Avaliando o Desempenho da Abordagem de Comitê na Análise de Sentimentos na Língua Portuguesa

Matheus Henrique Cardoso†
Ciência da Computação
Universidade do Vale do Itajaí - UNIVALI
São José, SC, Brasil
matheus.cardoso@edu.univali.br

Anita Maria da Rocha Fernandes
Ciência da Computação
Universidade do Vale do Itajaí - UNIVALI
São José, SC, Brasil
anita.fernandes@univali.com

ABSTRACT

Sentiment Analysis aims extract subjective information from texts that, when written in Portuguese face a number of difficulties related to grammatical nature and vocabulary diversity. In order to collaborate with researches in this area, this paper presents the proposal to evaluate the performance of the committee approach in relation to the traditional approaches of sentiment analysis in the Portuguese language context. As object of application we choose tweets about volleyball theme that will serve as basis for the approaches application. These texts will be treated using Natural Language Processing for better performance of the algorithms. Other approaches will also be used in this study to assist in the evaluation of the committee along with the aid of metrics such as accuracy, precision, recall and F-measure.

KEYWORDS

Sentiment Analysis, Natural Language Processing, Tweet Analysis

1 Introdução

A técnica de Análise de Sentimentos consiste em classificar a polaridade do sentimento existente em um texto como positivo, negativo e neutro [1], sendo que existem abordagens que utilizam outros rótulos tais como alegria, tristeza, raiva e surpresa [2]. Carvalho Filho [3] afirma que esta análise é bastante utilizada pelas empresas na preparação de estratégias de *marketing* extraindo informações úteis dos conteúdos publicados pelos consumidores para identificar possíveis melhorias em seus produtos e serviços.

Atualmente existem dois principais tipos de métodos de classificação presentes na literatura. O primeiro refere-se a técnicas supervisionadas, que são baseadas em aprendizado de máquina e que necessitam de uma grande base de dados rotulada para treinamento e teste. O segundo refere-se a técnicas não-supervisionadas, que fazem uso de tratamentos léxicos, cálculos e dicionários léxicos para classificação do sentimento contido em cada palavra do texto [4]. Também existe uma terceira abordagem que utiliza um comitê em que são agrupados dois ou mais classificadores com a finalidade de maximizar a precisão da predição [5]. Esta última abordagem será o foco deste trabalho.

Sabendo que o número de trabalhos relacionados a Análise de Sentimento em textos em português brasileiro é bem limitado, um estudo realizando uma avaliação da abordagem de comitê em relação as demais abordagens encontradas pode contribuir com a literatura das áreas de Aprendizado Supervisionado e Não-Supervisionado, Processamento de Linguagem Natural e Análise de Sentimentos.

Nas subseções a seguir serão descritas as etapas da Análise de Sentimento e serão resumidas as dificuldades encontradas na realização da Análise de Sentimento em textos extraídos das redes sociais.

1.1 Análise de Sentimentos em Redes Sociais

Tradicionalmente a Análise de Sentimentos é utilizada em textos mais longos, portanto, ambientes que possibilitam a criação de textos curtos e com linguagem coloquial como o Twitter são cenários que propõem desafios como o tamanho do texto, a variação da ortografia, esparsidade dos dados, existência de negação, existência de símbolos especiais, variação dos tópicos, grande quantidade de dados, estilo de linguagem variável e a utilização de mais um idioma em uma mesma frase [6].

O Twitter é uma das redes sociais mais populares e ela mantém uma rica base para Análise de Sentimentos por conta da sua limitação da quantidade máxima de caracteres das publicações que na sua grande maioria está na forma textual [5].

1.2 Abordagem de Comitê

O comitê, ou *ensemble*, é um sistema composto por um conjunto de classificadores que podem utilizar a mesma ou diferentes técnicas de classificação. A etapa de treinamento destes classificadores pode ser realizada utilizando algumas técnicas de organização e separação dos dados para treino e teste. As

principais encontradas na literatura são o *Bagging*, o *Boosting* e o *Stacking* [7].

Normalmente a abordagem de comitê tem um custo computacional elevado quando comparado a abordagens tradicionais [8]. Esta elevação do custo computacional se caracteriza na etapa de treinamento dos modelos que compõe o comitê. Esta desvantagem pode ser resolvida com técnicas de computação paralela [9]. Segundo Whitehead and Yaeger [10], esta abordagem aumenta a acurácia, porém em troca exige um esforço computacional. O comitê pode ser utilizado quando se deseja um alto percentual de precisão e quando o tempo de processamento é irrelevante. A simples combinação de classificadores em um comitê não garante um alto nível de acurácia e precisão, mas reduz a chance de se obter os piores resultados [6]. Segundo Bordin Junior [11], existem algumas pesquisas que mostram que a utilização do comitê é mais eficiente do que a utilização de um classificador único.

2 Trabalhos Correlatos

Serão apresentados neste tópico, os trabalhos correlatos que foram encontrados através de buscas realizadas utilizando o motor de pesquisa acadêmica do Google, chamado Google Acadêmico. Foram filtrados os trabalhos publicados entre o ano de 2014 e 2019, utilizando as palavras chave "análise de sentimento" e "língua portuguesa".

