Análise de Sentimento sobre Filmes no Contexto do Twitter

[1]Vinicius G. Rocha, [1]Anita Maria Fernandes, [2]Sandro Alexandre Ribeiro de Aguiar

[1]Curso de Ciência da Computação – Campus Kobrasol - Universidade do Vale do Itajaí (UNIVALI) – São José, SC – Brasil

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

viniciusfunny@gmail.com, anita.fernandes@univali.br, sandro.aguiar@gmail.com

***Abstract**: This paper presents a study on approaches to sentiment analysis in the Portuguese language, having as a case study the theme films. The paper proposes the comparison of framework approaches: Naive Bayes, which is a probabilistic classifier, OpLexicon, which is a lexical dictionary of the Portuguese language, and the Committee approach, composed by the Naive Bayes and OpLexicon algorithms.*

***Resumo:** Este artigo apresenta um estudo sobre abordagens para análise de sentimento na língua portuguesa, tendo como estudo de caso o tema filmes. O trabalho propõe a comparação de três abordagens: Naive Bayes, que é um classificador probabilístico, OpLexicon, que é um dicionário léxico da língua portuguesa, e a abordagem de Comitê, composta pelos algoritmos Naive Bayes e OpLexicon*

1. Introdução

A análise de sentimentos ou mineração de opinião refere-se à verificação dos termos que compõem um texto, isto é, se o documento analisado exprime uma opinião positiva, negativa ou neutra [França e Oliveira 2013]. "Também é possível classificar qual sentimento está presente em um texto (por exemplo, raiva, felicidade, tristeza, etc.)" [França e Oliveira 2013, p.129]. As redes sociais têm se tornado um "termômetro" sobre as opiniões das pessoas em relação aos mais variados assuntos. No Twitter, por exemplo, discussões são vistas como uma plataforma para divulgação e marketing e produtos. Este é o caso de filmes, onde produtores têm investido massivamente em publicidade e marketing voltado aos usuários do Twitter.

Este trabalho apresenta uma pesquisa que busca conhecer o sentimento das opiniões e classificar os pontos de vistas dos usuários do Twitter em positivo ou negativo ou neutro, usando três abordagens para determinar a polaridade: o Naive Bayes que é um algoritmo classificador probabilístico de aprendizagem de máquina, OpLexicon que é um dicionário léxico na língua portuguesa e o uso de um Comitê composto pelo Naive Bayes e OpLexicon, tendo como conjunto de dados opiniões de filmes na língua portuguesa.

A abordagem léxica também conhecida como classificação baseada em recursos léxicos tem como principal uso a utilização de dicionários de palavras e/ou expressões

que possuem classes predefinidas e valores. Esta técnica possui bons resultados quando usada em sentenças ou textos pequenos como o *twetts*, pois a característica é mais próxima a entidade.

Naive Bayes é um algoritmo probabilístico com base na regra de Bayes com seleção de recursos independentes. O algoritmo possibilita a não restrição do número de classes ou atributos. Como o algoritmo é baseado nas regras de Bayes ele possibilita que as probabilidades condicionais sejam importantes, pois é possível que alterne em torno da condição de uma forma conveniente ao problema [Weiand 2016].

Esta abordagem é composta por uma coleção de algoritmos de classificadores que tem como objetivo produzir uma saída final do sistema quando agregado a um método de combinação e conseguir uma classificação mais assertiva e com menos erro. Os comitês classificadores podem ser de dois tipos homogêneos ou heterogêneos. Neste trabalho, o comitê implementado é composto por dois algoritmos classificadores heterogêneos: o Naive Bayes e o OpLexicon.

2. Metodologia

Para a execução do trabalho, foi necessário realizar um levantamento sobre os trabalhos relacionados a análise de sentimento em língua portuguesa, a fim de identificar as abordagens mais utilizadas.

