Extrair Conhecimento em Comentários Gerados em Mídias Sociais Utilizando Análise de Sentimentos

Regenildo G. de Oliveira[2], Halysson Carvalho S. Junior[1], Artur F. da S. Veloso[2], Antônio A. Rodrigues[3], Marcello A. Silva,[1] Davi L. de Oliveira[4], Ricardo de Andrade L. Rabelo[4], Nicole U. Amorim[5], e José V. V. Sobral[6]

[1]Estácio CEUT – PI, Brasil

[2]Faculdade FAETE – PI, Brasil

[3]Faculdade FACID Devry – PI, Brasil

[4]Universidade Federal do Piauí – PI, Brasil

[5]Faculdade Estácio – PI, Brasil

[6]Instituto Federal do Ceará (IFCE), Fortaleza –CE, Brasil

[7]Faculdade Fatepi – PI, Brasil

[8]Instituto Federal do Maranhão (IFMA), São Luis – MA, Brasil

{arturfdasveloso, junioraraujo03, halysson1007, daviluis323, nildo.ngo jf.engtelecom, ayrtonoftitan, victormld}@gmail.com

Abstract. *The internet provide social media tools by which people communicate and influence the social, political and economic behavior of others. In this context, this work shows how the process of Feeling Analysis can obtain the evaluation of people in relation to products through the analysis of texts. The contributions of this article were the creation of the database and generation of the predictive model.*

Resumo. *A internet proporcionar ferramentas de mídias sociais pelas quais as pessoas se comunicam e influenciam o comportamento social, político e econômico de outras pessoas. Neste contexto, este trabalho mostra como o processo de Análise de Sentimentos pode obter a avaliação das pessoas em relação a produtos através da análise de textos. As contribuições deste artigo foram a criação da base de dados e geração do modelo preditivo.*

1. Introdução

A internet desempenha um papel importante no processo de tomada de decisão. as pessoas passaram a compartilhar conhecimento, experiência, críticas e opiniões por meio de mídias sociais, como portais, blogs, fóruns e outros. Com essa estratégia de auxílio na tomada de decisão, surgiu uma nova área de pesquisa chamada de Análise de Sentimentos ou Mineração de Opinião. As organizações passaram a usar análise de sentimentos para obter informações sobre seus negócios, com isso, detectar a avaliação do cliente e também averiguar o que os clientes pensam em relação a seus produtos [D'Adrea et al. 2015].

Partindo dessas considerações, a Seção 2 apresenta o estudo fornecido na literatura sobre análise de sentimentos. A seção 3 fornece uma aplicação de análise de sentimento e seus resultados. Finalmente, a seção 4 conclui o artigo.

2. Análise de Sentimentos

A Análise de Sentimentos é uma técnica que consiste em extrair computacionalmente novas informações, pela extração automática de informação em documentos textuais [Nitin and Damerau 2010]. O objetivo dessa técnica é obter, de forma automática, a polaridade de um texto ou sentença. O processo de análise de sentimento envolve 5 diferentes etapas para analisar dados de sentimento [D'Adrea et al. 2015]. A Fig.1 representa as etapas que envolvem análise de sentimentos.

Figura 1. Etapas de Análise de Sentimentos

- Coleta de dados: o primeiro passo da análise de sentimentos consiste em coletar dados gerado pelo usuário.
- Preparação de texto: melhoria na qualidade dados coletados.
- Detecção de sentimento: associa os termos da sentença a uma determinada polaridade.
- Classificação do sentimento: é realizado a conexão dos dados com o algoritmo que irá aprender com os dados históricos.
- Apresentação do resultado: converter texto não estruturado em informação significativa.

2.1. Processamentos de Linguagem Natural (PLN)

A análise de sentimentos faz uso de técnicas de PLN que segundo [Liddy 2001] "é um conjunto de técnicas computacionais para analisar e representar ocorrências naturais de texto em um ou mais níveis de análise linguística". Tokenização que é a divisão do texto em partes menores. Cada parte é chamada de token, e que de acordo com [Vinodhini and Chandrasekaran 2012], corresponde a um termo do texto formado por um grupo de caracteres. Por exemplo, a frase:

['o'.'notebook','e','bom','porem'.'nao','gosto','notebook','grandes']

Outra técnica utilizada é a remoção de *Stopwords*, São palavras que mais aparecem no texto sendo consideradas irrelevantes par o processo. Esse grupo compreende as classes gramaticais como artigos, preposições, conjunções, pronomes e alguns verbos que não necessitam ser indexados por possuírem frequência elevada [Bird et al. 2009]. A fig.2 demonstra esse procedimento.

Stemming é a técnica que realiza a redução de léxico agrupando tokens que compartilham de um mesmo padrão. Onde as palavras: gosta, gosto, gostando ficam respectivamente a "gost", "gost", "gostand". [Carvalho Filho 2014] utilizou as técnicas para melhorar o classificação de texto.

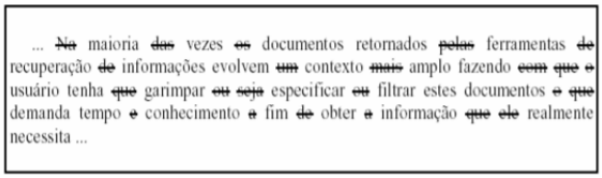

Figura 2. Remoção de Stopwords [Aranha 2007]

Conjuntamente pode utilizar a técnica denominada de *N-gram*: que pode usar *uni-gram, bi-gram, tri-gram* ou combinação destes para classificação de sentimentos. Na fig. 3 demonstra o par de palavras, onde cada termo é concatenado com o termo posterior [Kaur and Gupta 2013].

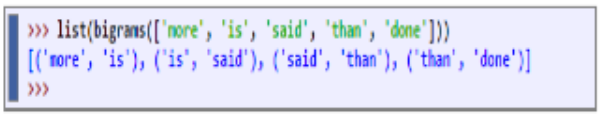

Figura 3. Técnica N-gram [Bird et al. 2009]

2.2. Abordagem de Classificação de Sentimentos

A abordagem de aprendizado de máquina supervisionada é o método usado para prever a polaridade de sentimentos [Hailong et al. 2014] com base no sentimento expresso em casa frase. Isto é, utiliza-se de todas as características da base de dados, a frequência de cada característica e associa isso a classe do treino (modelagem *Bag of words*) [Ali et al. 2016]. A principal vantagem deste método é a capacidade de adaptar e criar modelos treinados (classificadores) para fins e contextos específicos, sua principal desvantagem é que necessita de dados rotulados.

2.3. Naïve Bayes

Na etapa de classificação de sentimentos é utilizado o algoritmo de *Naïve Bayes*. Esse algoritmo tem por base o teorema de *Bayes* e utiliza dados de treino para formar um modelo probabilístico baseado na evidência das características no dado

$$P(A|B) = \frac{P((B|A)P(A))}{P(B)} \quad (1)$$

onde $P(A|B)$ é a probabilidade *aposteriori* (posterior), a frequência da palavra dada é B. $P(A)$ é a probabilidade da classe A e $P(B)$ é a probabilidade da palavra B.

Para cálcular $P(B|A)$ a probabilidade condicional , é necessário um conjunto de dados treinados que já foram classificados. Dessa forma, pode-se calcular a probabilidade de um determinado evento de acordo com sua frequência. Mas já sabendo qual foi o evento que o antecedeu. Isto é, avalia-se a probabilidade de um registro ter uma polaridade específica, ou a probabilidade de um evento acontecer de acordo como o especialista classificou os registros [Kaur et al. 2017].

