Análise de Sentimentos: Uma Comparação entre Diferentes Abordagens no Contexto da Língua Portuguesa

Matheus Henrique Cardoso[1], **Anita Maria da Rocha Fernandes**[1], **Sandro Alexandre Ribeiro de Aguiar**[2]

[1]Curso Ciência da Computação – Campus Kobrasol
Escola do Mar, Ciência e Tecnologia – Universidade do Vale do Itajaí (UNIVALI)
Caixa Postal 360 – 88.302-901 – São José – SC – Brasil

[2]MBA em Big Data – FATEC – SENAI – Cuiabá - MT

{matheus.henriq.c, sandro.aguiar}@gmail.com, anita.fernandes@univali.br

Abstract. *Sentiment Analysis aims extract subjective information from texts that, when written in Portuguese face a number of difficulties. In order to collaborate with research in the area, this project proposes the comparison of different approaches of sentiment analysis in the context of the Portuguese language. As object of application of the study, will be collected tweets related to the volleyball theme for creation of a database. The approaches considered in this study are based in machine learning, based in not supervisioned learning (lexical) and based in committee.*

Resumo. *Análise de sentimentos tem por objetivo extrair informações subjetivas de textos que, quando escritos em português enfrentam uma série de dificuldades. A fim de colaborar com as pesquisas na área, este projeto propõe a comparação de diferentes abordagens de análise de sentimentos no contexto da língua portuguesa. Como objeto de aplicação do estudo, serão coletados tweets relativos ao tema voleibol para criação de um database. As abordagens consideradas neste estudo são a baseada em aprendizado de máquina, baseada em aprendizado não-supervisionada (léxica) e baseada em comitê.*

1. Introdução

A técnica de Análise de Sentimentos consiste em classificar a polaridade do sentimento existente em um texto como positivo, negativo e neutro [Rosa 2015], sendo que existem abordagens que utilizam outros rótulos tais como alegria, tristeza, raiva e surpresa [Brum 2015]. 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 utilizam tratamentos léxicos, cálculos e dicionários léxicos para classificação do sentimento contido em cada palavra do texto [Benevenuto et al. 2015]. Também existe uma terceira abordagem que utiliza um comitê que agrupa dois ou mais classificadores com a finalidade de maximizar a precisão da predição [Aguiar et al. 2018].

Considerando a complexidade encontrada no processo de Análise de Sentimentos em textos da língua portuguesa, este trabalho pretende expor as três diferentes abordagens

citadas à textos referente ao contexto esportivo extraídos do Twitter, 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 comparar os resultados obtidos na operação de classificação da polaridade dos sentimentos encontrados nos textos como positivo e negativo. A seguir são apresentados os trabalhos correlatos encontrados através de uma revisão sistemática na literatura, bem como os materiais e métodos e os resultados esperados.

2. Trabalhos Correlatos

Em Aguiar et al. (2018) foi proposta a utilização da abordagem de comitê utilizando os algoritmos Naive Bayes, SVM (*Support Vector Machine*), Árvore de Decisão, *Random Forest* e Regressão Logística. Foi utilizada uma base de dados com *tweets* 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 Tabela 1 mostra os resultados do experimento destacando o comitê com 86% de acurácia.

Tabela 1. Resultados obtidos por Aguiar et al. (2018)

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
Regressão Logística	0.842	0.845	0.842	0.842	0.206
Comitê	**0.865**	**0.866**	**0.864**	**0.865**	**0.184**

Em um estudo similar utilizando algoritmos da abordagem supervisionada, Silva (2016) propõe a utilização do método do comitê para classificação de frases extraídas das redes sociais na língua inglesa utilizando os seguintes algoritmos: Naive Bayes Multinomial, *Support Vector Machine*, *Random Forest* e Regressão Logística. Os resultados dos experimentos confirmaram que a abordagem de comitê tem o potencial igual ou superior a abordagens de classificação tradicionais no contexto de redes sociais. Já Benevenuto et al. (2015) realizou experimentos relacionando algoritmos da abordagem não supervisionada (léxica) com os *datasets* onde as métricas utilizadas foram a acurácia, a precisão, a revogação e F1 Score. Após a execução dos experimentos, foi calculada a média dos resultados para geração de um ranking geral no qual se destacaram os algoritmos SentiStrength, AFinn, OpinionLexion, Umigon e VADER (*Valence Aware Dictionary and sEntiment Reasoner*).

3. Materiais e Métodos

A metodologia proposta para este trabalho engloba oito etapas que serão descritas abaixo.

1. Revisão da literatura. A revisão será focada nos algoritmos que abrangem as três abordagens a serem consideradas.
2. Análise e seleção dos algoritmos de cada uma das abordagens que serão utilizados na comparação. Esta etapa será realizada com base na revisão da literatura e trabalhos correlatos.

3. Criação de 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. Será utilizada uma API (Application Programming Interface) que coletará os *tweets*. Alguns *tweets* serão utilizados para treinamento e outros para testes.
4. Após a coleta dos *tweets*, será necessário escolher um método para rotular as frases do *dataset* de treinamento. Será necessário analisar os métodos mais utilizados (com base na literatura) para selecionar o que melhor se aplica ao problema.
5. Uma vez escolhido o método de rotulação, as frases do *dataset* serão preparadas para serem utilizadas pelos algoritmos. Esta preparação se baseia nas técnicas de Processamento de Linguagem Natural.
6. Implementação dos algoritmos selecionados na Etapa 2. Para isto serão utilizadas as bibliotecas Python para a abordagem de aprendizado de máquina.
7. Treinamento e teste dos algoritmos.
8. Comparação dos resultados obtidos.

Para realização deste trabalho foi desenvolvido um programa escrito em Python que captura os *tweets* que possuem, em seu conteúdo, um dos seguintes termos: ”VNL”, ”*Volleyball Nations League*”, ”liga das nacoes de volei”, ”pre olimpico”, ”pre-olimpico”e ”copa do mundo de volei”. Os *tweets* capturados são armazenados em formato CSV *(Comma-separated values*) em um serviço de armazenamento em nuvem chamado One-Drive. A captura dos *tweets* começou a ser realizada no dia 06 de maio de 2019 fornecendo, até a presente data, 11.705 textos que serão tratados e, posteriormente, rotulados para que possam servir como recurso de entrada para os experimentos. Os programas para aplicação das três abordagens serão desenvolvidos em Python.

4. Resultados Esperados

Como resultado deste estudo, espera-se disponibilizar à literatura um trabalho com resultados das comparações entre as abordagens mencionadas para que possa orientar trabalhos futuros e nortear leitores durante a escolha da abordagens e algoritmos a serem utilizados nos seus trabalhos.

Referências

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

Benevenuto, F., Ribeiro, F., and Araújo, M. (2015). Métodos para análise de sentimentos em mídias sociais. In *Anais do Simpósio Brasileiro de Sistemas Multimidia e Web*, Manaus, Brasil.

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

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