

Comparativo entre Modelos de Classificação: Identificando Sentimentos de Textos em Português Brasileiro

Antonia Claésia da Costa Souza¹

¹ Instituto Federal de Educação, Ciência e Tecnologia do Ceará Campus Crato
CEP 63115-500 - Crato - CE - Brasil

claesia.costa@gmail.com

Abstract. *This paper aims to perform a comparative analysis of classification models, and to analyze the popular reaction to a crime of a severe nature. The aim is to show how the text mining process was used to make the data collection, processing of the extracted text and the creation of a classification model that allowed to make the mapping of the opinion of Twitter users. With the analyzed data, this work has demonstrated that 55,04% of the tweets represent repudiation.*

Resumo. *Este trabalho tem por objetivo realizar um comparativo entre modelos de classificação, e analisar a reação popular frente a um crime de natureza grave, mostrando como o processo de mineração de textos foi utilizado para fazer a coleta de dados, processamento do texto extraído e a criação de um modelo de classificação que permitiu fazer o mapeamento de opinião dos usuários do Twitter. Com os dados analisados, demonstrou-se que uma expressante parcela (55,04%) dos tweets foram de repúdio ao crime acontecido.*

1. Introdução

As redes sociais *online* têm se tornado um importante ambiente informacional e de debates, onde as pessoas expressam e compartilham suas experiências e opiniões, se tornando impressionante a quantidade de usuários ativos e o volume de dados criados diariamente nessas redes, e deixaram de ser apenas receptores passivos de informações e tornaram-se os reais responsáveis pela produção, compartilhamento, interação e avaliação do conteúdo produzido na Web [Oliveira et. al. 2016].

Este trabalho objetiva realizar um comparativo entre modelos de classificação e com isso analisar a polaridade expressa nas mensagens dos usuários do Twitter. Para isso, foi feito um comparativo de diferentes modelos de classificação, cujo o que apresentou melhor desempenho foi utilizado para a análise. Desta forma, foi realizado um estudo de caso acerca do assassinato da vereadora Marielle Franco que aconteceu no Rio de Janeiro na noite do dia 14 de março de 2018.

2. Experimento

2.1. Coleta de Dados

A coleta de dados está relacionada com o processo de recuperação das mensagens referente ao assassinato. A coleta de dados foi feita através de *hashtags* (marcadores que permitem o agrupamento de tweets). As *hashtags* utilizadas para a consulta foram escolhidas através de uma observação dos termos mais utilizados para agrupar os *tweets* referente ao tema. Os resultados das consultas realizadas foram armazenados em uma

base dados, a coleta foi feita em *tweets* postados até um mês após o crime, com publicação entre 15 de Março e 14 de Abril de 2018. As *hashtags* utilizadas para coleta foram #marielle, #mariellefranco e #mariellepresente, contando com 32.997 *tweets* coletados.

2.2. Pré-Processamento

O pré-processamento é responsável por remover informações irrelevantes que não contribuem para a classificação de sentimento. Foi realizada a remoção das *hashtags*, URLs e caracteres de pontuação presente no *tweet*, além da remoção de menções a perfis do Twitter, onde caracteriza-se pela utilização de @ (at). Vale ressaltar que *retweets* foram preservados, pois foi considerado que ao um usuário compartilhar uma mensagem está em concordância com o que ela expressa, demonstrando assim seu sentimento através da mesma.

Após isso, as mensagens passaram por um procedimento de remoção de *stopwords*, palavras com pouca relevância e que não contribuem para distinção do texto [Francisco 2018]. Para isso, foi utilizado uma *stoplist* (grupo de palavras classificado como *stopwords*) do português brasileiro presente do módulo NLTK¹, e a operação de *stemming* ou radicalização, que é o processo de reduzir as variações de uma palavra ao seu provável radical ou *stem* com o objetivo de associar palavras semelhantes e diminuir a dimensionalidade da representação do texto [Francisco 2018].

