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

Ana Carolina Bras Costa; Êmile Cunha Lopes
Departamento de Informática
Universidade Federal do Maranhão – UFMA
E-mails: carolina.bcosta1@gmail.com; emile.clopes@gmail.com

INTRODUÇÃO

A internet tornou-se um espaço onde as pessoas compartilham suas opiniões, sentimentos e experiências. A 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.
A análise de sentimento tem como principal objetivo identificar sentimentos e emoções contidos em um determinado texto. A maioria técnicas atuais utilizam aprendizado de máquina para determinar o quão positivo ou negativo é um texto.

OBJETIVOS

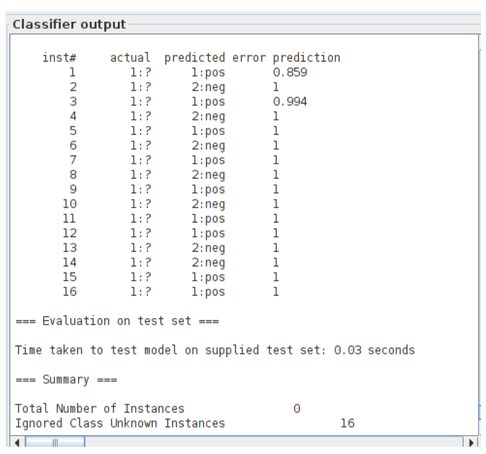
O objetivo deste trabalho é demonstrar a aplicabilidade da análise de sentimento para permitir que a máquina possa realizar processamento de linguagem natural e determinar o sentimento expresso por uma manchete ou subtítulo de uma notícia, para tal é utilizado o classificador SVM (Maquina de Vetores de Suporte) e o software Weka.

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 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.

RESULTADOS

1. Machine Learning: A métrica que deve ser 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. Como podemos observarna figura abaixo, o classificador Naive Bayes foi o que menos se desviou da classificação real de cada notícia.



```
Classifier output

    inst#     actual  predicted error prediction
        1        1:?       1:pos       0.859
        2        1:?       2:neg       1
        3        1:?       1:pos       0.994
        4        1:?       2:neg       1
        5        1:?       1:pos       1
        6        1:?       2:neg       1
        7        1:?       1:pos       1
        8        1:?       2:neg       1
        9        1:?       1:pos       1
       10        1:?       2:neg       1
       11        1:?       1:pos       1
       12        1:?       1:pos       1
       13        1:?       2:neg       1
       14        1:?       2:neg       1
       15        1:?       1:pos       1
       16        1:?       1:pos       1

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Total Number of Instances               0
Ignored Class Unknown Instances                16
```

2. Dicionário 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 abaixo:

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

CONCLUSÃO

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 a nível de software 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.

REFERÊNCIAS

Bing, L. (2012). {\it 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).{\it 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.{\it Twiter Sentiment Analysis}. Computation and Language, 2015.