PAPER • OPEN ACCESS

Sentiment analysis of feature ranking methods for classification accuracy

To cite this article: Shashank Joseph *et al* 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **263** 042011

View the article online for updates and enhancements.

Related content

- Aspect level sentiment analysis using machine learning
  D Shubham, P Mithil, Meesala Shobharani et al.

- Sentiment Analysis on Textual Reviews
  Mirsa Karim and Smija Das

- Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques
  Redouane Karsi, Mounia Zaim and Jamila El Alami

Sentiment analysis of feature ranking methods for classification accuracy

Shashank Joseph, Calvin Mugauri, Sumathy S[*]

School of Information Technology and Engineering, VIT University, Vellore-632014, India.

*Email:ssumathy@vit.ac.in

Abstract. Text pre-processing and feature selection are important and critical steps in text mining. Text pre-processing of large volumes of datasets is a difficult task as unstructured raw data is converted into structured format. Traditional methods of processing and weighing took much time and were less accurate. To overcome this challenge, feature ranking techniques have been devised. A feature set from text preprocessing is fed as input for feature selection. Feature selection helps improve text classification accuracy. Of the three feature selection categories available, the filter category will be the focus. Five feature ranking methods namely: document frequency, standard deviation information gain, CHI-SQUARE, and weighted-log likelihood –ratio is analyzed.

Keywords: datasets, text pre-processing, feature selection, classification accuracy, feature ranking

1. Introduction

The web stores vast amount of content generated by the users. The content describes customer opinions on products and services through reviews, tweets, blog, etc. Customers are able to make purchasing decisions basing on reviews [1]. That is, before a customer makes a buying decision, he /she can visit the related website and browse online reviews of alternative products. The reviews also help manufactures to make an improvement on products or services. Therefore, accurate understanding of expressed sentiments can help give business a competitive advantage over competitors [2].

Feature weighting is an important stage in sentiment analysis as it passes data as input for sentiment classification. Basing on the feature ranking or weighting method, accuracy is improved for classification. That is, the smallest subset of features is found that maximally improves the performance of the model. [3] Prior to feature ranking is pre-processing of the dataset. Text pre-processing is a vital part of any Natural Language Processing system as characters, words, and sentences identified are crucial units that are passed further for processing. Text preprocessing starts with data filtering followed by data cleansing. The resultant file is exported for feature weighting. In feature weighting, there are three categories of feature selection namely filter methods, wrapper methods and embedded methods. Working directly on the dataset, filter methods are able to provide a weighting, ranking or subset as output. Wrapper methods, guided by the outcome of the system they search in the space for features. And by the use of internal information of the classification model, embedded method performs feature selection. This paper is based on filter methods which provide a weightage or ranking on the features. Five ranking or feature selection methods are analyzed.

2. Literature Review

Many researches have been done to achieve best results performing sentiment analysis in multilingual context. The text classification is based on the feature ranking. But the feature ranking has two issues as accuracy on categorization and identification of features. To resolve these issues, feature selection is retrieving the important features from the data set before

classification of data set. If the results of feature selection are robust, specific and reliable then classification process will be improved. If the number of features increases then it will affect the complexity leading to decreased accuracy of classification. The feature selection process plays essential role for classification issues to increase the computation and accuracy. [4]
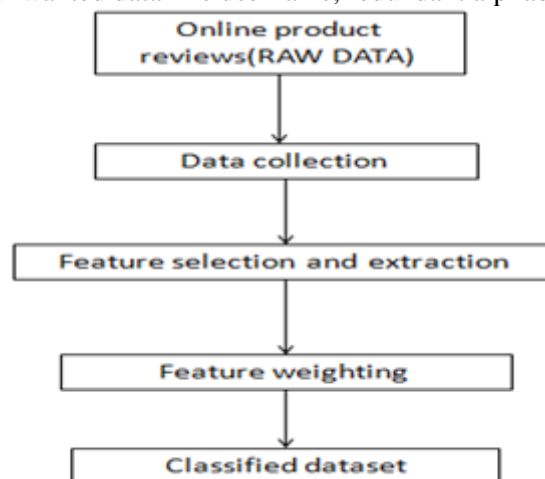
The text classification machine workflow consist the preprocessing methods, feature extraction techniques, feature selection and classification of data set. Brief information about the text classification workflow stages (Fig.1) are given below.

