Análise de Sentimentos em Notícias Utilizando Dicionário Léxico e Aprendizado de Máquina

Ana Carolina Bras Costa[1], **Êmile Cunha Lopes**[1]

[1]Departamento de Informática – Universidade Federal do Maranhão (UFMA)
Av. dos Portugueses, 1966 - Bacanga, São Luís - MA, 65080-805

carolina.bcosta1@gmail.com, emile.clopes@gmail.com

Resumo. *Por meio de textos, postagens em redes sociais as pessoas expressam suas opiniões sobre diversos aspectos, economia, política ou até mesmo um produto que acabou de ser lançado. Identificar os sentimentos por trás dessas palavras têm sido vantajoso para diversas empresas, afim de conhecer melhor seu usuário e a opinião deste sobre o produto que estes oferecem, por exemplo. Entretanto, identificar esses sentimentos requer uma análise computacional para automatizar e poder lidar com uma grande quantidade de informações. Este artigo mostra os resultados de duas abordagens computacionais. A primeira um método léxico e a segunda por aprendizado de máquina utilizando o software WEKA aplicando os classificadores SVM, Naive Bayes e árvore de decisão.*

1. Introdução

Nos últimos anos um novo estudo tem surgido graças ao crescimento das redes sociais. A internet tornou-se um espaço onde as pessoas compartilham suas opiniões, sentimentos e experiências. Esse novo estudo, análise de sentimentos ou mineração de opinião, analisa essas informações compartilhadas para entender as necessidades do usuário e também o nível de satisfação sobre determinado assunto ou produto. De acordo com [Bing 2012], análise de sentimentos é uma área de pesquisa que investiga as opiniões das pessoas para diferentes matérias: produtos, eventos, organizações. Afim de entender o impacto de uma análise de sentimentos, [Asur e Huberman 2010] foi capaz de prever a partir de análise de dados gerados por usuários do Twitter, a quantidade de vendas de ingressos no fim de semana de abertura para filmes com precisão de 97,3%, maior do que o obtido por Hollywood como uma ferramenta de previsão que eles utilização para esse tipo de estimativa.

A análise de sentimento tem como principal objetivo identificar sentimentos e emoções contidos em um determinado texto. Essa análise eleva seu grau de dificuldade de acordo com o número de palavras contida na fonte analisada. Podendo este ser, um tweet, um post no Facebook ou uma notícia, que é o foco deste trabalho. As técnicas atuais, a maioria delas utilizando aprendizado de máquina, determinam o quão positivo ou negativo é um texto. Neste artigo, apresentamos uma abordagem passo-a-passo baseada no artigo publicado [Olga Kolchyna et al 2015] chamado *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*. Na primeira parte implementamos uma abordagem utilizando um dicionário léxico, afim de ser comparado com o desempenho da classificação utilizando aprendizado de máquina.

Este trabalho irá seguir a seguinte seções: Na 2 seção, apresentamos fundamentação teórica, na seção 3 é apresentada a metodologia aplicada na abordagem do dicionário léxico e na de aprendizado de máquina, na seção 4 Resultados e a comparação dos dois métodos, na seção 5 conclusão e trabalhos futuros, por fim referências do artigo.

2. Fundamentação Teórica

2.1. SentiWordNet

SentiWordNet é um recurso léxico para a mineração de opinião. SentiWordNet atribui a cada verbete de WordNet - dicionário online - três contagens de sentimento: positividade, negatividade, neutralidade. Ele se encontra disponível para download gratuito na sua versão mais atual - a 3.0 -. [1]

2.2. WEKA

Waikato Environment for Knowledge Analysis (WEKA) é um software livre e gratuito amplamente utilizado para mineração de dados. O Weka tem como objetivo reunir algoritmos de diferentes paradigmas na sub-área da inteligência artificial referente ao estudo aprendizagem de máquinas. O Weka analisa computacionalmente e estatísticamente os dados fornecidos utilizando técnicas de mineração de dados para indutivamente gerar as soluções esperadas a partir dos padrões encontrados. Ele se encontra disponível para todas as plataformas.[2]

