Análise de Sentimento em Redes Sociais para a Língua Portuguesa Utilizando Algoritmos de Classificação

Erikson Júlio de Aguiar [1], Bruno S. Faiçal [1,2], Jó Ueyama [2], Glauco Carlos Silva [1], André Menolli [1]

[1]Cento de Ciências Tecnológicas – Universidade Estadual do Norte do Paraná (UENP)
Caixa Postal 261 – 86360-000 – Bandeirantes – PR – Brasil

erjulioaguiar@gmail.com, {bsfaical, glauco, menolli}@uenp.edu.br

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Carlos – SP – Brasil

bsfaical@alumni.usp.br

joueyama@icmc.usp.br

Abstract. *Sentiment Analysis in social network has been explored in different types of research, and its main intention is to extract user's opinion on a wide range of subjects, making possible to obtain significant information. For Portuguese, Sentiment Analysis research is still being established. In this context, this work proposes a method to estimate sentiment in social networks for the Portuguese language, focusing on Twitter. For such purpose, we used a Committee, which is implemented through a set of machine learning algorithms for classification. The evaluation of the proposed method was performed using statistical and performance tests. The results indicate that the Committee had better accuracy when compared to other machine learning algorithms for the performance tests. Nevertheless, there was no proven statistical difference between the Committee's algorithm and some of the algorithms, indicating that these methods could achieve equivalent accuracy to the Committee in some specific situations, such as a larger database.*

Resumo. *A Análise de Sentimento em redes sociais vem sendo explorada em diferentes tipos de pesquisas, tendo como principal intuito extrair opiniões dos usuários sobre os mais diversos assuntos, possibilitando que informações significativas sejam obtidas. No Brasil, as pesquisas sobre Análise de Sentimento ainda estão se estabelecendo. Com base nesse contexto, este trabalho propõe um método para estimar sentimentos em redes sociais para a língua portuguesa, tendo como foco o Twitter. Para tal, é utilizado um Comitê, que é implementado por meio de um conjunto de algoritmos de aprendizagem de máquina para classificação. A avaliação do método proposto foi realizada utilizando testes estatísticos e de desempenho. Os resultados obtidos indicam que o Comitê teve melhor acurácia se comparado a outros algoritmos de aprendizagem de máquina para os testes de desempenho. Contudo, não foi comprovada diferença estatística entre o Comitê e alguns dos algoritmos, o que pode indicar que estes métodos podem alcançar acurácia equivalente ao Comitê em algumas situações específicas, como por exemplo, uma base de dados maior.*

1. Introdução

As redes sociais vêm se tornando uma importante fonte para análise de dados, uma vez que nestas estão dados advindos de diferentes perfis de usuários sobre os mais diversos assuntos. Especificamente para análise de sentimento, o Twitter é uma base rica, pois concentra publicações de usuários com tamanho padronizado, e na grande maioria estão na forma textual. A partir deste cenário, surgiram abordagens baseada na análise de sentimento, que objetivam extrair informações úteis sobre as publicações, como a proposta de [Gonçalves et al. 2012], que tem a finalidade de analisar os textos e identificar a opinião do usuário sobre ele. A técnica de 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.

A análise de sentimento vem sendo tema de diversos estudos, tanto no ambiente acadêmico como em [Ding et al. 2009], bem como na esfera da indústria, como afirma [Rosa 2015]. Esta técnica pode ser de grande valia para as empresas, pois é possível entender, por exemplo, a opinião de clientes sobre determinado produto. Dessa forma, com o uso intensivo das redes sociais, a análise de dados para esse meio vem ganhando cada vez mais importância, como afirma [França et al. 2014]. 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 e o números de pesquisas sobre o assunto vem crescendo. Todavia, [da Silva 2016] declara que o número de trabalhos sobre o tema ainda é bem limitado, principalmente devido à complexidade de se trabalhar com texto, principalmente advindos de redes sociais, que é desestruturado e o processo para fazer o tratamento do mesmo exige um custo de tempo alto.

Especificamente para a língua portuguesa, o número de trabalhos neste tema é ainda mais limitado. Encontram-se poucos trabalhos que fazem a análise de textos na língua portuguesa, e os trabalhos existentes em sua maioria não apresentem resultados satisfatórios. Sabendo dessa realidade, este trabalho tem a finalidade de apresentar uma abordagem, baseada em algoritmos de classificação, para estimar o sentimento em comentários do Twitter para a língua portuguesa.

2. Trabalhos Relacionados

Na literatura existem diversos estudos envolvendo o tema análise de sentimento em texto, alguns utilizam abordagens que empregam métodos léxicos, outros o aprendizado de máquina. Alguns desses estudos envolvem a análise de sentimento em texto de modo geral, isto é para qualquer tipo de documento de texto. Por outro lado, outros estudos apresentam características mais parecidas com este trabalho, tratando da análise de sentimento em redes sociais, como o Twitter.