3. Experimento e Resultados

Os textos que serviram de base para criação do corpus deste estudo foram comentários de portais, blogs e fóruns relacionados a marcas de notebooks. A avaliação dos comentários, leva em conta os sentimentos, opiniões, avaliações e emoções das pessoas em relação a performance, preço eventuais problemas. Neste estudo, os textos são classificados de acordo com o seguinte critério: a polaridade do sentimento expresso (em positivo, negativo e neutro).

A coleta dos comentários foi a primeira parte da execução deste trabalho. Para esta etapa, utilizou-se do *crawler* (WebHarvy) coletor de dados web. Os caracteres especiais que poderiam prejudicar a análise foram removidos. Foram coletados 1660 comentários, destes, 705 comentários rotulados como positivos, 640 como negativos e 315 como neutros. Após a adequação dos comentários na etapa de coleta aplicou-se as técnicas de PLN, tokenização, *stopwords* e *stemming* e *n-gram*. Como visto na seção 2.

3.1. Gerando o Modelo

Em seguida foi gerado o modelo usando a abordagem *Bag of Words* e o algoritmo *Naïve Bayes*. Como visto nas seções 2.2 e 2.3, essa modelagem usa a frequência das características para treinar um modelo. Por exemplo, imagine as seguintes sentenças:

Sentença 1: O notebook é bom, porém não gosto de notebook grandes, Negativo.

Sentença 2: Gostei bastante desse notebook, mesmo sendo caro, Positivo.

O modelo nesse caso, usa todos os termos (características) da base, ou seja, todos os termos das duas sentenças: {o, notebook, é, bom, porém, não, gosto, de, grandes, gostei, bastante, desse, mesmo, sendo, caro}

Em seguida, é feita a contagem do número de vezes que cada característica apareceu na sentença. Para representar a Sentença 1, o termo "notebook" ficaria com a contagem 2, pois este apareceu duas vezes na frase. As outras características ficaram com a contagem 1 e 0 quando essas não foram encontradas na sentença.

{1,2,1,1,1,1,1,1,1,0,0,0,0,0,0, Negativo}

A sentença 2 seria representada da seguinte forma:

{0,1,0,0,0,0,0,0,0, 0, 1,1,1,1,1,1,1,1, Positivo}

Agora, essa pontuação pode ser integrada ao classificador *Naïve Bayes* para calcular a probabilidade posterior, de modo a verificar quais palavras tenham maior probabilidade ao determinar o sentimento geral do comentário [Kaur et al. 2017].

3.2. Treinamento e Teste

A validação do modelo foi feito pela técnica *Cross Validation*, ou validação cruzada. Que consiste em dividir toda a base de dado em k subamostras, que serão chamadas de folds. Dessas k subamostras, uma será separada para ser utilizada na validação (conjunto de teste) e as k-1 subamostras restantes serão usadas para treinar o modelo (conjunto de treinamento). O resultado final é média do desempenho do classificador nas k interações. Esse processo é repetido até que o modelo seja treinado e testado com todas as partes do dado. Com essa estratégia, evitamos problemas de variância nos dados [Vasinek et al. 2016].

3.3. Performance do Modelo

Apresentação dos resultados é a última etapa do processo de análise de sentimentos [D'Adrea et al. 2015]. É nessa fase que os autores do estudo fizeram a análise do resultado e verificaram se o classificador atingiu o resultado esperado. Em seguida, é validado performance do classificador através da mátrica acurácia do modelo, que é, basicamente o percentual de acertos que o mesmo teve.

$$Accuracy = \frac{(TP+TN+TNE)}{(TP+FP+TN+FN+TNE+FNE)}$$

A termologia (TP), (TN) e (TNE) da equação 2 significa classificação correta das classes positivo, negativo e neutro. E (FP), (FN) e (FNE) significa classificação errada das classes respectivamente [Ali et al. 2016].