Em paralelo a isto, iniciou-se a coleta de tweets utilizando o Standard Search (API do Twitter). Foram selecionados cinco filmes dos indicados a melhor filme no Oscar 2019. Foram eles Bohemian Rhapsody, Infiltrado na Klan, Nasce Uma Estrela, Pantera Negra, Green Book: O Guia. Foram coletadas cerca de 1421 opiniões a respeito dos filmes, através da API do twitter no dia 27 de fevereiro de 2019. O intervalo que as opiniões foram feitas, foi nos dias 24 de fevereiro de 2019 a 26 de fevereiro 2019. Este intervalo e o tema de “melhores filmes indicados ao Oscar” foi escolhido devido a API do twitter fornecer tweets somente sete dias atrás a partir do dia atual, pois segundo a documentação da API do twitter os dados que são relevantes para a rede social são os dados mais atuais. Logo, o tema dos melhores filmes indicados ao Oscar de 2019, escolhido pelo trabalho estava em alto debate nas redes sociais, devido à proximidade do evento.

Depois recolher 1421 opiniões, elas foram salvas no Banco de Dados MYSQL de maneira original para preservar os dados. Partir de então foi copiado os dados originais para outra tabela e classificadas de forma manual a sua polaridade. Após a rotulação das polaridades das frases armazenadas no MYSQL, os tweets passaram por um pré-processamento linguístico para remoção de *ruídos*, *stopwords* e *hashtags*, para então ser aplicado as abordagens propostas. Assim os dados foram classificados nas seguintes categorias de sentimentos: positivos, negativos e neutros. A Tabela 1 descreve como ficou a classificação dos dados.

Tabela 1. Descrição da polaridade dos dados coletados

Filmes	Total por filme	Positivo	Negativo	Neutro
Bohemian Rhapsody	300	106	19	175
Green Book: O Guia	229	114	7	108
Infiltrado na Klan	298	93	12	193

Nasce Uma Estrela	294	90	11	193
Pantera Negra	300	169	22	109
Total	1421	572	71	778

Do total dos dados recolhidos foram usados cerca de 80% de cada categoria (positivo: 458, negativo: 57, neutra: 622) para o treino do algoritmo Naive Bayes. Já os 20% restantes de cada categoria (positivo: 114, negativo: 14, neutra: 156) foram usados para a validação das abordagens para o resultado final. A forma de validação dos algoritmos foi feita de forma quantitativa separada pelas seguintes polaridades (positiva, negativa e neutra). Após a geração dos resultados dos algoritmos, eles foram comparados com as opiniões originais já polarizadas para então determinar o resultado final.

3. Conclusões

No presente momento a pesquisa está na fase preliminar de testes. Estes primeiros testes apresentam que o algoritmo OpLexicon teve um melhor desempenho (Tabela 2). Os testes com a abordagem de Comitê ainda não foram realizados.

Tabela 2. Descrição dos resultados por algoritmos

Algoritmos	Acertos Positivos	Acertos Negativos	Acertos Neutro
OpLexicon	61%	14%	72%
Naive Bayes	49%	21%	86%

Referências

Graça Neto. (2016). Sentimentalista: Um framework para análise de sentimentos baseado em processamento de linguagem natural. Dissertação (Programa de Pós-Graduação em Ciência e Tecnologia da Computação) - Universidade Federal de Itajubá, Minas Gerais, Brasil.

Weiand. (2016). Análise de sentimentos do Twitter com Naive Bayes e NLTK. In *Revista Científica Trajetória Multicursos*, Osório, Brasil.

França e Oliveira. (2013). Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre junho e agosto de 2013. In *BRASNAM - III Brazilian Workshop on Social Network Analysis and Mining*, Brasília, Brasil.

Nascimento e Medeiros. (2016). Aplicação de Comitês de Classificadores no Aprendizado Semissupervisionado Multidescrição. In *XIII ENIAC*, Recife. Brasil.

Sklearn. (2018). **Scikit-learn.** Acessado em 15 de agosto de 2018. Disponível em: <http://scikit-learn.org>.