No trabalho de [Gonçalves et al. 2012], é discutida 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 e eventos esportivos. Para a realização dessa tarefa, foram coletados cerca de 1,8 bilhões de *tweets*, a partir de 2006 até agosto de 2009. Como primeiro passo, os autores aplicaram a limpeza do texto, que consistiu em aplicar a radicalização, remover as *stop words*, caracteres indesejados e URLs. Em seguida aplicou-se a *tokenização*

do texto, utilizando o espaço como separador dos *tokens*. Os sentimentos que podem ser classificados pelas técnicas são: (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 o H1N1 e a queda do avião da AirFrance em 2009. Como resultado, os testes realizados com a técnica visando o H1N1, identificaram sentimentos de atenção e medo. Para os textos referentes ao acidente da AirFrance, foram identificados medo e tristeza.

Sobre a classificação de sentimento utilizando abordagens de aprendizado de máquina, pode-se combinar algoritmos de classificação, como [Silva 2016] apresenta em seu trabalho, que propõe o método de Comitê para a análise de sentimento em texto das redes sociais. O Comitê proposto foi implementado a partir dos seguintes algoritmos: Naive Bayes Multinomial, Máquina de vetor de suporte (SVM), *Random Forest* e Regressão Logística, por serem bem citados na literatura para esse tipo de aplicação, como afirma [Silva 2016]. O trabalho tem como foco a análise de textos obtidos do Twitter, e utilizou as seguintes bases de *tweets* já rotuladas em inglês: (i) Sanders, (ii) Stanford Twitter Corpus, (iii) Obama-McCain Debate e (iv) *Health Care Reform*. Por fim, após realizada a análise de desempenho da técnica proposta, resultados interessantes foram obtidos, alcançando um valor máximo de acurácia de 84.89% para a base de *tweets* Sanders.

Diferente dos trabalhos apresentados, que na sua maioria são para a análises de texto na língua inglesa, o presente estudo tem foco voltado para a análise de sentimento em textos da rede social Twitter na língua portuguesa. Também é proposta a utilização da abordagem do Comitê, empregando algoritmos de aprendizado de máquina para rotular o sentimento no texto. Assim, pretende-se obter um método de análise de sentimentos para a língua portuguesa, que apresente resultados satisfatórios tanto de acurácia como confiança nas predições.

3. Abordagem Proposta

Para alcançar o objetivo deste trabalho, nesta seção é apresentada a abordagem proposta para estimar o sentimento em textos curtos na língua portuguesa, mais especificamente nas publicações da rede social Twitter. A estrutura da abordagem é apresentada na Figura 1, e é utilizada a técnica que combina a predição de algoritmos de classificação, Comitê.

O Comitê empregado nesse trabalho é formado pelos seguintes algoritmos: (i) Naive Bayes, (ii) SVM, (iii) Árvore de decisão, (iv) Random Forest e (v) Regressão logística (detalhado na seção 3.3). Ainda, o peso (a importância) da classificação individual de cada classificador é atribuído com base na acurácia de classificação apresentada por estes (este processo está descrito na seção 3.5).

3.1. Base de Treinamento

Com o objetivo de treinar os algoritmos de aprendizado de máquina para classificar o sentimento contido nos textos, foi necessário obter uma base de *tweets* já rotulada. Assim, foi realizada uma busca na literatura e em sites específicos, com a finalidade de encontrar conjuntos de dados abertos e em português para a utilização neste trabalho. Foi encontrada uma base, rotulada manualmente, disponibilizada pelo grupo de pesquisa Mi-

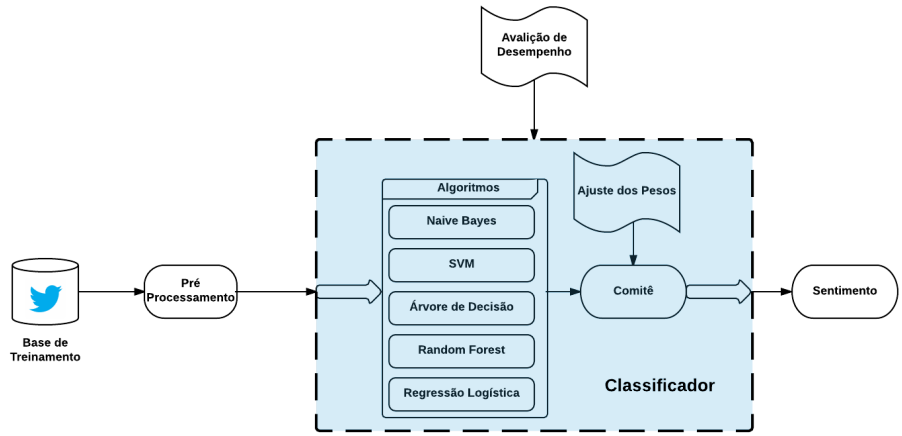


Figura 1. Estrutura da abordagem proposta

ningBR [1] e com sentimentos identificados e classificados como: Negativo (-1); Neutro (0) ou Positivo (1).

