍBERO*. ISSN impresso: 1517-3283/ISSN online: 2525-3166, n. 24, p. 81–94, 2016.
- REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Barueri SP: Editora Manole Ltda, 2003.
- RIHAB, B.; ELKHELIFI, A.; FaBouchlahem Rihabiz, R. A Machine Learning Approach For Classifying Sentiments in Arabic tweets. In: *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS '16*. New York, New York, USA: ACM Press, 2016. p. 1–6.
- RIOS, R. A. et al. Analyzing the public opinion on the brazilian political and corruption issues. In: *6th Brazilian Conference on Intelligent Systems, BRACIS 2017*. [S.l.: s.n.], 2017.
- RODRIGUES, H. J. F. *Ferramenta para text mining em textos completos*. 2016.
- ROSA, R. L. *Análise de sentimentos e afetividade de textos extraídos das redes sociais*. Tese (Doutorado) — Universidade de São Paulo, 2015.
- ROSA, R. L.; RODRIGUEZ, D. Z.; BRESSAN, G. Sentimeter-br: A social web analysis tool to discover consumers' sentiment. In: *2013 IEEE 14th International Conference on Mobile Data Management*. [S.l.: s.n.], 2013. v. 2, p. 122–124. ISSN 1551-6245.

- ROSSI, A. L. D. *Ajuste de parâmetros de técnicas de classificação por algoritmos bioinspirados*. Tese (Doutorado) — Universidade de São Paulo, 2009.
- SANTOS, A. G. L. dos; BECKER, K.; MOREIRA, V. Um estudo de caso de mineração de emoções em textos multilíngues. 2014.
- SANTOS, E. M. *Teoria e Aplicação de Support Vector Machines à Aprendizagem e Reconhecimento de Objetos Baseado na Aparência*. 2002.
- SANTOS, T. M. dos. *Avaliação do desempenho de modelos preditivos no contexto de análise de sobrevivência*. São Paulo: Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, 2013.
- SILVA, L. M. O. *Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN Para a Identificação de Modelos ARMA Não-Sazonais e Sazonais*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil, set 2005.
- SILVA, N. F. D.; HRUSCHKA, E. R.; HRUSCHKA, E. R. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, Elsevier, v. 66, p. 170–179, 2014.
- SILVA, N. F. F. d. *Análise de sentimentos em textos curtos provenientes de redes sociais*. Tese (Doutorado) — Universidade de São Paulo, 2016.
- SILVA, N. F. F. da. *Análise de sentimentos em textos curtos provenientes de redes sociais*. Tese (Doutorado) — Universidade de São Paulo, mar 2016.
- SKLEARN. *Multinomial Naive Bayes*. 2017. Disponível em: <http://scikit-learn.org/stable/modules/naive_bayes.html>.
- SOARES, H. B. *Análise e Classificação de Imagens de Lesões da Pele Por Atributos de Cor, Forma e Textura Utilizando Máquina de Vetor de Suporte*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2008.
- SOUZA, B. A. et al. For or against? polarity analysis in tweets about impeachment process of brazil presidente. In: *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web - Webmedia '16*. New York, New York, USA: ACM Press, 2016. p. 335–338. ISBN 9781450345125.
- SOUZA, E. et al. Topie: An open-source opinion mining pipeline to analyze consumers' sentiment in brazilian portuguese. In: _____. *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*. Cham: Springer International Publishing, 2016. p. 95–105.
- SOUZA, E. et al. Characterizing user-generated text content mining: A systematic mapping study of the portuguese language. In: _____. *New Advances in Information Systems and Technologies*. Cham: Springer International Publishing, 2016. p. 1015–1024.
- SOUZA, E. et al. Characterizing opinion mining: A systematic mapping study of the portuguese language. In: _____. *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*. Cham: Springer International Publishing, 2016. p. 122–127.

TORMAN, V. B. L.; COSTER, R.; RIBOLDI, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Clinical & Biomedical Research*, v. 32, n. 2, 2012.

Turchetti Maia, T. *Aprendizado Transdutivo Baseado em Teoria da Informação e Teoria do Aprendizado Estatístico*. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2007.

TWITTER, D. *GET search/tweets*. 2017. Acessado em: 19-06-2017. Disponível em: <https://dev.twitter.com/rest/reference/get/search/tweets>.

WANG, Y. et al. Exploring cross-cultural crowd sentiments on twitter. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM, 2017. (HT '17), p. 321–322. ISBN 978-1-4503-4708-2.