Os resultados de classificação das polaridades positivo, negativo e neutro estão apresentados na Tabela 1. Chamada de matriz de confusão [Ali et al. 2016], também conhecida como tabela de contingência, esta é uma matriz simples que apresenta os resultados de um classificador estatístico, dessa forma, permite a visualização do desempenho de um algoritmo [Kaur et al. 2017].

Tabela 1. Resultados de Classificação

Preditivo	Positivo	Negativo	Neutro	Total
Positivo	566	97	42	705
Negativo	147	438	55	640
Neutro	52	72	192	315
Total	765	607	288	1660

$$Accuracy = \frac{1196}{1661} X 100\% = 72,00\%$$

O modelo classificador mostrou ser eficiente de modo geral, apresentado um bom percentual de 72%. No entanto, existiram problemas na classificação onde uma característica que obteve maior probabilidade (ou relevância), estava presente nas polaridades positivo, negativo e neutro dificultando a avaliação do modelo classificador. Por conseguinte houve muitos erros que podem ser observados na Tabela 1, ou seja, os Falsos positivos, negativos e neutros (FP, FN e FNE respectivamente).

4. Conclusão e trabalhos futuros

Neste artigo, foi apresentado o processo de Análise de Sentimentos para gerar um modelo preditivo que prever a polaridade de sentimentos com base em conjuntos de dados treinados e de teste. O corpus deste trabalho foi criado a partir avaliação de comentários coletados em portais, blogs e fóruns relacionados a marcas de notebooks. Para este fim, foi utilizado a abordagem de aprendizado de máquina supervisionada que correlaciona diversas áreas como: Processamento de Linguagem Natural (PLN), modelagem de dados *Bag of words*, classificador probabilístico *Naïve Bayes* e entre outras. Além disso,

a eficácia do resultado é avaliada com um percentual ótimo de acurácia no cenário de treinamento e teste.

Um desafio futuro na aplicação do modelo de classificação implementado neste trabalho é a escolha de outros algoritmos classificadores como KNN, Árvores de Decisão, Máquina de Vetor e Suporte e Redes Neurais. Assim, permita extrair conhecimento útil e tenha excelente desempenho.

Referências

Ali, N. M., Jun, S. W., Karis, M. S., Ghazaly, M. M., and Aras, M. S. M. (2016). Object classification and recognition using bag-of-words (bow) model. In *Signal Processing & Its Applications (CSPA), 2016 IEEE 12th International Colloquium on*, pages 216–220. IEEE.

Aranha, C. N. (2007). *Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional*. PhD thesis, PUC-Rio.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Carvalho Filho, J. A. (2014). Mineração de textos: Análise de sentimento utilizando tweets referentes à copa do mundo 2014.

D'Adrea, A., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. In *International Journal of Computer Applications*, page 0975 – 8887. IJCA.

Hailong, Z., Wenyan, G., and Bo, J. (2014). [ieee 2014 11th web information system and application conference (wisa) - tianjin, china (2014.9.12-2014.9.14)] 2014 11th web information system and application conference - machine learning and lexicon based methods for sentiment classification: A survey.

Kaur, A. and Gupta, V. (2013). A survey on sentiment analysis and opinion mining techniques. In *Journal of Emerging Technologies in Web Intelligence*, pages 367–371. JETWI.

Kaur, H., Mangat, V., et al. (2017). A survey of sentiment analysis techniques. In *I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017 International Conference on*, pages 921–925. IEEE.

Liddy, E. (2001). Natural language processing. In *In Encyclopedia of Library and Information Science*, New York: Marcel Decker.

Nitin, I. and Damerau, F. J. (2010). Handbook of natural language processing. ISBN 9781420085921.

Vasinek, M., Plato, J., and Snasel, V. (2016). Limitations on low variance k-fold cross validation in learning set of rules inducers. In *Intelligent Networking and Collaborative Systems (INCoS), 2016 International Conference on*, pages 207–214. IEEE.

Vinodhini, G. and Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining. In *International Journal 2.6*.