2.3. Classificadores

Segundo Witten et al.(2016): (1) *Naive Bayes*: É uma das técnicas de classificação mais utilizadas. O classificador assume que um determinado atributo tem um valor que independe de todos os outros atributos, ou seja, os atributos são condicionalmente independentes dado o valor da classe [Costa e Lopes 2015]. (2) *Support Vector Machine (SVM)*: São classificadores em que é fornecido um conjunto de dados X e cada elemento desse conjunto corresponde a “um ponto em um espaço euclidiano R que serão separados em positivos e negativos” [Costa e Lopes 2015]. (3) *Decision Tree*: São classificadores onde um conjunto de características é transformada em uma árvore para depois ser usada “prever se uma determinada instância da base de teste pertence a determinada classe.” [Costa e Lopes 2015]. (4) *Random Forest*: Também um dos algoritmos mais utilizados, devido à sua simplicidade e pela flexibilidade em pode ser utilizado para classificação e também de regressão, além de que o *random forest* cria diversas árvores de decisão. (5) *Logistic Regression*: É uma técnica estatística que tem como objetivo produzir, um modelo que permita realizar previsões de valores tomados por uma variável categórica, permite estimar a probabilidade associada à ocorrência de determinado evento.

2.4. Mineração de Texto

(1) Para a classificação manual, foi selecionado aleatoriamente uma amostra de 500 *tweets*. O critério para classificação foi: (a) positivas - mensagens que continham apoio às famílias, resistência perante o acontecido, luta, (b) negativas - mensagens de repúdio e/ou tristeza ao acontecido, indignação, ofensas, impressões negativas no geral, (C) neutrais - mensagens impossíveis de identificar. O resultado dessa etapa foi 160 *tweets* positivos,

¹ <https://www.nltk.org/data.html>

236 negativos e 104 neutros. (2) Na etapa de treinamento e teste, foi selecionada 70% dos *tweets* classificados manualmente para treinamento e 30% para testes. Esses dados foram usados para o processo de aprendizagem do algoritmo e verificar sua precisão em relação às mensagens da base de dados. Primeiramente, foi realizada o treinamento dos 5 modelos, onde obteve-se os seguintes resultados:

Tabela 2 – Acurácia (A%), Precisão (P%), Recall (R%), Erro (E%), Cohen Kappa (CK%) e F1-Score (F%) do classificador para as polaridades testadas.

Modelo	A%	P%	F%	R%	CK%	E%
Naive Bayes	78.6%	81.7%	80.1%	78%	65%	55%
Decision Tree	71%	72%	75.1%	71%	55.1%	65%
Support Vector Machine	74%	75.2%	74.2%	74%	59.2%	58%
Random Forest	71%	75.1%	73%	71.4%	56.1%	62%
Logistic Regression	77%	78.9%	78.3%	77%	63.6%	58%

(3) Com base nesses resultados, o método que obteve melhor desempenho nas métricas foi o algoritmo Naive Bayes, apresentando 78.6% de acurácia na classificação, que é o percentual de acertos que ele teve. O resultado obtido é satisfatório considerando que a capacidade de avaliação humana correta da subjetividade de um texto varia de 72% [Wiebe et al 2006] a 85% [Golden 2011].

3 - Por fim, depois da etapa de aprendizado do algoritmo foi realizado a classificação de todos os 32.997 *tweets* contidos na base de dados.

2.5. Resultados

Após fazer o treino do algoritmo fornecendo ao Scikit-Learn² a base separada para treino passados por classificação humana, a base de treino foi utilizada para analisar a acurácia. Realizada a avaliação do método adotado, foi feita a análise de todos os dados presentes na base coletada, dos 32.997 *tweets* presente na base de dados, 18.162 (55.04%) foram classificados negativos, 13.108 (39.72%) como positivos e 1.727 (5.23%) como neutros. A Figura 2 mostra a quantidade de mensagens por dia por polaridade, onde o eixo horizontal apresenta as datas contidas nos *tweets* coletados e o eixo vertical a quantidade.