O grupo utilizou esses dados para a elaboração de alguns trabalhos, tais como [Souza et al. 2016b, Souza et al. 2016c, Souza et al. 2016a]. No site do grupo, pode-se encontrar a base dividida entre balanceada e desbalanceada. Contudo, neste trabalho foi feita a junção das bases, com o propósito de expandi-la e deixá-la com um número maior de registros. O *dataset* utilizado possui 2516 instâncias, no qual, temos 332 para a classe positiva, 719 neutro e 1465 para negativo.

3.2. Pré-Processamento

A etapa de Pré-Processamento foi dividida em duas sub-etapas (*normalização* e *transformação do texto*) que são descritas na sequência.

Normalização

Nesta etapa foram aplicadas técnicas de processamento de linguagem natural, com a finalidade de deixar o texto mais limpo para as próximas etapas. Para isso, foi empregada a biblioteca *Natural Language Toolkit* (NLTK)[2] do Python, na versão 3.5. Essa biblioteca conta com diversos recursos que são de grande valia para o pré-processamento do texto. Os recursos selecionados para a normalização foram:

1. **Limpeza do texto com expressão regular:** Nesse passo foram retirados sentenças indesejadas do texto, como *links*, que não são significativos para as etapas futuras.
2. **Extração dos *tokens*:** Os *tokens* foram separados utilizando uma função do NLTK, que 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 foram retiradas as *stop words* de cada *tweet*.

[1] `https://sites.google.com/site/miningbrgroup/home/publications`
[2] `http://www.nltk.org/`

Transformação do Texto

Algoritmos de aprendizado de máquina na prática geralmente não entendem o texto puro para realizar suas ações, logo é necessários converter esse texto em algo mais significativo para os mesmos, em que essa função pode ser realizada por meio da técnica conhecida como *Bag of words*, ou saco de palavras em português. Essa técnica transforma a coleção de textos em uma tabela, que indica as palavras e o número de ocorrência dos termos dentro do corpo textual, assim o valor textual é transformado em numérico para ser utilizado pelos algoritmos de classificação [da Silva 2016]. Para realizar esse processo na prática, neste trabalho foi utilizada a biblioteca *Scikit-Learn* (SKLEARN) do Python, que conta com vários recursos, como algoritmos de classificação e agrupamento, bem como de pré-processamento.

3.3. Algoritmos

Na literatura relacionada ao aprendizado de máquina, existem vários algoritmos de classificação. Portanto, realizou-se um levantamento na literatura e, a partir dos trabalhos de [Silva 2016], [Grandin and Adan 2016] e [Augustyniak et al. 2014], foi observado que os classificadores utilizados em estudos correlatos ao contexto deste trabalho foram: (i) Naive Bayes, (ii) Máquina de vetor de suporte (SVM), (iii) Árvore de decisão, (iv) *Random Forest* e (v) Regressão logística. Assim, neste trabalho foi aplicado os algoritmos citados para compor o Comitê.

Nos algoritmos de aprendizagem de máquina, geralmente os valores dos parâmetros são especificados pelo usuário. Contudo, esse processo manual, pode diminuir a taxa de acerto do algoritmo, pois a escolha normalmente se baseia em testes empíricos regidos pela aleatoriedade [Oliveira and Prati 2013].

Na literatura, existem pesquisas focadas somente no objetivo de desenvolver algoritmos otimizados para buscar os melhores parâmetros, isto é, os que proporcionem um erro generalizado menor [Rossi 2009]. Neste trabalho foi empregada uma técnica de otimização de parâmetros, de maneira a conseguir bons parâmetros para os algoritmos selecionados. Para esta etapa foi utilizado o *Grid Search*, que faz a busca direcionada em alguns parâmetros preestabelecidos manualmente, que auxiliarão como espaço de busca para o mesmo. Ou seja, a partir do desempenho alcançado manipulando esses parâmetros o algoritmo sabe por quais deve começar e finalizar sua busca [Farias 2016].

A biblioteca SKLEARN possui uma função que fornece suporte ao *Grid Search*, no qual tem como entrada o algoritmo e o espaço de busca. Essa técnica foi selecionada por ser conhecida na literatura para a identificação de hiperparâmetros [Bergstra et al. 2011]. Apesar de ser uma técnica custosa pela quantidade de avaliações realizadas, o tempo de processamento não resultou em um obstaculo significativo, pois o *dataset* utilizado não possui quantidade de instâncias elevadas.

3.4. Comitê

O Comitê combina N classificadores seguindo alguma regra, que tem como finalidade maximizar a precisão da predição, ou seja, esse método tem como propósito ser melhor do que uma escolha aleatória de algoritmos [Faceli et al. 2011]. A técnica do Comitê pode ser utilizada de algumas formas, no entanto, será citada apenas uma delas, pois as outras não são significativas para este trabalho. A técnica da votação, é apresentada

em [Silva 2016], no qual é descrita da seguinte forma: considerando um conjunto de N classificadores E_1,...,E_N, que realizam a predição para M classes C_1,...,C_M, quando combinados realizam a predição para as classes indicadas, em que a classe com o maior número de votos é o valor do rótulo predito pelo Comitê. Outro ponto que é válido considerar, é que pode-se utilizar o sistema de votação simples ou com pesos entre os algoritmos.

