EAComp
Encontro Acadêmico de Computação

VI Encontro Acadêmico de Computação - UFMA
Copyright (c) EAComp. ISSN: 2594-486X

Mineração de Opinião Aplicada ao Cenário Político

Leandro Massetti Ribeiro Oliveira[1], **Vandecia Rejane Monteiro Fernandes**[1]

[1] Engenharia da Computação - Universidade Federal do Maranhão (UFMA)
São Luis – MA – Brasil

massetti.leo@gmail.com, vandecia@ecp.ufma.br

Abstract. *Opinion Mining (or sentiment analysis) is a computational analysis of texts found on the web. It arose with the need to classify the opinions automatically obtaining a wide range of results that would not be possible manually. This work aims to make an application of Opinion Mining in the political environment. It intend to classify Twitter's opinions extracted about the mayor's candidates of São Luís with the objective of obtaining the opinion of the voters in relation of each candidate and compare it with the 2016 elections results.*

Resumo. *A mineração de opinião (ou análise de sentimentos) trata da análise computacional de textos encontrados na web. Surgiu com a necessidade de classificar opiniões de maneira automatizada com o objetivo de obter uma grande margem de resultados que manualmente não seria possível. Este trabalho visa fazer uma aplicação da mineração de opinião em um cenário político. Deseja-se classificar opiniões extraídas do Twitter a respeito dos candidatos a prefeitura de São Luís com o objetivo de obter a opinião do eleitores em relação aos candidatos, comparando com o resultado das eleições do ano de 2016.*

1. Introdução

Opiniões sempre estiveram presentes no cotidiano da humanidade como influência das escolhas que fazemos. A Web tem se mostrado um repositório importante de opiniões sobre diversos temas. A cada segundo, milhões de pessoas publicam informações ou opiniões sobre diversos assuntos. Essas opiniões estão dispostas em blogs, fóruns, e, principalmente, em redes sociais.

Conforme Liu (2012), a opinião de outros usuários sobre certo serviço ou produto é um fator importante na tomada de decisões. Hoje, muitas pessoas estão se tornando emissoras de informações, pois ao navegarem pela rede compartilham conhecimentos, críticas e opiniões, baseados na própria experiência. Até certo tempo atrás, o processo de tomada de decisão era baseado em conversas com pessoas conhecidas. Hoje a Web está repleta de sites especializados em revisões de produtos e/ou locais, tornando-se uma grande fonte de coleta de opiniões. Esse grande volume de dados faz a busca de opiniões relevantes um processo trabalhoso e até exaustivo. Para auxiliar esse processo, surgiu a área de estudo chamada de Análise de sentimentos ou Mineração de opinião (MO).

Este trabalho aplicou a mineração de opinião no cenário político. Classificou-se as opiniões extraídas do Twitter a respeito dos candidatos a prefeitura de São Luís visando obter a opinião do eleitores em relação aos candidatos.

 ISSN: 2594-486X

2. Mineração de Opinião

A mineração de opinião, também chamada de análise de sentimentos ou análise de subjetividade, é um estudo computacional das opiniões, atitudes e emoções das pessoas, expressas em forma de texto em relação a uma entidade (filme, livro, local, pessoa, produto). [Medhat et al. 2014].

A MO é uma disciplina relativamente recente que engloba pesquisas de mineração de dados, linguística computacional, recuperação de informações, inteligência artificial, entre outras [Becker and Tumitan 2013].Pode ser resumida como um conjunto de várias técnicas realizadas sobre textos e documentos em geral, para classificá-los de maneira que represente o sentimento do autor do texto [Oliveira 2016].

Através da mineração de opinião, pode-se saber a opinião do público em geral sobre um político, um filme (podendo prever a rentabilidade do mesmo na bilheteria) e até prever o comportamento da bolsa de valores [Becker and Tumitan 2013]. Outra aplicação comum é na revisão de produtos (*reviews*),na qual existe um grande volume de dados e onde há maior interesse das empresas. De acordo com [Pang and Lee 2008], muitas pessoas afirmam que a pesquisa online sobre informações de produtos foram determinantes na decisão de compra.