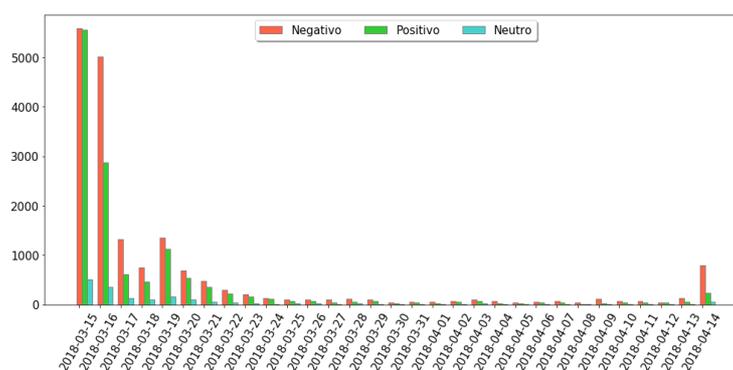


Figura 2 – Classificação dos sentimentos por dia

² <https://scikit-learn.org/>

É possível observar que no segundo dia as mensagens negativas não diminuíram tanto em consideração as positivas, na qual nesse dia notícias divulgaram imagens de câmera mostrando Marielle Franco deixando evento antes do crime e outro carro saindo atrás logo depois. O primeiro pico corresponde a data do dia 15 de março onde foi o primeiro dia após o acontecimento. O outro pico é o do dia 19 de março, onde o PSOL (Partido Socialismo e Liberdade) se mobiliza contra notícias falsas sobre Marielle Franco e o deputado federal Alberto Fraga desativa seu perfil nas redes sociais após publicar uma fake news a respeito da vereadora. O último pico observado foi no dia 14 de abril, na qual completa-se 1 mês do acontecimento.

3. Conclusões e Trabalhos Futuros

Este trabalho teve como objetivo fazer um comparativo entre modelos de classificação e analisar a polaridade expressa nas mensagens dos usuários do Twitter relacionado ao assassinato da vereadora Marielle Franco. O classificador que teve melhor desempenho foi Naive Bayes, ele depende de uma fase de treino supervisionada, na qual para ser realizada uma amostra de *tweets* foi rotulada pelo autor como positivo, negativo e neutro.

Os resultados obtidos mostraram-se satisfatórios considerando os aspectos de classificação por humanos de acordo com o demonstrado na literatura, verificou-se que expressante parte das mensagens foram classificadas como negativas (55,04%), e um número significativo de mensagens positivas (39,72%).

Visando aprofundar as investigações nessa temática, como trabalhos futuros pretende-se criar um conjunto de dados maior, aumentando assim o número de mensagens de treino e teste para comparar os métodos de classificação automática. Por fim, pretende-se realizar análises mais detalhadas, utilizando outras métricas e fazer uso de outros métodos de análise de sentimentos para fazer um comparativo com os resultados obtidos entre os métodos.

Referências

- Golden, P. (2011). Write here, write now. Disponível em: <http://www.research-live.com/features/write-here-writenow/4005303.article>. Acesso em: 09 de Abril de 2019.
- Wiebe, J., Wilson, T., Cardie, C. (2006) “Annotating Expressions of Opinions and Emotions in Language”, *Language Resources and Evaluation*, v. 39, n. 2-3, pp. 165–210.
- de Oliveira, Aletheia Machado. (2016). *Redes Sociais Virtuais, Blog, Wiki e Moocs, como parte de uma arquitetura pedagógica*. Rehutec, Vol 5, No. 1, p12.
- Francisco, Ovídio José. (2018). *Recuperação de informação em atas de reunião utilizando segmentação textual e extração de tópicos*. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de São Carlos.
- Ortigosa, A.; Martín, J. M.; Carro, R. M. (2014). *Sentiment analysis in Facebook and its application to e-learning*. *Computers in Human Behavior*, Vol. 31, p. 527-541.
- Witten, I. H., Frank, E., Hall, M. A., e Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann