www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 2 February 2018, Page No. 23653-23655
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i2.22

A Performance Evaluation of Attribute Extraction Technique for the Analysis of Sentiment

G.Divyaprathi[1], Prof.S.Kuppuswami[2]

[1]PG Scholar, Department of Computer Science &Engineering,Kongu Engineering College, Erode, TamilNadu,
[2]Principal, Kongu Engineering College, Erode,TamilNadu,

Abstract:
Sentiment Analysis is the process of computation identifying and categories the thought expressed by people in order to determine whether the writer's attitude towards a particular product is positive or negative. The results of this analysis can be used in computing customer satisfaction metrics, marketing, contextual advertising, suggestion systems based on the user likes and rating, recommendation systems etc. It can be done on reviews, blogs, tweets, forums etc. Term Weighting is the process of assigning weights to the terms in the documents. Different term weighting schemes like Term Frequency, Term Frequency – Inverse Document Frequency and Binary Weighting methods are used on the reviews to weigh the sentiments words present in the reviews. It is done considering the terms as unigrams, bigrams, trigrams and Syntactic n-grams. The huge data presented to extract sentiments are unstructured in nature and they require processing like classification or clustering to get some meaningful information out of it. Supervised machine learning approaches like Naïve Bayes are used to classify the human sentiments present in the reviews as positive and negative.

Keywords: Sentiment analysis, Feature extraction, Naïve Bayes.

INTRODUCTION

Opinions are central to almost all human action because they are the key impact of our demeanour. Whenever we need to make a decision, we want to know others opinions.

In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. individual client also want to know the sentiment of existing users of a product before leverage it, and others sentiment about political nominee before making a voting decision in a political election. In the past, when an individual needed opinions, he/she approached friends and family. When an establishment or a business needed public or customer opinions, it conducted surveys, opinion polls and focus groups. Getting public and customer opinions has long been a huge business itself for marketing, public relations and political campaign companies. Analysing such collected reviews and mining certain patterns from the text is in prime focus that lead to the growth of sentiment analysis.

OVERVIEW

The workflow of the proposed work is given in the figure 1. The first step involves the collection of reviews. Electronic Multi Domain Review dataset is used. The reviews are pre-processed to remove the stop words, punctuations, numerals and are converted into a matrix format using various weighting methods. This matrix is passed as the input to the Naïve Bayes classifiers along with their class labels.

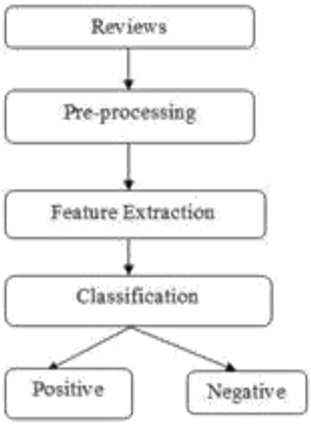

Figure 1: Workflow Diagram

A.PREPROCESSING

If they have irrelevant and excess information present or noisy and undependable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, Instance selection, normalization, transformation and feature extraction .

B.FEATURE EXTRACTION

a)Unigram: Unigrams are features of single words similar to the Bag-of-words feature. Unigrams are simple to create and has been used in most sentiment analysis task as mentioned in the related works section. Every word in a text documents is considered as a feature. Unigrams can be easily formed by tokenization where a text documents is read as a string of words with spaces in between. The tokenize will separate the each word in the string based on the blank.

b) Bigram: Bigrams are part of N-gram features where a feature is made up of two words in this work. Unigrams disregard word structure and sequence, causing the concern that semantics of the sentence is distorted. Bigram is a way of preserving a little of the sentence structure. They are formed by pairing off each word with its left and right neighbour. So each word forms two bigrams.

c)Trigram: Trigrams work similarly as bigrams but they are made up of three words instead . Each word is paired off with both neighbouring word to form a trigram. Trigrams are phrase features that preserve even more of the sentence structure and sequence. Collocations and phrases that express stronger sentiment can be easily captured with trigram. Although trigrams are less popular than unigrams and bigrams, it does show potential from some literatures.

d)Syntactic Bigram: The result of this one-to-one correspondence dependency grammars are word grammars. A bigram is a succession of two neighboring elements from a string of nominal, which are typically letters, sentences, or words.

e) Syntactic Trigram: A Trigram is a sequence of three adjacent elements from a string of tokens, which are typically letters, sentences, or words.

C. TERM WEIGHTING

Term weighting is the assignment of numerical values to terms that represent their importance in a document in order to improve effectiveness. It considers the relative importance of individual words in a sentiment analysis system, which can improve system effectiveness, since not all the terms in a given document collection are of equal importance. Different types of term weighting schemes that are used in this sentiment analysis system are,

a)Binary Weighting: Binary weighting is the process of representing the occurrences of terms in the document with the help of either 0 or 1. When the term is present in the document, it is given a term weight of 1 and when the term is absent in the document, it is given a term weight of value 0. In this way, a document term matrix is built using BIN scheme.

b) Term Frequency: Term Frequency weighting scheme takes the number of occurrences of a particular term in a document. In this scheme, the total number of occurrences of a particular term in the document is counted and the occurrence count is used as the term weights. Term Frequency evaluate how frequently a term occurs in a text file.

c)TF-IDF: The TF-IDF weight is composed of two terms: the first work out the normalisation Term Frequency which is the number of times a word appears in a document, divided by the total number of words in that document and the second term is the Inverse text file Frequency, computed as the logarithm of the number of the text file in the corpus divided by the number of text file where the specific term appears. Highlight author and affiliation lines of affiliation 1 and copy this selection.