De acordo com Liu (2012), o foco principal da MO são aquelas opiniões que expressam ou implicam um sentimento positivo ou negativo sobre determinado alvo, para que assim, possa-se classificar sua polaridade. A polaridade de uma opinião refere-se ao sentimento do detentor da opinião em relação a entidade alvo, podendo ser representado como positivo, negativo, ou neutro.

A principal vantagem da classificação do sentimento por polaridade é justamente a sua objetividade, pois o uso de apenas três opções ajuda a obter uma resposta rápida e fácil de sumarizar. Essa técnica, no entanto, não permite obter uma resposta mais profunda quanto ao sentimento dos detentores da opinião [Liu 2012].

O processo da mineração de opinião pode ser dividido em três etapas: identificação, classificação e sumarização.

- **Identificação**: A primeira etapa é a de identificação, que ocorre após a coleta dos textos de suas respectivas fontes (anúncios, jornais, redes sociais, documentos). Nessa etapa é feita a identificação das opiniões contidas neles e a associação dessas opiniões a sua respectiva entidade [Becker and Tumitan 2013]. Na etapa de identificação é onde ocorre também a separação das opiniões de textos sem importância (neutros).
- **Classificação**: A etapa de classificação é a parte principal da MO. É nela que se define o sentimento de uma opinião identificada na etapa anterior e onde se classifica computacionalmente um sentimento em uma opinião. Normalmente essa etapa trata de um problema binário, onde se distingue opiniões positivas de opiniões negativas [Tsytsarau and Palpanas 2012]. Pode-se utilizar níveis de intensidade para compor o sentimento, como palavras com grau de moderadamente positivo ao muito positivo, ampliando ainda mais a precisão da classificação [Becker and Tumitan 2013]. Para a classificação das opiniões são usadas quatro abordagens principais: As abordagens léxicas (ou de dicionário), de aprendizado de máquina, abordagem estatística e abordagem semântica [Oliveira 2016].

- **Sumarização**: Na etapa de sumarização é onde os resultados serão mostrados através do uso de métricas e sumários de maneira que representem o sentimento do público alvo [Becker and Tumitan 2013].

3. Metodologia

Neste trabalho aplicou-se a MO em *tweets* opinativos sobre os candidatos a prefeitura de São Luís no ano de 2016, utilizando a abordagem de aprendizado de máquina para classificar as opiniões quanto ao sentimento dos eleitores.

A abordagem de aprendizado de máquina necessita de uma base de treinamento para poder gerar um modelo de classificação capaz de prever a classe de novas instâncias. Sabendo disso, é necessário a obtenção de um *corpus* com o intuito de servir de modelo para o classificador. Neste trabalho utilizou-se o twitter como fonte de opiniões. Para a coleta dos *tweets*, foi utilizado o *Twitter Search API* e *Twitter Streaming API* através do *Twitter Archiving Google Sheet* (TAGS) desenvolvido por [Hawksey 2014]. Esse *script* coleta automaticamente os *tweets* de acordo com os parâmetros de pesquisa selecionados.

Os *tweets* coletados foram separados em dois grupos. *Corpus A* com os *tweets* do primeiro debate na TV Mirante até o dia anterior ao segundo debate, e o *Corpus B* com apenas os *tweets* do segundo debate. Com esses dados, o objetivo foi utilizar o *Corpus A* para treinar o classificador e analisar o *Corpus B*.

A metodologia utilizada nos *tweets* coletados consiste em fazer uma seleção dos *tweets* (etapa de identificação) que realmente são opiniões, realizar um pré-processamento nos *tweets* selecionados e fazer a extração das características no intuito de montar uma matriz de atributos para o uso de um algoritmo de aprendizado de máquina.

3.1. Seleção dos Tweets

Os *tweets* coletados resultaram em dois *corpus* com muitos textos que não representam uma opinião. Cerca de 78% dos *tweets* provém de blogs de notícias, *tweets* feitos pelos próprios candidatos, textos neutros e resultado de pesquisas. Esses *tweets* que não representam opiniões foram descartados.

Realizou-se então a classificação manual do *Corpus A* e do *Corpus B* quanto ao sentimento de cada opinião. Cada opinião foi associada ao número +1 (representando o sentimento positivo) ou -1 (representando o sentimento negativo). Essa etapa foi necessária para poder utilizar técnicas de aprendizado supervisionado no conjunto de dados.