Portanto, para esta abordagem é utilizado o Comitê por votação, e foi optado pela utilização do sistema de votos com peso. Para definir os pesos dos algoritmos, a regra utilizada foi que os algoritmos com maior acurácia tem maior peso.

3.5. Ajuste dos Pesos

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

$$peso_i = \frac{ac_i}{\sum_{j=0}^{n-1} ac_j} \tag{1}$$

Onde:

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

3.6. Avaliação do Desempenho

Ao se aplicar algoritmos de 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, sendo possível assim concluir se o método é válido para se utilizar [Faceli et al. 2011].

Portanto, foi realizado uma pesquisa com a finalidade de explorar na literatura as medidas comumente utilizadas para avaliação de modelos preditivos em aprendizado de máquina. Várias métricas foram encontradas, como [dos Santos 2013] apresenta em seu trabalho. Entretanto, foi decidido utilizar as seguintes métricas para avaliação: (i) para amostragem, foi empregada a validação cruzada (*cross-validation*); (ii) para medidas de desempenho foram: a acurácia, o erro, a precisão, a sensibilidade (*recall*) e o $F1$ *score*; (iii) para a avaliar os classificadores, foi utilizada a curva *Receiver Operating Characteristic* (ROC).

A validação cruzada é empregada segundo [Tavares et al. 2007] com a finalidade de evitar que o resultado da predição centralize em valores que podem ser prejudiciais a predição, quando se utiliza um espaço reduzido de amostras. O processo de validação cruzada particiona a base de dados em conjunto de treinamento e conjunto de teste, no qual, é comumente utilizado na literatura a validação cruzada com o *K-Fold* [Bespalov et al. 2011]. Esse método decompõe a base em treino e teste aleatoriamente, com a possibilidade de sobreposição, refazendo esse processo K vezes.

Para a validação cruzada neste trabalho, utilizamos o K = 10, gerando 10 grupos de teste, que foram executados de uma forma lógica e sequencial a partir da implementação utilizada que vem agregada a biblioteca SKLEARN. Assim são iniciadas as predições seguindo esse processo de divisão, no qual as métricas desempenho dos algoritmos são inseridas para que sejam mensuradas K vezes e após esse processo para gerar os valores finais das métricas é calculada a mediana dos K valores [Faceli et al. 2011].

O gráfico ROC produz valores entre 0 e 1, que formará uma curva baseada nas predições do classificador. Um fator importante para mensurar esse classificador é a área abaixo da curva (AUC), em inglês *Area Under Curve*. Cujo é um índice que varia entre 0 e 1, no qual quanto mais próximo do valor 1, melhor será o classificador. Nesse gráfico é definida uma linha limite, que indica qual o valor mínimo de AUC deve ser considerado para que o mesmo ainda seja viável para a aplicação [Prati et al. 2008].

4. Resultados e Discussões

Esta seção visa apresentar os principais resultados obtidos com a aplicação da abordagem proposta na base selecionada. Além disso, a partir de análises realizadas sobre os resultados, também são apresentadas algumas discussões.

4.1. Análise da Abordagem Proposta

Como descrito na Seção 3.6, algumas métricas foram definidas a fim de avaliar o desempenho dos algoritmos utilizados, tal como a abordagem proposta utilizando o Comitê. A Tabela 1 apresenta os resultados para as métricas de desempenho para cada um dos algoritmos selecionados, assim como para o Comitê. Os 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

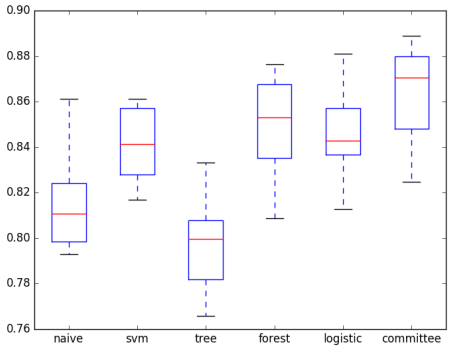
Algoritmos	Acurácia	Precisão	Recall	F1 Score	Erro
Naive Bayes	0.810	0.814	0.810	0.812	0.272
SVM	0.841	0.846	0.841	0.844	0.212
Árvore de Decisão	0.800	0.827	0.800	0.815	0.283
Random Forest	0.856	0.862	0.856	0.859	0.206
Reg. Logística	0.842	0.845	0.842	0.842	0.206
Comitê	**0.865**	**0.866**	**0.864**	**0.865**	**0.184**

Como pode ser observado na Tabela 1, o algoritmo que apresenta melhor acurácia é o Comitê, com 86%. Assim, podemos concluir que o Comitê foi o algoritmo que teve a maior taxa de acertos para a base de dados utilizada (contém 2516 registros para a língua portuguesa). Graficamente pode-se visualizar a diferença das acurácias entre os algoritmos pelo *boxplot* na Figura 2(a).