XIA, R.; ZONG, C.; LI, S. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, v. 181, n. 6, p. 1138 – 1152, 2011. ISSN 0020-0255.

ZAGO, G. d. S.; SILVA, A. L. M. da. Sites de Rede Social e Economia da Atenção: Circulação e Consumo de Informações no Facebook e no Twitter. *Vozes e Diálogo*, v. 13, n. 01, jul 2014. ISSN 2237-4531.

Apêndices

APÊNDICE A – MATRIZ DE CONFUSÃO DOS MÉTODOS UTILIZADOS

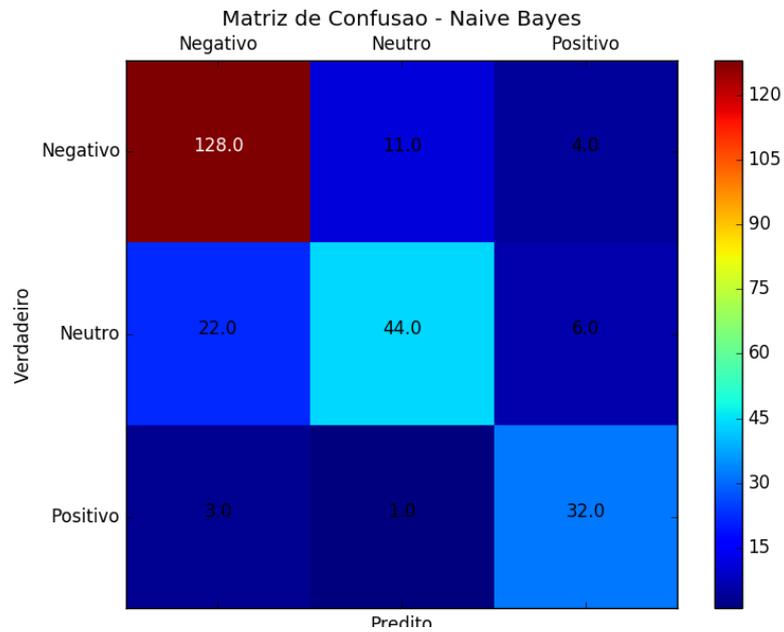