Realizada a etapa de seleção nos dois *Corpus*, o conjunto de dados ficou distribuído como apresentado na Tabela 1

Tabela 1. Opiniões coletadas.

	Corpus A	*Corpus B*
Positivos	225 *Tweets* - 39,47%	37 *Tweets* - 21,76%
Negativos	345 *Tweets* - 60,53%	133 *Tweets* - 78,24%
Total	570 *Tweets*	170 *Tweets*

3.2. Pré-Processamento

A fase de pré-processamento realizado no texto foi feita utilizando a linguagem de programação Python através da biblioteca de tratamento de texto NLTK (*Natural Language Toolkit*) [Bird et al. 2009] que oferece uma variedade de técnicas de processamento de linguagem natural. Dessa forma, o pré-processamento se deu nos seguintes passos:

- **Padronização dos termos**: Nessa etapa é onde as palavras de cada *tweet* são padronizadas através da eliminação de pontuação, eliminação de acentos, caracteres especiais e todas as letras passam para caixa baixa.
- **Remoção de *StopWords***: Nessa etapa são removidas as palavras que são consideradas irrelevantes, pois não tem sentido próprio quando consideradas sozinhas (como exemplo temos: ”de”, ”o”, ”a”, ”para”, ”com”).
- ***Stemming***: *Stemming* é o processo de reduzir palavras flexionadas ao seu radical (*stem*), dessa forma, nessa etapa as palavras flexionadas (como plural, sufixo temporal, formas de gerúndio) são reduzidas ao seu radical.

Feito o pré-processamento, tem-se uma lista onde cada índice será uma lista com os radicais de cada opinião. Essa lista é necessária para a criação da matriz de atributos.

3.3. Geração da matriz de atributos

Uma matriz de atributos refere-se ao conjunto de vetores de características de cada opinião. Para isso, é necessário guardar as palavras que aparecem em todo o *corpus* de treinamento. Esse conjunto de palavras é denominado *Bag Of Words*.

A *Bag Of Words* deste trabalho contem palavras que apareceram no *corpus* com frequência mínima de 3 vezes. Com isso, a dimensionalidade da matriz de atributos é diminuída. O vetor de características de cada opinião será formado utilizando a técnica TF-IDF [Ramos et al. 2003]. Ao final teremos uma matriz $m \times n$ onde m é a quantidade de opiniões contida no corpus e n o número de palavras na *bag of words*. Após formada a matriz, a mesma foi normalizada para melhora de resultados.

4. Resultados

Os resultados iniciais foram feitos usando o *K-fold cross validation* com $k = 5$ no conjunto inteiro. Os algoritmos tiveram os seus parâmetros de treino estimados a partir da função *GridSearchCV* para obtenção dos melhores resultados. A tabela 2 mostra os resultados iniciais.

Tabela 2. Resultados - *5-fold cross validation*.

Algoritmo	Acurácia	Precisão	Recall	Especificidade	F-measure
Naive Bayes	63.919%	0.49442	**0.78621**	0.55849	0.60683
Árvore de decisão	81.351%	0.74661	0.71742	0.86616	0.73048
SVM	**82.568%**	**0.80109**	0.67554	**0.90794**	**0.73217**

O algoritmo com melhor desempenho foi o SVM, que é um algoritmo que lida muito bem com dados de dimensionalidade muito alta (A *bag of words* teve tamanho médio de 380 palavras). Porém, como pode-se observar, o menor valor do SVM foi a taxa de *recall*, ou seja, dos exemplos que era positivos, o SVM classificou corretamente

apenas 67.554%. O motivo para esse baixo valor pode ter sido o desbalanceamento das classes, que em sua maioria, eram opiniões negativas.

O algoritmo de Árvore de decisão teve um desempenho um pouco menor que o do SVM, porém teve seus valores de *recall* e especificidade mais balanceados, enquanto que o *Naive Bayes* teve um desempenho abaixo do desejável. No geral, o SVM foi melhor devido ao seu *F-measure* ter sido o mais alto.

A próxima classificação foi realizada utilizando o *Corpus A* para treinar os mesmos algoritmos do teste anterior e o *Corpus B* para classificação. Os resultados das classificações podem ser vistos na Tabela 3.