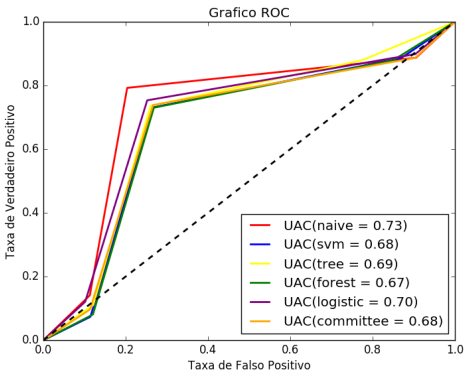
No entanto, o gráfico ROC, como mostra a Figura 2(b), apresenta um comportamento diferente do resultado obtido pela medida de acurácia. De acordo com gráfico ROC, o classificador que gera melhores modelos de classificação para o *dataset* utilizado neste estudo é o Naive Bayes, apresentando um valor de AUC igual a 0.73, sendo superior ao resultado do Comitê (que ficou no meio do *ranking*, com um valor de AUC igual

a 0.68). Para gerar todos os valores das métricas apresentadas, utilizou dos valores da matriz de confusão dos algoritmos.

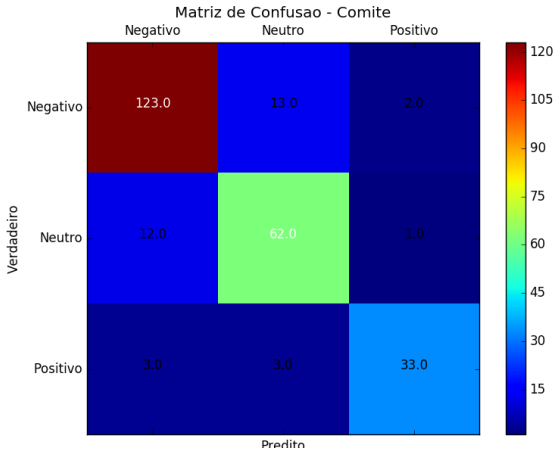
Na Figura 2(c), é apresentado um exemplo que descreve a matriz de confusão do algoritmo que obteve a maior acurácia, ou seja, o Comitê. Para obter os valores da matriz na validação cruzada com k = 10, foi necessário calcular a mediana desses valores, para que assim se obtenha a mediana da matriz de confusão para as iterações da validação cruzada.



(a) *Boxplot* das acurácias dos algoritmos utilizados



(b) Gráfico ROC



(c) Matriz de confusão do Comitê

Figura 2. Desempenhos de classificações alcançados pelos algoritmos

Análise Estatística

Com o propósito de validar os experimentos estatisticamente, foi realizada a análise dos resultados utilizando alguns testes, seguindo as sugestões de [Demsar 2006, Faceli et al. 2011]. Primeiramente foi aplicado o teste de Shapiro Wilk, que segundo [Torman et al. 2012] tem o propósito de verificar se uma amostra segue a distribuição normal. Para o teste de Shapiro Wilk são apresentadas duas hipóteses [Mesquita Lopes et al. 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 a execução do teste de Shapiro Wilk para as amostras, a hipótese H_0 foi negada, aderindo a hipótese H_1. Assim, segundo [Malhotra 2015] é necessário utilizar um teste não paramétrico. Foi escolhido o teste de Friedman, pois o mesmo é utilizado para um conjunto de mais de duas amostras dependentes (neste trabalho queremos comparar seis amostras dependentes). O teste de Friedman [Friedman 1937] é um tipo de teste de hipótese não paramétrico, no qual as amostras são ordenadas seguindo um *ranking* de desempenho. Por exemplo, o algoritmo com melhor acurácia é o número 1 do ranking, o segundo melhor 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 têm diferença, se o p-valor ≤ 0.05.

A execução do teste de Friedman ocorreu com o ordenamento das amostras da seguinte forma: Comitê, *Random Forest*, Regressão Logística, SVM, Naive Bayes e Árvore de Decisão, a partir dos valores de acurácia. Se obteve um p-valor igual a 0.015, em que nega a hipótese nula H_0, logo segundo [Benavoli et al. 2016] tem que se usar um teste *post-hoc* para verificar as diferença significativas entre as amostras.

O teste *post-hoc* Nemenyi segundo [Demsar 2006], é indicado como pós teste de Friedman, o mesmo realiza a verificação comparando as amostras duas as duas, a fim de verificar se as mesmas apresentam diferenças significativa estatisticamente. Portanto, considerando o p-valor 0.05, temos se caso o p-valor seja ≥ 0.05, não há diferença e se for < 0.05, há diferença. A Tabela 2 apresenta os p-valores resultantes deste 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	1.000	0.011		
N. Bayes	0.896	0.026	0.999	0.024	
A. Decisão	0.000	0.140	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 que o Comitê apresenta diferença para todos os outros algoritmos. Analisando a Tabela 2, percebe-se que a hipótese definida não é verdadeira, pois não houve diferença estatística entre o Comitê e a maioria dos algoritmos. Apenas a árvore de decisão apresentou diferença estatística significativa em relação ao Comitê. Portanto, para o *dataset* utilizado, pode-se afirmar que os algoritmos são estatisticamente equivalentes para realizar predições.