Figura 16 – Matriz de Confusão do Naive Bayes

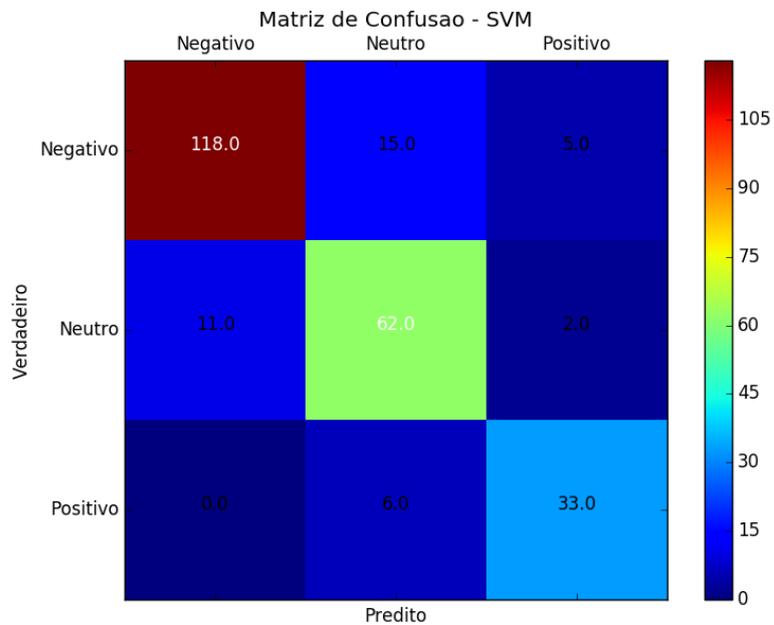


Figura 17 – Matriz de Confusão do SVM

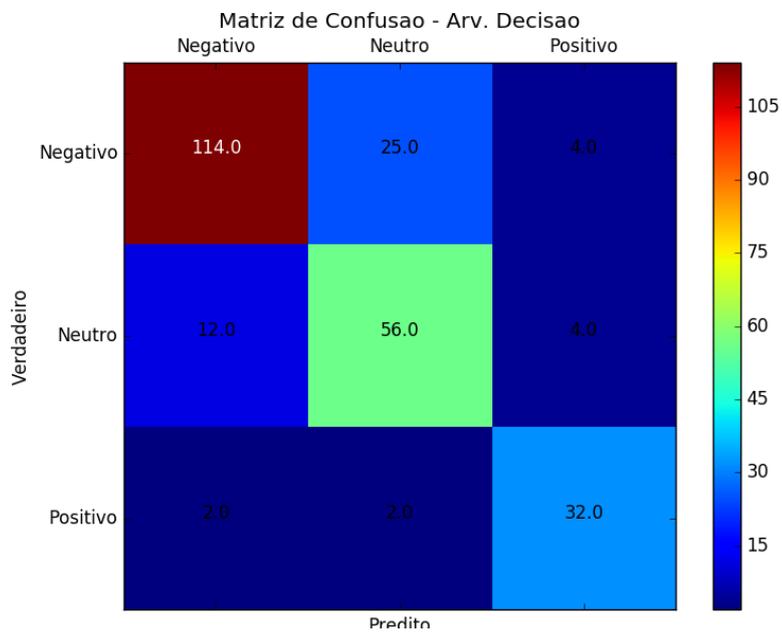


Figura 18 – Matriz de Confusão da Árvore de Decisão

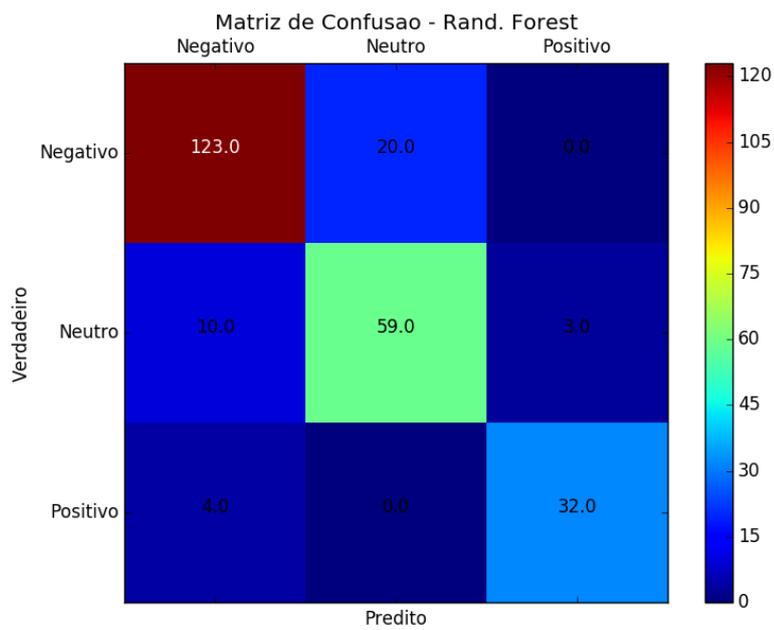


Figura 19 – Matriz de Confusão da Random Forest

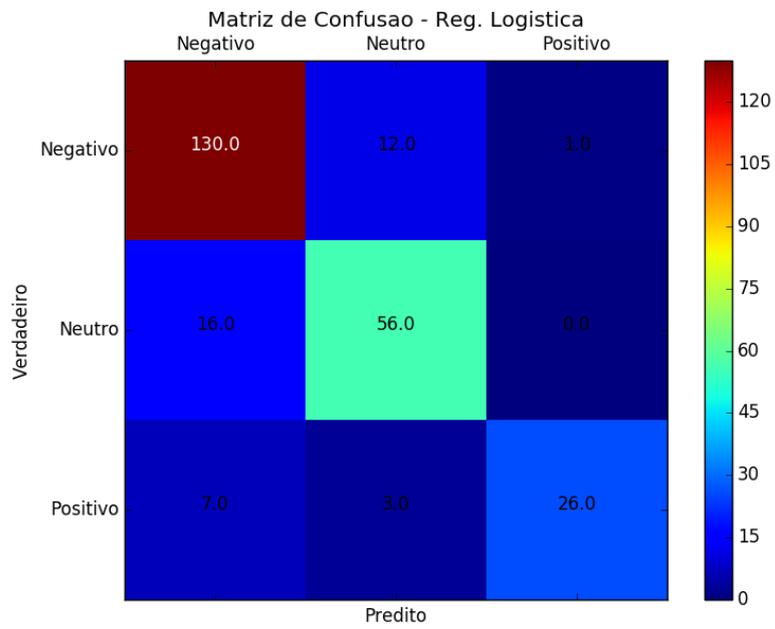


Figura 20 – Matriz de Confusão da Regressão Logística