Tabela 3. Resultados usando *Corpus A* em *Corpus B*.

Algoritmo	Acurácia	Precisão	Recall	Especificidade	F-measure
Naive Bayes	51.765%	0.27723	**0.75676**	0.45113	0.40580
Árvore de decisão	72.353%	0.37500	0.40540	0.81203	0.38961
SVM	**73.529%**	**0.41667**	0.54054	**0.78947**	**0.47059**

Percebe-se nesses novos resultados uma diferença significante da Tabela 2, embora novamente o resultado usando o SVM tenha sido o melhor, é possível observar que as taxas tiveram um desempenho menor. Vale ressaltar que, como visto na Tabela 1 a proporção de exemplos no *Corpus A*, utilizado para treino, está com um balanceamento aceitável considerando o teste anterior. O desbalanceamento se encontra agora nos dados de teste (*Corpus B*).

Uma possível causa para a baixa precisão e *recall* pode ser dada ao domínio dos dados. O *Corpus B* contém apenas *tweets* do momento do segundo debate na TV Mirante. Esse contexto próprio pode levar a características que o classificador não obteve com a base de treino e palavras que não existem na *bag of words*.

Além dos problemas citados anteriormente, existem as dificuldades usuais no ramo da MO: negações, gírias, ironias e sarcasmos, que não foram tratadas neste trabalho. De acordo com [Liu 2012], as ironias e sarcasmos estão bastante presentes em opiniões relacionadas a política. Existem também os casos de ambiguidade, que são opiniões que ao mesmo tempo em que demonstram um sentimento negativo quanto a um candidato, exaltam o adversário.

O resultado do algoritmo SVM neste trabalho pode ser considerado satisfatório, pois se compara com os resultados obtidos em outros trabalhos como [Pang et al. 2002], [Sousa 2012] e [Nascimento et al. 2012]. Além disso, classificações de conteúdo subjetivo possuem baixo grau de consenso, um exemplo são as anotações de sentimento feito por humanos que dificilmente tem consenso maior que 75% [Becker and Tumitan 2013].

5. Conclusão

A Mineração de opinião é utilizada para identificação e classificação do conteúdo opinativo criado pelos usuários em redes sociais, classificando opiniões em positivas, negativas ou neutras.

Neste trabalho foram aplicadas técnicas de mineração de opinião a fim de obter uma avaliação do sentimento do público em relação aos candidatos à prefeitura de São

Luís. Para isso foram utilizadas técnicas de pré-processamento e aprendizado de máquina para alcançar o objetivo. Ao contrário do esperado, não foi possível obter um conjunto grande suficiente para analisar estatisticamente a aceitação de cada candidato ou até comparar com o resultado das eleições. No entanto, sua sumarização pode concluir que os usuários do Twitter não estavam satisfeitos com o debate.

Quanto ao desempenho dos classificadores, embora de certa forma tenham sido satisfatórios, apresentaram problemas devido as dificuldades usuais da mineração de opinião, como a presença de gírias, o desbalanceamento dos dados e a ambiguidade. Ao final do presente trabalho conclui-se que as técnicas de aprendizado de máquina são eficientes na classificação de textos. Considerando também que a mineração de opinião ainda tem poucos trabalhos para a língua portuguesa, considera-se este trabalho uma contribuição interessante para a literatura da área.

Referências

Becker, K. and Tumitan, D. (2013). Introdução à mineração de opiniões: Conceitos, aplicações e desafios. *Simpósio Brasileiro de Banco de Dados*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. "O'Reilly Media, Inc.".

Hawksey, M. (2014). Twitter archiving google spreadsheet tags v6. *JISC CETIS MASHe: The Musing of Martin Hawksey (EdTech Explorer)*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Nascimento, P., Aguas, R., Lima, D., Kong, X., Osiek, B., Xexéo, G., and Souza, J. (2012). Análise de sentimento de tweets com foco em notícias. In *Brazilian Workshop on Social Network Analysis and Mining*.

Oliveira, L. M. R. (2016). Um estudo sobre mineração de opinião: Conceitos e abordagens.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Sousa, G. L. S. d. (2012). Tweetmining: análise de opinião contida em textos extraídos do twitter.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.