Em suma, podemos concluir que para conjuntos de dados pequenos, como o utilizado neste trabalho (possui 2516 registros), o Comitê pode ser mais indicado, pois como apresentado na Figura 2(a) obtém uma maior acurácia. No entanto, para conjunto de dados maiores, pode-se optar por outro algoritmo, como o Naive Bayes, pois como mostrado na Figura 2(b), gera melhores modelos de classificação para a tarefa em questão. Além disso, o tempo de execução do Naive Bayes é menor se comparado aos outros algoritmos que não apresentam diferença para o Comitê, como mostrado a Tabela 3.

Tabela 3. Tempo de execução dos algoritmos utilizados

	N. Bayes	SVM	A. Decisão	R. Forest	R. Logística	Comitê
Tempo (s)	**0.208**	5.624	3.277	1.779	0.725	11.183

4.2. Comparações Frente a Ferramentas Existentes

Existem algumas implementações disponíveis para classificar sentimento em texto, dessa forma, foi realizada uma pesquisa que resultou na seleção de três implementações: (i) método de classificação de sentimento da plataforma TSviz [Rios et al. 2017], utiliza abordagem léxica; (ii) o IBM Watson[3], que tem um módulo de processamento de linguagem natural; e (iii) o módulo de análise de sentimentos dos serviços cognitivos da Microsoft,[4]. Com o intuito de comparar a abordagem proposta com as implementações disponíveis no mercado, o conjunto de dados utilizado nos experimentos anteriores foram testados nas implementações selecionadas, comparando o resultado com o da abordagem proposta.

A abordagem descrita em [Rios et al. 2017] utiliza o Corpus chamado Reli. Este Corpus contém várias palavras extraídas de resenhas de livros clássicos em português. Para calcular o sentimento, primeiramente cada palavra da frase é transformada em um vetor de características, utilizando o conceito de *Word2vec* [5], ou seja, é transformada em um vetor numérico. Assim, pode-se calcular a similaridades das palavras do texto com as palavras do Corpus, a fim de verificar a qual classe pertence. Dessa forma, se a palavra mais similar estiver na classe positiva, a palavra recebe o rótulo positivo dentro da frase, sendo esse procedimento feito para todas as palavras da frase. Após as palavras serem rotuladas, é utilizado de duas equações para a calcular o número de palavras positivas (P) e o número de palavras negativas (N), isto é, soma quantas palavras de cada classe tem dentro da frase (F). Por fim, as equações são normalizadas, cujo para a classe positiva é (NP) e para negativa (NN), logo, esse valores são subtraídos (NP – NN), que resultará no valor do sentimento da frase, que poderá ser -1, 0 ou 1.

As métricas de desempenho obtidas por essa implementação é retratada na Tabela 4. Pode-se perceber, que o mesmo atingiu uma acurácia igual a 43%, apresentando um resultado pior que a abordagem proposta nesse aspecto.

O IBM Watson e o Microsoft *Text Analytics* são implementações mais voltados a aplicações comerciais, pois pertencem a empresas privadas. Contudo, ambas disponibilizam versões de teste, em que seus usuários têm liberdade para utilizar da ferramenta com algumas limitações.

O IBM Watson é composto por várias APIs integradas, bem como a *Natural Language Understanding* (NLU), que contém uma *feature* para análise de sentimentos. Nessa *feature* o usuário insere um texto, e partir desse texto, é identificada a linguagem do texto e o sentimento é analisado, atribuindo os rótulos positivo, negativo ou neutro. Em relação as métricas de desempenho, pode-se ver na Tabela 4 que a acurácia do Watson foi de 42%, apresentando também um resultado pior que a abordagem proposta nesse aspecto.

[3] `https://www.ibm.com/watson/services/natural-language-understanding/`

[4] `https://azure.microsoft.com/pt-br/services/cognitive-services/text-analytics/`

[5] `https://code.google.com/archive/p/word2vec/`

Os serviços cognitivos da Microsoft disponibilizam diversas ferramentas, uma delas é a *Text Analytics*, que contém o módulo para análise de sentimento. Esta ferramenta recebe como entrada textos, por uma API REST, e retorna um arquivo Json, contendo a classificação do sentimento em forma numérica decimal, no intervalo entre 0 e 1. Assim sendo, para utilizar essa ferramenta foi necessário realizar a conversão do resultado apresentado pela ferramenta da seguinte forma:

- 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 as métricas de desempenho, como pode-se observar na Tabela 4, o Microsoft *Text Analytics* teve uma acurácia de 58%. Portanto, obteve uma taxa de acerto inferior à abordagem proposta.