2.3. Classificadores

:

- **Naive Bayes**: Naive Bayes é uma das técnicas de classificação mais utilizadas atualmente. Isto acontece porque um classificador Naive Bayes assume que o valor de um determinado atributo(feature) é independente de todos os outros atributos. A teoria em que o classificador se baseia foi concebida pelo inglês Thomas Bayes no século XVII.
- **Support Vector Machine(SVM)**: Support Vector Machines (SVM's) são classificadores onde dado um um conjunto de treino X cada elemento deste conjunto é um ponto em um espaço euclidiano R que serão separados em positivos e negativos. Esta técnica de classificação foi introduzida por Vapnik e é muito utilizada atualmente especialmente em problemas de análise de sentimento em texto.
- **Árvore de decisão(J48 trees)**: As Árvores de decisão são classificadores onde um conjunto de características é transformado em uma árvore que depois será usada para prever se uma determinada instancia da base de teste pertence a determinada classe. As árvores de decisão têm sido muito utilizadas pelos softwares de mineração de dados, como o que foi utilizado para a execução deste artigo.

3. Metodologia

Este trabalho utilizou a bases de dados em língua Inglesa, composta por manchetes e subtítulos de notícias de jornais do Canadá, EUA e Reino Unido. O conjunto total de

[1]Disponível em: http://sentiwordnet.isti.cnr.it/t

[2]Disponível em: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

notícias foi aproximadamente 180 notícias, divididas em conjunto de dados de treino e conjuntos de dados de teste. Para indicar a classe da notícia ao final do texto era acrescentado o identificador 'pos' ou 'neg'. No caso do aprendizado de máquina, a base de dados foi adaptada para o formato exigido pelo software do WEKA com a extensão .arff.

3.1. Dicionário Léxico

1. **Pre-Processamento**: O algoritmo recebeu um arquivo .txt com a base de dados e essa foi tokenizada - dividida em palavras individuais -, cada palavra foi comparada com um arquivo no formato .txt contendo stop words - conectores, artigos, etc. -, após terem sido eliminados caracteres indesejados (pontuações, números, símbolos) das strings, mantendo só os essenciais à compreensão semântica do texto. Uma vez feito isso, a string composta pelo token é considerada uma palavra. Esse objeto possui atributos como o valor que está em um outro arquivo .txt, SentiWordNet, com o valor negativo e o valor positivo daquela palavra no dicionário de palavras SentiWordNet. Foram utilizadas 151 notícias.
2. **Cálculo de Sentimento**: Um vez que cada palavra já agora possui o valor (positivo e/ou negativo) correspondente a ela no SentiWordNet, é possível calcular o positivoScore da palavra e o seu negativoScore pelos seguintes fórmulas:

$$Score_{AVG} = \frac{1}{m}\sum_{i=1}^{m} W_i.$$

Figure 1. Fórmula total do Score

3. **Classificação**: O score de cada palavra é utilizado para calcular o ScoreTotal da notícia quando mais próximo de 1 mais positiva é a noticía, quanto mais próximo de -1 mais negativa.

3.2. Aprendizado de máquina

Neste trabalho foi utilizado a Maquina de Vetores de Suporte (SVM) como instrumento de classificação, e o programa Weka que contém os principais métodos necessário para a fase de treinamento e classificação. Ela recebe como features (características) uma bag-of-words(vetor de N-grams) e logo após de selecionar as melhores features, treina e classifica o dataset de teste no modelo previamente treinado. No contexto geral da utilização da técnica de aprendizado de máquina, foram executados os seguintes passos:

1. **Pre-Processamento**: Nesta etapa, utilizamos o filtro StringToWordVector do Weka, que é um filtro não-supervisionado para filtrar os dados. Este filtro, como sugere o nome, transforma uma determinada string em um vetor de uni-grams. Isso acontece pois o NGramTokenizer, com o numero mínimo e o numero máximo de 'n' setados em 1(uni-grams), tokeniza a string de entrada de acordo com os outras condições selecionadas. Estas condições foram setadas de acordo com a Figura 2:

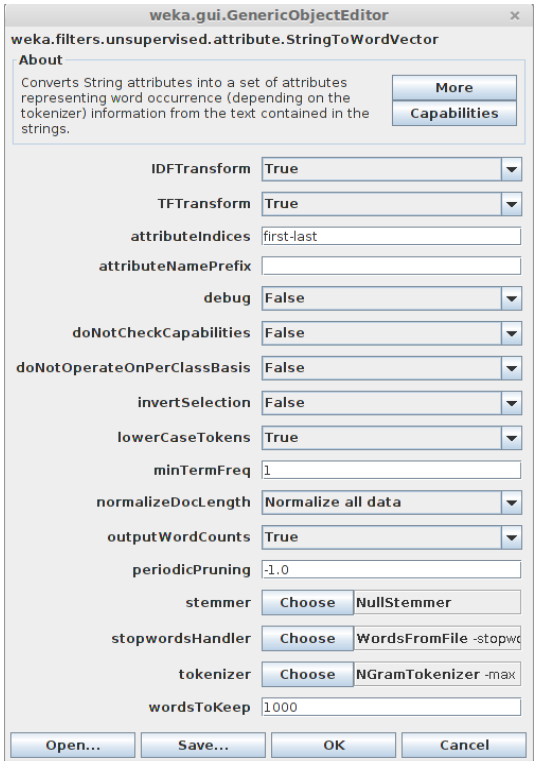

Figure 2. Configuração do filtro StringToWordVector

2. **Seleção de características**: Para fazer a etapa de Seleção de características foi utilizado o filtro supervisionado AttributeSelection, que recebe como entrada um feature evaluator e um método de busca. O feature evaluator (avaliador de características) utilizado foi o InformationGainAttributeEvaluator utilizando o método de busca Ranker. Quando este avaliador é aplicado, ele verifica quais os melhores features a serem utilizadas na etapa de classificação, como mostrado na Figura 3:

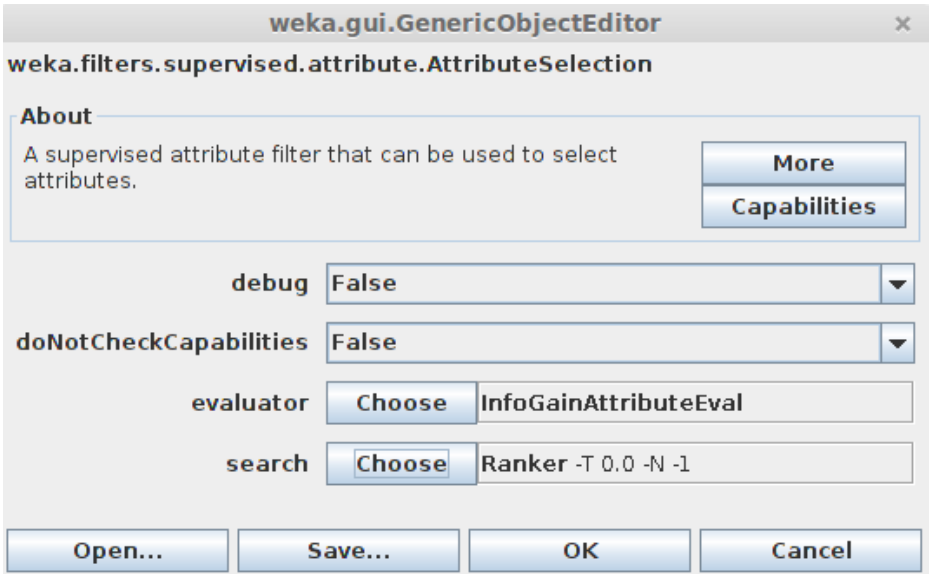

Figure 3. Configuração do filtro AttributeSelection

3. **Treinamento e teste**: Logo depois de selecionadas as melhores características, o conjunto de dados de treino foi usado para treinar e testar três classificadores diferentes: Naive Bayes, Support Vector Machine (que no Weka é representado pelo classificador SMO) e com J48 trees.