2.1 Data Collection

The data collection form the online websites is the raw dataset for sentiment analysis. The online website could be like http://www.amazon.in, http://www.flipkart.in, http://www1.ap.dell.com/content/default.aspx?c=in&l=en&s=&s=gen&~ck=cr, etc. The collected raw data has been arranged as per preprocessing requirements [5].

2.2 Preprocessing Methods

The collected datasets first of all goes for preprocessing where the datasets are processed with filter data, Data Cleaning and Extract to text file. The filter data is a process where raw data within keyword is filtered using the name of a product. i.e. 'Hp laptop', 'iphone', 'Cisco wireless router' etc. Data cleansing is the process of removing the noisy data from the raw data, like @, #... all special characters and unwanted data like username, redundant alphabets, URL etc.

Fig. 1 Workflow for Text Classification of online product review

2.3 Feature selection and extraction

Feature selection and extraction is the process where select the preprocessed datasets and transform into the specific set. The text classification process is based on the featured datasets; hence it is most important to select the extract features. The main objective of feature extraction is to transform data into well represented format based on their feature. [6] There are some steps to obtain the feature selection and extraction:

2.3.1   Case Normalization

Mostly the textual review contains combination of upper case and lower case character, which creates the major problem for the feature selection. To overcome from this issue case normalization converts the all character into the lowercase. [18, 19]

2.3.2   Tokenization

Tokenization is the process where separation occurs on the dataset, then assign the distinct tokens for each feature. The datasets which is based on the English language, punctuation and blank space are the delimiters for tokenization [18].

2.3.3   Stemming process

Stemming is process to reduce the distinct type of tokens and convert into appropriate token type. Stemming works based on the prefixes, pluralization and suffixes of the tokenized dataset.[17]

2.3.4   Generate N-Grams

The n-grams are based on the number of character present on the particular stemmed sets. For example take trigram (3-gram) of phrases as "NICE" could be '_ _ N','_ N I','NIC', 'C E _','E _ _'. The N-grams of phrases has dimensions as 1 dimensional, 2 dimensional, 3 dimensional and they called as 'unigrams', 'bigrams', 'trigrams' respectively. The N-Grams are mostly using for the speech pattern recognition [14,15] and identify the particular language. Also text classification requires the N-Grams for effective classification [16].

2.4 Feature weighting

The feature weighting process providing the weightage to individual feature based on the frequency of the feature. Hence performance of the feature classification process will increase. There are two approaches for feature weighting:

2.4.1   Term Frequency

Term frequency observes the frequency of single term on the set of features, while dataset is having regularity based on the set of feature for specific term. The dataset regularity is depending on the respective binary value of the term. We can calculate term frequency with the formula where TF(t) is term frequency function, t is number of terms, $TF(t, o_i)$ is percentage of datasets belonging to a class o_i in which term t occur. And |c| is number of available categories[7,9].

$$TF(t) = \sum_{i=0}^{|b|} TF(t, o_i) = \sum_{i=0}^{|b|} R(t|o_i) \quad (1)$$

2.4.2   Term Frequency Inverse Document Frequency (TF-IDF)

The term frequency is more valuable technique then term frequency. TF-IDF is based on the term frequency to collect the frequency weightage of particular term for the modeling process. The inverse document frequency is based on the number of documents has the particular phrase or term in the recommend dataset. To calculate the TF-IDF there is a formula below[8,9].

$$TF - IDF = f(t) \times log\frac{P_r}{p_r} \quad (2)$$

The variables are associated with t is term to weight, f(t) is frequency of term to weight in the dataset of data collection, D_r A collection of data that recommended to observer, P_r Number of dataset in D_r, p_r number of dataset in D_r that contain t.