Tabela 4. Métricas de desempenho dos serviços disponíveis

	Acurácia	**Precisão**	**Recall**	**F1 Score**	**Erro**
Watson	0.424	0.558	0.424	0.482	0.986
T. Analytics	0.589	0.654	0.589	0.620	0.631
[Rios et al. 2017]	0.430	0.494	0.430	0.460	0.712

Análise Estatística

A análise estatística é inicialmente realizada para comparar a abordagem de [Rios et al. 2017] com a abordagem proposta. Primeiramente foi empregado o teste de Shapiro Wilk, em que obtivemos um p-valor que recusou a hipótese nula H_0. Assim, seguiu-se para o teste de Friedman, no qual teve um p-valor igual 0,018, logo utilizamos o *post-hoc* Nemenyi para verificar as diferenças significativas entres as técnicas. Pelo teste *post-hoc* Nemenyi foi-se criada a Tabela 5, e partir desta pode-se concluir que a abordagem de [Rios et al. 2017] tem diferença estaticamente significativa em relação a classificação realizada pelo Comitê. Portanto, podemos concluir que o Comitê, por ter uma maior taxa de acerto, apresentou um comportamento mais preciso que a abordagem de [Rios et al. 2017]. É importante ressaltar que tais resultados são vinculados as condições em que o experimento foi realizado.

Assim como para os algoritmos anteriores, as implementações da IBM e Microsoft também foram comparadas por meio de análises estatísticas com a abordagem proposta. Inicialmente aplicamos o teste de Shapiro Wilk, mostrando que as amostras eram não normais, assim foi aplicado o teste de Friedman, que obteve o p-valor igual a 0.0008. Sabendo disso, aplicamos o *post-hoc* Nemenyi, que gerou a Tabela 5 para o Watson, *Text Analytics* e para a abordagem de [Rios et al. 2017]. Como pode ser observado, ambas implementações apresentaram diferença estatística em relação ao Comitê. Assim, podemos concluir que o Comitê apresentou maior eficiência na classificação de sentimentos em comentários do Twitter que as ferramentas Watson NLU e *Text Analytics*, nas condições que foram utilizados.

5. Conclusão e Trabalhos Futuros

A análise de sentimentos é um tema vem ganhando importância com o crescente uso das redes sociais, e necessita de métodos que obtenham boa precisão em suas predições,

Tabela 5. Tabela com os p-valores gerados a partir do teste *Post-hoc* de Nemenyi para os métodos semelhantes

	Comitê	**R. Forest**	**R. Logística**	**SVM**	**N. Bayes**	**A. Decisão**
R. Forest	0.522					
R. Logística	0.908	0.028				
SVM	0.502	1.000	0.026			
N. Bayes	0.985	0.084	0.999	0.026		
A. Decisão	0.000	0.254	6.4e-7	0.269	5.1e-16	
[Rios et al. 2017]	**7.1e-14**	5.9e-14	2e-16	6e-14	2e-16	1.5e-8
Watson	**2e-16**	2e-16	2e-16	2e-16	2e-16	2e-16
T. Analytics	**2.3e-14**	8.9e-16	2e-16	9.2e-14	2e-16	2e-10

principalmente para a língua portuguesa. Considerando essa necessidade, este trabalho propôs uma abordagem para a língua portuguesa, focada na análise de sentimentos em comentários do Twitter. A abordagem proposta é baseada em um Comitê, que é composto de diferentes algoritmos de aprendizado de máquina, e foi comparado o desempenho da abordagem com implementações existentes, que funcionam para a língua portuguesa.

Os testes iniciaram pela avaliação de desempenho do Comitê e dos algoritmos que o compõem. Os resultados indicaram que a abordagem proposta, utilizando o Comitê, tem maior taxa de acerto, precisão e menor erro em relação aos algoritmos utilizados de forma isolada. Posteriormente, a abordagem foi comparada com soluções comercias disponíveis. Os resultados indicaram que a acurácia da abordagem proposta foi superior a todas as soluções comerciais disponíveis. Outro fato de destaque foi que, para o *dataset* utilizado o Comitê apresentou uma acurácia acima de 80%.

Após mensurar a diferença de desempenho, entre o Comitê e os algoritmos que o compõem, o Comitê também foi confrontado com os algoritmos por meio de testes estatístico. Este testes foram realizados a fim de validar se a abordagem proposta apresenta diferença estatística significante para os algoritmos isolados, bem como para as ferramentas disponíveis. Considerando os testes estatísticos para o *dataset* utilizado, verificou-se que o Comitê não apresenta diferença estatística para os algoritmos Naive Bayes, *Random Forest*, SVM e Regressão Logística. No entanto, apresenta diferença para a Árvore de Decisão e para as demais ferramentas, [Rios et al. 2017], Watson e Microsoft *Text Analytics*.

Portanto, é possível concluir que o Comitê é uma proposta interessante para ser utilizada em um o conjunto de dados que contenha poucos registros. Uma possível aplicação da abordagem proposta é o uso do comitê para rotular um grande conjunto de dados a partir do padrão inferido em um conjunto de dados com poucos registros. Tal abordagem permite uma rotulação mais rápida e menos suscetível a erros resultantes a fadiga quando comparado a rotulação manual. Por fim, as rotulações errôneas podem ser consideradas ruídos do conjunto de dados e poderiam ser quantificadas a baixo de 20%, dado a acurácia do Comitê. Em situações em que existe uma grande base de dados é possível investigar o comportamento do algoritmo Naive Bayes. Essa hipótese é sugerida a partir do gráfico ROC, o qual indica que o Naive Bayes permite alcançar os melhores modelos de classificação e não ter sido possível identificar diferença estatisticamente significativa entre o Naive Bayes e o Comitê. Por fim, Naive Bayes apresentou tempo de execução inferior aos outros algoritmos utilizados.