Feito isso, o algorítimo exibiu uma descrição completa dos resultados.

4. Resultados

1. **Método Léxico**: Na aplicação do método léxico foi verificado que ele conseguiu identificar 85,59% das notícias corretamente. Entretanto apresentou algumas dificuldades que afetaram diretamente a sua acurácia na classificação das notícias. A matriz de confusão está ilustrada na figura 4. Dentre as dificuldades encontradas, a primeira foi a limitação do dicionário. Ape-

	a	b	
a	12	16	a = pos
b	95	28	b =neg

Figure 4. Matriz de Confusão

sar da grande quantidade de palavras muitas não se encontram lá fazendo apenas algumas palavras terem peso na notícia na hora do cálculo do sentimento. A forma com que ele é organizado faz com que uma mesma palavra apareça no dicionário diversas vezes e com valores diferentes. Isso se dá a análise do contexto que esta palavra está inserida. O problema disso é que como se trata de uma ferramenta automatizada, necessitaria da interferência humana para poder escolher qual melhor score é melhor para aquela palavra. Uma vez que poucos palavras foram identificadas ao invés de dividir pelo número apenas de palavra relevantes aplicamos a divisão por quantidade de palavras total de uma notícia.

2. **Método Aprendizado de Máquina (Machine Learning)**: Depois que o modelo foi treinado e testado, ele precisou ser analisado a fim de determinar a efetividade do mesmo. Quanto aos resultados, este trabalho se baseia no trabalho de [Olga Kolchyna et al 2015] chamado *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*, onde a métrica usada é o F-Score. Porém o F-Score não pode ser visualizado pois a classificação em abos os três classificadores teve que ser encapsulada em um classificador genérico chamado FilteredClassifier como estratégia para contornar o erro de incompatibilidade de atributos. Porém, é possível visualizarmos na Figura 5 como cada algorítimo classificou as 16 mensagens da base de teste:

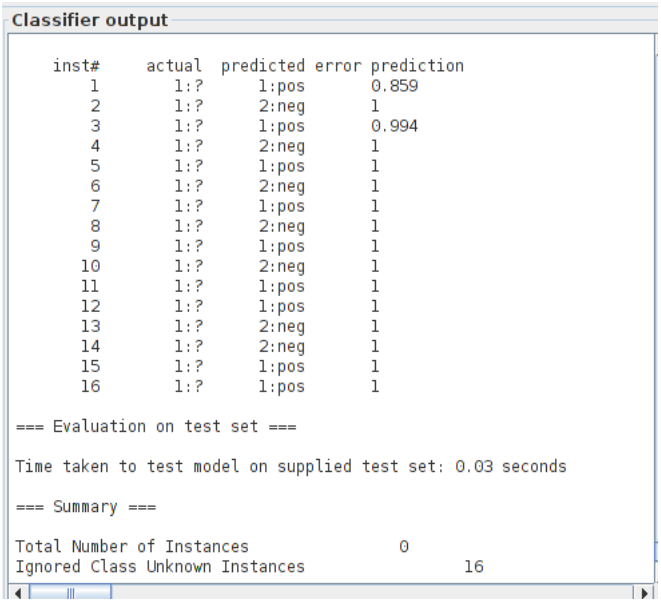

Figure 5. Resultados da classificação utilizando Naive Bayes

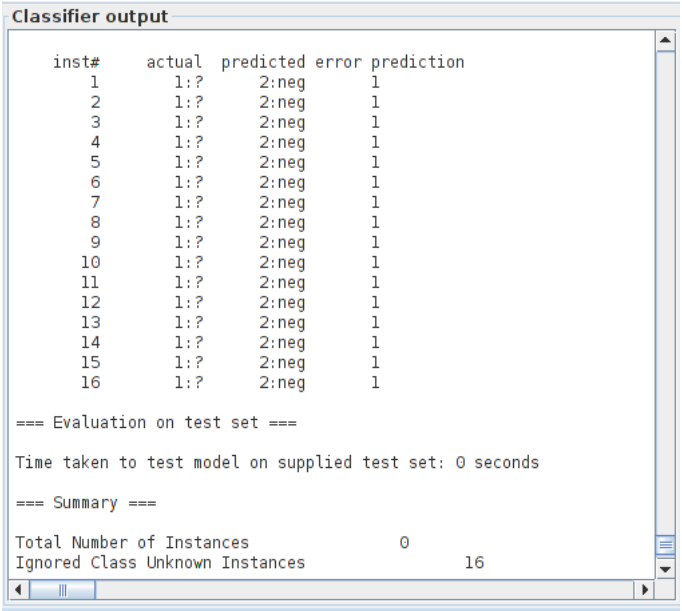

Figure 6. Resultados da classificação utilizando SVM

Como podemos observar, o classificador Naive Bayes foi o que menos se desviou da classificação real de cada notícia.É importante ressaltar também que a base de treino contém aproximadamente 8x mais notícias que a base de teste, o que contribuiu para a acurácia dos testes.

5. Conclusão e Trabalhos Futuros

Com este trabalho foi possível perceber que tanto a técnica do dicionário léxico quanto a técnica de aprendizado de máquina são ferramentas muito eficientes na análise de sentimento de um texto. Como trabalhos futuros é proposto uma combinação das duas técnicas

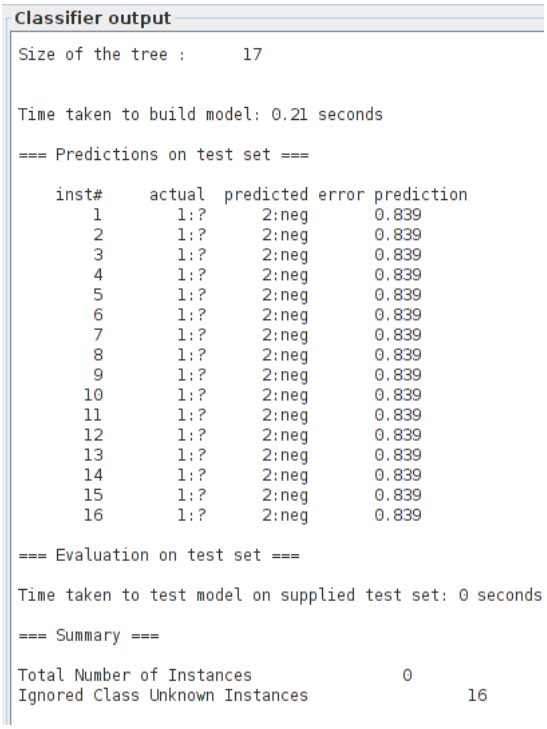

Figure 7. Resultados da classificação utilizando Árvore de decisão(J48)

a nível de software, um aumento da base de teste e de treino, e a adição de uma etapa de clusterização para verificar se há uma melhora nos resultados apresentados neste artigo a fim de que o modelo seja utilizado de forma confiável para analisar o sentimento de um título ou subtítulo de uma notícia.

6. Referências

Bing, L. (2012). *Sentiment analysis: A fascinating problem. In Sentiment Analysis and Opinion Mining*, pag. 7–143. Morgan and Claypool Publishers.

Asur, S. and Huberman, B. A. (2010).*Predicting the future with social media*. In Proceedings of the 2010 IEEE/WIC/A CM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '10, pages 492–499, Washington, DC, USA. IEEE Computer Society.

Olga Kolchyna, Tharsis Souza, Philip Treleaven, and Tomaso Aste.*Twiter Sentiment Analysis*. Computation and Language, 2015.