Em Aguiar et al. [5] foi proposta a utilização da abordagem do comitê utilizando os algoritmos Naive Bayes, SVM (*Support Vector Machine*), Árvore de Decisão, *Random Forest* e Regressão Logística na sua composição. Devido a estes algoritmos utilizarem aprendizado de máquina, foi utilizada uma base de dados com *tweets*, escritos na língua portuguesa, já rotulados disponibilizada pelo grupo de pesquisa MiningBR contendo 2.516 textos, sendo 1.465 com sentimento negativo, 719 com sentimento neutro e 332 com sentimento positivo. A comparação entre o comitê e os algoritmos que o constitui apresentou que o comitê teve melhor acurácia, com 86%. Em um estudo similar utilizando algoritmos da abordagem supervisionada, Silva [6] propõe a utilização do método do comitê para classificação de frases escritas na língua inglesa que foram extraídas das redes sociais. O comitê foi implementado utilizando os seguintes algoritmos: Naive Bayes Multinomial, SVM (*Support Vector Machine*), *Random Forest* e Regressão Logística. Foram realizadas comparações entre algoritmos de classificação, os sistemas com os melhores resultados encontrados na literatura e o comitê. A pesquisa confirmou que a abordagem de comitê tem o potencial igual ou superior a abordagens de classificação tradicionais no contexto de redes sociais.

3 Solução Proposta

Considerando a complexidade encontrada no processo de Análise de Sentimentos em textos da língua portuguesa advindos das redes sociais, este trabalho pretende expor as três diferentes abordagens citadas à textos extraídos do Twitter referente ao contexto esportivo, mais especificamente no contexto da Liga das Nações de Vôlei de 2019, do Torneio Pré-Olímpico de Voleibol de 2019 e da Copa do Mundo de Voleibol de 2019, com a finalidade de avaliar o desempenho da abordagem de comitê em relação as abordagens tradicionais da análise de sentimentos na operação de classificação da polaridade dos sentimentos encontrados nos textos como positivo e negativo.

A Figura 1 representa resumidamente a aplicação da solução.

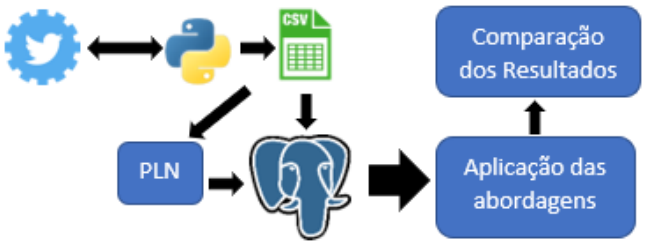

Figura 1: Solução proposta.

Cada uma das abordagens será representada por um ou mais algoritmos com o intuito de possibilitar a realização de experimentos que resultarão dados que serão avaliados e sumarizados para melhor apresentação à literatura. A abordagem supervisionada será representada neste trabalho pelos algoritmos Naive Bayes, SVM (*Support Vector Machine*), Árvore de Decisão e Regressão Logística. Já a abordagem não-supervisionada será representada pelos algoritmos PANAS-t, SenticNet, SentiWordNet, LIWC (*Linguistic Inquiry and Word Count*), a versão em português do LIWC, Sentiment140 Lexicon, Opinion Lexicon e VADER (*Valence Aware Dictionary for sEntiment Reasoning*). A abordagem de comitê representada pelo agrupamento de alguns destes algoritmos onde serão organizados utilizando *Bagging*, *Boosting* e *Stacking* que, segundo Silva [6], são técnicas que visam melhorar a distribuição das amostras de dados para obter uma melhor generalização dos modelos.

4 Metodologia

Foi realizada uma revisão sistemática da literatura e levantamento dos trabalhos correlatos para analisar e selecionar as abordagens de análise de sentimentos juntamente com seus algoritmos de classificação mais utilizados para utilização neste estudo. Em

paralelo, foi criado um *dataset* de treinamento com frases extraídas do Twitter relacionadas a Liga das Nações de Voleibol, ao Torneio Pré-Olímpico de Voleibol e a Copa do Mundo de Voleibol. Para captura dos *tweets* foi desenvolvido um programa em Python que realiza conexão com a API (*Application Programming Interface*) da própria rede social e armazena os resultados em um arquivo no formato de planilha CSV (*Comma-separated values*). A captura dos *tweets* relacionados aos temas aconteceu durante o período em que estas competições estivavam sendo realizadas e parte dos dados serão utilizados para treinamento e outros para testes.

A captura dos textos iniciou no dia 6 de maio e finalizou no dia 31 de outubro totalizando cerca de 15.485 tweets distribuídos da seguinte forma entre os campeonatos: 2.218 sobre a Copa do Mundo de Voleibol, 6.379 sobre a Liga das Nações de Voleibol e 6.888 sobre o Pré-Olimpico. Os *tweets* foram categorizados manualmente e receberam rótulos de negativo ou neutro ou positivo. Após o término da etapa de rotulação, 2.979 textos foram descartados devido não pertencerem aos temas pesquisados, 10.176 textos foram classificados como neutro (não apresentam sentimento), 1.298 textos foram classificados como positivo e 1.032 como negativo.

