Text Classification for Twitter Sentiment

1st Igor Mourão Ribeiro
Computer Engineering Departament
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brazil
igormr98mr@gmail.com

2nd Isabelle Ferreira de Oliveira
Computer Engineering Departament
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brazil
isabelle.ferreira3000@gmail.com

3rd José Luciano de Morais Neto
Computer Engineering Departament
Instituto Tecnológico de Aeronáutica
São José dos Campos, Brazil
zluciano.t19@gmail.com

***Abstract*—Esse relatório documenta a implementação de algoritmos de Processamento de Linguagem Natural, aplicando diferentes técnicas de Machine Learning para classificar Tweets entre sentimentos positivos e negativos. Os algoritmos implementados foram Naive Bayes e Support Vector Machine, utilizando diferentes features produzidas a partir de um dataset do Kaggle, e comparando-os pelas métricas de acurácia e coeficiente Kappa.**

***Index Terms*—Processamento de Linguagem Natural, Naive Bayes, Aprendizagem Supervisionada, Support Vector Machine**

I. INTRODUCTION

Um dos aspectos relevantes da interação entre as pessoas na atualidade é a expressão de sentimentos por meio de textos nas mídias sociais. Nesse contexto, o monitoramento das redes sociais pode ser explorado como forma de extrair a aceitação e/ou aprovação de produtos e também obter conhecimento dos usuários. A análise de sentimentos surge da necessidade de tratar e interpretar textos, opiniões e comentários realizados pelos usuários em redes sociais. Por meio das informações subjetivas extraídas textos em linguagem natural, pode ser gerado conhecimento estruturado, auxiliando a tomada de decisão.

A expansão da Internet e a utilização das redes sociais definiram um ecossistema de interação, no qual os usuários deixaram de ser receptores passivos e se tornaram produtores, compartilhadores e avaliadores de conteúdo. Em um cenário onde as reputações de empresas e a aceitabilidade de produtos no mercado são diretamente afetadas pela repercussão de opiniões de seus clientes na web, tanto quanto pelas campanhas de publicidade, a análise de sentimentos surge como um diferencial para rastreamento do conteúdo emocional daquilo que se escreve e compartilha nas redes sociais. Nesse sentido, a análise de sentimentos alia-se à publicidade promovendo subsídios para definição de estratégias e garantia da vantagem competitiva.

A análise de sentimentos ser aplicada na gestão de informação por exemplo, fornecendo feedback do cliente a partir do conteúdo dos diversos canais de comunicação e entregando informações úteis para tomada de decisão e definição de estratégias para satisfação dos clientes.

O Twitter [2] é uma rede social e servidor para microblogging muito utilizada, que será utilizada para essa análise de sentimentos. Atualmente, o limite máximo de um *tweet* (mensagem postada no blog) é de 280 caracteres e tem-se um total de 6000 *tweets* por segundo o que implica em 200 bilhões por ano. Exemplos de *tweets* com emoções podem ser vistos nas Figuras 1 e 2.

Fig. 1. *Tweet* com mensagem positiva

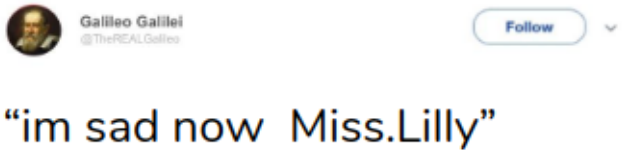

Fig. 2. *Tweet* com mensagem negativa

Neste sentido, este trabalho tem como objetivo apresentar a análise de sentimentos aplicada a textos em linguagem natural de uma rede social, usando diversos modelos e entradas para posterior comparação.

II. RESULTS AND DISCUSSIONS

III. CONCLUSION

Pode-se concluir através dos resultados que o SVM, mesmo exigindo um processamento muito grande e trabalhando com 10% do tamanho padronizado de dados, foi o que teve o pior resultado, sendo assim um classificador ruim se comparado aos demais.

Sendo assim, o método de *Naive-Bayes* foi consideravelmente se baseando nos valores de Kappa obtidos. Sendo o método em *Word-Level* ligeiramente melhor que os demais, embora todos, tenham tido um desempenho similar, com exceção do método em *Char-Level* que teve um desempenho um pouco menor entre os que foram feitos baseados em *Bayes*.

REFERENCES

[1] Dataset de Sentimentos, https://www.kaggle.com/kazanova/sentiment140.

[2] Twitter, https://twitter.com

[3] Kibriya, Ashraf & Frank, E. & Pfahringer, Bernhard & Holmes, Geoffrey. (2004). Multinomial naive Bayes for text categorization revisited. Advances in Artificial Intelligence. 488-499.

[4] SciKit-Learn, https://scikit-learn.org/stable/

[5] Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1): 37–46. doi:10.1177/001316446002000104

[6] Implementing SVM and Kernel SVM with Python's Scikit-Learn https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/

[7] sklearn.metrics.cohen_kappa_score https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html