2.4.3   Chi-square

Chi-square is normalized value, which represents the degree of relationship heterogeneous categories and features. Calculate the value of Chi-square as,

$$CHI(S, t_i) = \frac{D \times (LM - NO)}{(L+N) \times (L+O) \times (M+N) \times (M+O)} \quad (3)$$

Where S is feature, t_i is category, D is number of datasets in the collected text, n is number of categories, L is number of times S and t_i occur, M is number of time neither t_i nor S occurs, N is number of times S occurs without t_i and O is the number times 't_i occurs without S. [9, 10]

2.4.4   Information gain

Information gain is parameter. It contains the number of bits of information that gained prophecy of classification based on the features availability in the dataset [9,12,13]. The information gain defines as :

$$IG(S) = -\sum_{i=1}^{n} P(t_i) \times logP(t_i) + P(S) \times \sum_{i=1}^{n} P(t_i|S) \times logP(t_i|S) + P(\bar{S}) \times \sum_{i=1}^{n} P(t_i|\bar{S}) \times logP(t_i|\bar{S}) \quad (4)$$

Where S is feature, P(S) is probability that feature S occurs, $P(\bar{S})$ is probability that feature S does not occurs and $P(t_i)$ is probability that class t_i occurs.

2.4.5 *Standard deviation*

Standard deviation is a method which calculates the feature distribution from mean of feature space. The heavy standard deviation presents features that are extensive over the huge rang of values. And the light standard deviation provides the feature points which are closer to the mean value [11]. The standard deviation defines as.

$$mean_j(S_k) = \frac{1}{M}\sum_{i=1}^{M} Y_{i\,k} \quad (5)$$

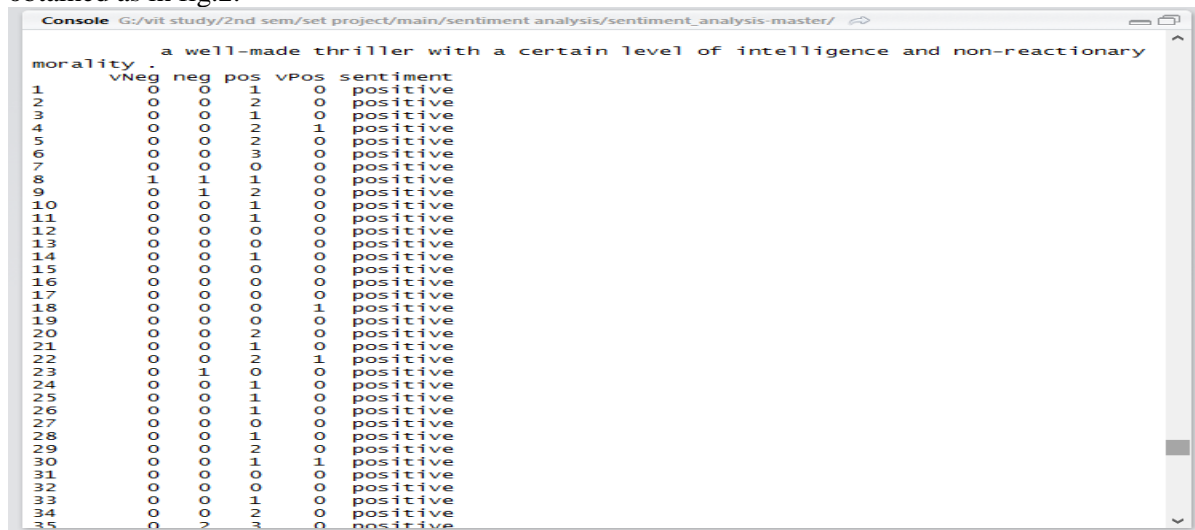$$stdDv_j(S_k) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(Y_{i\,k} - mean_j(S_k))^2} \qquad (6)$$

$$j = 1, 2\dots J \text{ and } k = 1, 2\dots n$$

$$SD(S_k) = |stdDv_1(S_k) - stdDv_2(S_k)| \qquad (7)$$

Where S is feature, M is number of sample in a class, n is number of feature available, J is number of categories, $Y_{i\,k}$ signifies weight of kth feature which is based on the TF-IDF to i^{th} sample, $mean_j$ and $stdDv_j$ are mean and standard deviation respectively[9,11].