Após a rotulação dos textos, as frases do *dataset* serão preparadas para serem utilizadas pelos algoritmos. Esta preparação basicamente é a aplicação das técnicas de processamento de linguagem natural onde os dados serão normalizados a partir da aplicação destas técnicas. Em cada um dos tweets, todas as letras maiúsculas serão transformadas em minúsculas com a aplicação da função *LowerCase*. Com a remoção dos *stopwords*, serão extraídos dos textos alguns elementos dispensáveis para identificação e polarização do sentimento contigo no texto como *links*, menção de usuários, *hashtags*, conjunções, preposições, pronomes e artigos. Outra função será encarregada de retirar os numerais que também não possuem valor sentimental. Devido os textos serem extraídos de redes sociais, será utilizado um corretor ortográfico para evitar palavras inexistentes no vocabulário. Também serão aplicadas as técnicas de tokenização, onde o texto é segmentado em pequenos agrupamentos de palavras (uni-grama, bi-grama, tri-grama, etc) e stemização, em que cada palavra do texto é transformada em seu radical, sendo retirada a sua terminação.

Serão implementados sistemas na linguagem de programação Python a fim de desenvolver a lógica dos algoritmos de classificação que foram selecionados. Para implementação dos algoritmos que utilizam aprendizado de máquina serão utilizadas algumas bibliotecas disponíveis para a linguagem. Estes recursos da linguagem Python são bastante utilizados pelos pesquisados tanto da área de Mineração de Texto como da área de Análise de Sentimentos e também foram utilizados pelos autores dos trabalhos correlatos.

Com os dados rotulados e dados tratados, será criado um banco de dados PostgreSQL contendo o texto original, o texto tratado e sua respectiva polaridade. Com estes dados serão realizadas algumas etapas de treinamentos e de testes dos algoritmos de aprendizado de máquina. Após todas as etapas anteriores estiverem concluídas, serão realizados alguns experimentos para comparação das abordagens e seus algoritmos. Para sumarização dos resultados obtidos serão utilizadas algumas métricas como acurácia, precisão, *recall* (revocação ou sensibilidade) e F-*measure* (F1-score ou F-score) para melhor acompanhamento dos experimentos e comparação dos resultados.

5 Considerações Finais

Como resultado deste trabalho, espera-se disponibilizar à literatura um trabalho com resultados das avaliações e comparações entre as abordagens mencionadas, principalmente da abordagem de comitê, para que possa nortear os leitores em suas pesquisas durante a escolha das abordagens e algoritmos a serem utilizados nos seus experimentos e também, servir de auxílio para trabalhos futuros relacionados à área de Análise de Sentimentos.

REFERÊNCIAS

[1] R.L. Rosa, 2015. Análise de sentimentos e afetividade de textos extraídos das redes sociais. Tese (Doutorado em Sistemas Digitais) – Universidade de São Paulo, São Paulo, Brasil.

[2] H.B. Brum, 2015. Análise de sentimentos para o português usando redes neurais recursivas. Monografia (Graduação em Ciência da Computação) – Universidade Federal do Pampa, Alegrete, Brasil.

[3] J. A. Carvalho Filho, 2014. Mineração de textos: análise de sentimentos utilizando tweets referentes à Copa do Mundo de 2014. Monografia (Graduação em Engenharia de Software) – Universidade Federal do Ceará, Cerará, Brasil.

[4] F. Benevenuto, F. Ribeiro and M. Araújo, 2015. Métodos para análise de sentimentos em mídias sociais. *In* Simpósio Brasileiro de Sistemas Multimidia e Web, Manaus, Brasil.

[5] E.J. Aguiar et al., 2018. Análise de sentimento em redes sociais para a língua portuguesa utilizando algoritmos de classificação. *In* XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, Porto Alegre Brasil.

[6] N.F.F. Silva, 2016. Análise de sentimentos em textos curtos provenientes de redes sociais. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo, São Paulo, Brasil.

[7] W.K.N. Silva and A.M Santos (2017). Estratégias de Construções de Comitês de Classificadores Multirrótulos no Aprendizado Semissupervisionado Multidescrição. Revista de Informática Teórica e Aplicada, 24(2):71-100.

[8] R. Xia, C. Zong and S. Li (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138-1152.

[9] S. Sun, C. Zhang and D. Zhang (2007). An experimental evaluation of ensemble methods for EEG signal classification. Pattern Recognition Letters, 28(15), 2157-2163.

[10] M. Whitehead and L. Yaeger, 2008. Sentiment mining using ensemble classification models. *In* Innovations and Advances in Computer Sciences and Engineering, pages 509-514.

[11] A. Bordin Junior, 2018. Aplicação de programação genética na análise de sentimentos. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal de Goiás, Goiânia, Brasil.