Como trabalhos futuros pretende-se (i) utilizar uma base de dados maior para o

treinamento dos algoritmos; (ii) utilizar base de dados em outros idiomas, para comparar o desempenho com a base em português e (iii) investigar a classificação de sentimentos multilinguagens, isto é, classificar *tweets* em outros idiomas, como inglês e espanhol sem utilizar classificadores específicos.

Referências

Augustyniak, L., Kajdanowicz, T., Szymanski, P., Tuliglowicz, W., Kazienko, P., Alhajj, R., and Szymanski, B. (2014). 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)*, pages 924–929. IEEE.

Benavoli, A., Corani, G., and Mangili, F. (2016). Should we really use post-hoc tests based on mean-ranks. *Journal of Machine Learning Research*, 17(5):1–10.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554.

Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, page 375, New York, New York, USA. ACM Press.

da Silva, N. F. F. (2016). *Análise de sentimentos em textos curtos provenientes de redes sociais*. PhD thesis, Universidade de São Paulo.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Ding, X., Liu, B., and Zhang, L. (2009). Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1125–1134, New York, NY, USA. ACM.

dos Santos, T. M. (2013). Avaliação do desempenho de modelos preditivos no contexto de análise de sobrevivência.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. (2011). Inteligência artificial: Uma abordagem de aprendizado de máquina. *Livros Técnicos e Científicos*.

Farias, V. A. E. d. (2016). Uma abordagem para a modelagem de desempenho e de elasticidade para bancos de dados em nuvem.

França, T. C., de Faria, F. F., Rangel, F. M., de Farias, C. M., and Oliveira, J. (2014). Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. *XXIX Simpósio Brasileiro de Banco de Dados–SBBD*, 14.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Gonçalves, P., Dores, W., Benevenuto, F., and Preto-MG-Brasil, O. (2012). Panas-t: Uma escala psicométrica para mediç ao de sentimentos no twitter.

Grandin, P. and Adan, J. M. (2016). Piegas: A systems for sentiment analysis of tweets in portuguese. *IEEE Latin America Transactions*, 14(7):3467–3473.

Malhotra, R. (2015). *Empirical Research in Software Engineering: Concepts, Analysis, and Applications*. Chapman & Hall/CRC.

Mesquita Lopes, M., Branco, V. T. F. C., and Soares, J. B. (2013). Utilização dos testes estatísticos de kolmogorov-smirnov e shapiro-wilk para verificação da normalidade para materiais de pavimentação. *Transportes*, 21(1):59–66.

Oliveira, G. M. G. and Prati, R. C. (2013). Ajuste de parâmetros em algoritmos de aprendizado de máquina utilizando transferência de aprendizado. *X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, page 3.

Prati, R., Batista, A., and Monard, M. (2008). Curvas ROC para avaliação de classificadores.

Rios, R. A., S., L. C., Pagliosa, P. A., and Mello, R. F. (2017). Analyzing the public opinion on the brazilian political and corruption issues. In *6th Brazilian Conference on Intelligent Systems, BRACIS 2017*.

Rosa, R. L. (2015). *Análise de sentimentos e afetividade de textos extraídos das redes sociais.* PhD thesis, Universidade de São Paulo.

Rossi, A. L. D. (2009). *Ajuste de parâmetros de técnicas de classificação por algoritmos bioinspirados*. PhD thesis, Universidade de São Paulo.

Silva, N. F. F. d. (2016). *Análise de sentimentos em textos curtos provenientes de redes sociais*. PhD thesis, Universidade de São Paulo.

Souza, E., Alves, T., Teles, I., Oliveira, A. L. I., and Gusmão, C. (2016a). *TOPIE: An Open-Source Opinion Mining Pipeline to Analyze Consumers' Sentiment in Brazilian Portuguese*, pages 95–105. Springer International Publishing, Cham.

Souza, E., Castro, D., Vitório, D., Teles, I., Oliveira, A. L. I., and Gusmão, C. (2016b). *Characterizing User-Generated Text Content Mining: A Systematic Mapping Study of the Portuguese Language*, pages 1015–1024. Springer International Publishing, Cham.

Souza, E., Vitório, D., Castro, D., Oliveira, A. L. I., and Gusmão, C. (2016c). *Characterizing Opinion Mining: A Systematic Mapping Study of the Portuguese Language*, pages 122–127. Springer International Publishing, Cham.

Tavares, L. G., Lopes, H. S., and Lima, C. R. E. (2007). Estudo comparativo de métodos de aprendizado de máquina na detecção de regiões promotoras de genes de escherichia coli. *Anais do I Simpósio Brasileiro de Inteligência Computacional*, pages 8–11.

Torman, V. B. L., Coster, R., and Riboldi, J. (2012). Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Clinical & Biomedical Research*, 32(2).