3.   Results

AFFIN-96 and AFFIN-111 datasets are used for feature ranking classification. With the help of RStudio the results as the classification like very negative, negative, positive and very positive are obtained as in fig.2.

Fig 2. Classification of data

4. Conclusion

Text mining is the process of extracting meaningful and useful information from unstructured data. It figures out some interesting patterns from very large databases. Different techniques are used in order to come out with a feature set useful for feature ranking.Different feature ranking techniques have been highlighted in this paper which may provide an insight for text mining researchers.

References

[1]   Zhao, Yanyan, Bing Qin, and Ting Liu. 2015 "Creating a fine-grained corpus for Chinese sentiment analysis." *IEEE Intelligent Systems* 30.1 (2015): 36-43.

[2]   Lizhen, Liu, et al. 2014 "A novel feature-based method for sentiment analysis of Chinese product   reviews." *China Communications* 11.3 (2014): 154-164.

[3]   Saeys, Yvan, Thomas Abeel, and Yves Van de Peer. 2008 "Robust feature selection using ensemble feature selection techniques." *Machine learning and knowledge discovery in databases* :313-325.

[4]   Ciurumelea, Adelina, et al. 2017 "Analyzing reviews and code of mobile apps for better release planning." *Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on*. IEEE.

[5]   Parkhe, Viraj, and BhaskarBiswas. 2014 "Aspect based sentiment analysis of movie reviews: finding the polarity directing aspects." *Soft Computing and Machine Intelligence (ISCMI), International Conference on*.IEEE.

[6]   Zainuddin, Nurulhuda, and Ali Selamat. 2014 "Sentiment analysis using support vector machine." *Computer, Communications, and Control Technology (I4CT), 2014 International Conference on*.IEEE.

[7]   Chen, Yifei, Bingqing Han, and Ping Hou. 2014 "New feature selection methods based on context similarity for text categorization." *Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on*.IEEE.

[8]   Beel, Joeran, Stefan Langer, and BelaGipp.March 2017 "TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' PersonalDocumentollections." *Proceedings of the 12th iConference.To appear in March 2017*.

[9]   Yousefpour, Alireza, Roliana Ibrahim, and HazaNuzly Abdel Hamed. "Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis." *Expert Systems with Applications* 75 (2017): 80-93.
[10]   Kiliç, Selim. 2016 "Chi-square Test." *Journal of Mood Disorders* 6.3 (2016): 180.
[11]   Leys, Christophe, et al.2013 "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median." *Journal of Experimental Social Psychology* 49.4 (2013): 764-766.
[12]   Lee, Changki, and Gary Geunbae Lee. 2006 "Information gain and divergence-based feature selection for machine learning-based text categorization." *Information processing & management* 42.1 (2006): 155-165.
[13]   Wen, Ping-Ping, et al. 2016 "Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization." *Bioinformatics* 32.20 (2016): 3107-3115.
[14]   Bahdanau, Dzmitry, et al. 2016 "End-to-end attention-based large vocabulary speech recognition." *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*.IEEE, 2016.
[15]   Zinchenko, Kateryna, Chien-Yu Wu, and Kai-Tai Song. 2016 "A Study on Speech Recognition Control for a Surgical Robot." *IEEE Transactions on Industrial Informatics* (2016).
[16]   Gadag, Ashwini I., and B. M. Sagar. 2016 "N-gram based paraphrase generator from large text document." *Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on*. IEEE, 2016.
[17]   Ismailov, A., et al. 2016 "A comparative study of stemming algorithms for use with the Uzbek language." *Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on*.IEEE, 2016.
[18]   Titus, Nikhil George, Tinto AntoAlapatt, and NiranjanRao. 2016 "A Case Study on the Different Algorithms used for Sentiment Analysis." *International Journal of Computer Applications* 138.12 (2016).
[19]   Cheng, Zhijin, et al. 2016 "Case studies of fault diagnosis and energy saving in buildings using data mining techniques." *Automation Science and Engineering (CASE), 2016 IEEE International Conference on*.IEEE, 2016.