RESULTS AND ANALYSIS

a) PERFORMANCE COMPARISON

The performance datasets are presented in different term weighting schemes shows different accuracy rates of which Naïve Bayes classifier with unigrams, bigrams, trigram, syntactic bigram and syntactic trigram using Binary weighting produces the maximum result. Accuracy metric is used to measure the performance of the proposed methodology. Accuracy metric is calculated using confusion matrix.

When comparing accuracy in binary weighting schemes, Unigram gives better accuracy then bigram, trigram, syntactic-bigram, syntactic-trigram.

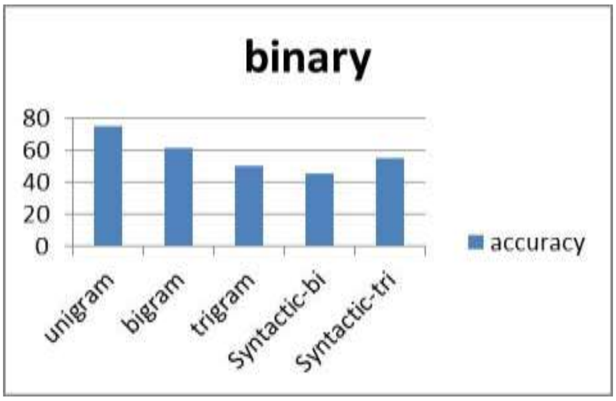

FIGURE 2: COMPARING ACCURACY IN BINARY SCHEMES

When comparing accuracy in TF weighting schemes, Unigram gives better accuracy then bigram, trigram, syntactic-bigram, syntactic-trigram.

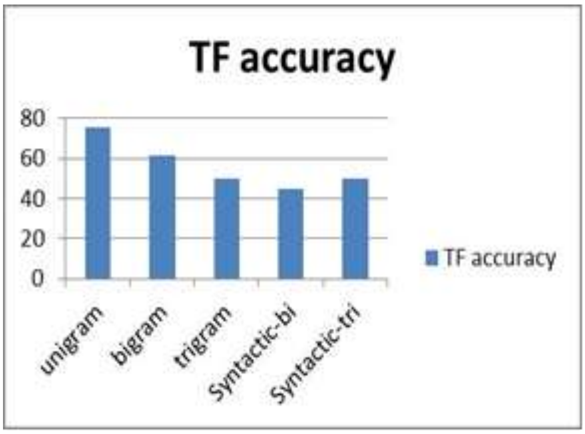

FIGURE 3: COMPARING ACCURACY IN TF SCHEMES

When comparing accuracy in TF-IDF weighting schemes, Unigram gives better accuracy then bigram, trigram, syntactic-bigram, syntactic-trigram.

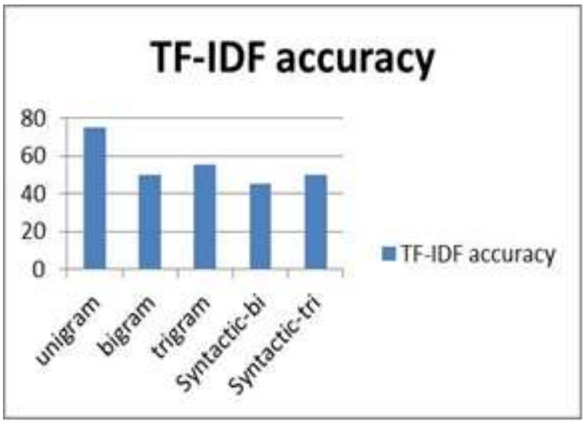

FIGURE 4: COMPARING ACCURACY IN TF-IDF SCHEMES

As a overall analysis of proposed methods, unigram approach provides better performance when compared to all other approaches.

CONCLUSION AND FUTURE WORK

Electronics datasets of Multi Domain Review dataset that is analysed portrays that Naive Bayes classifier achieves the best accuracy among the three classifiers using the Unigrams, Bigrams, Trigrams, Syntactic ngrams. Naive Bayes classifier shows a moderate fall in accuracy rates. The TF – IDF term weighting scheme outperformed amongst the five term weighting schemes used. The new term weighting scheme referenced from the text classification methods proved to be an efficient one in sentiment analysis too. The term weighting scheme proved to be a better solution to the n grams approach since it proves an efficient result in considering only the unigrams. In future, the same architecture flow with a variation of using syntactic n grams along with the different weighting schemes are planned to be implemented.

REFERENCES

[1] Agarwal, B., Mittal, N.: Categorical probability proportion difference (CPPD): a feature selection method for sentiment classification. In: Proceedings of the 2nd Workshop on Sentiment Analysis where AI Meets Psychology (SAAIP 2012), pp. 17–26. Mumbai (2012)

[2] Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders:domain adaptation for sentiment classification. In: ACL (2007)

[3] Mohan, P., & Thangavel, R. (2013). Resource Selection in Grid Environment Based on Trust Evaluation using Feedback and Performance. *American Journal of Applied Sciences*, *10*(8), 924.

[4] Esuli, A., Sebastiani, F.: SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of Language Resources and Evaluation (LREC) (2006)

[5] Hassan, A., Abbasi, A., Zeng, D.: Twitter sentiment analysis: a bootstrap ensemble framework. In: International Conference of Social Computing (SocialCom), pp. 357–364 (2013)M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.