"Review on study model on personality analysis from social media using Machine Learning Algorithms-KNN, SVM, NB"

Mr. Prashant P. Ghantiwala
Research Scholar
Sabarmati University

Dr. Mukta Agarwal
Assistant Professor
Sabarmati University

Abstract:

Blogs and social networks have recently become a valuable resource for mining sentiments in fields as diverse as customer relationship management, public opinion tracking and text filtering. In fact knowledge obtained from social networks such as Twitter and Facebook has been shown to be extremely valuable to marketing research companies, public opinion organizations and other text mining entities. However, Web texts have been classified as noisy as they represent considerable problems both at the lexical and the syntactic levels. In this research we used a random sample of people of daily post and tweets on social media

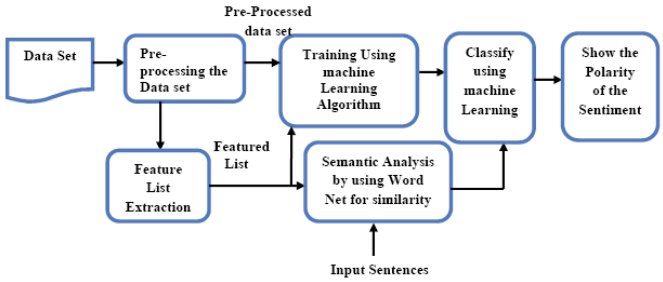
Keywords:

Machine Learning, Semantic Orientation, Naïve Bayes, Support Vector Machine, Personality Analysis Sentiment Analysis, Twitter, Facebook, Neural Network.

1. INTRODUCTION

The current research paper covers the analysis of the contents on the Web covering lots of areas which are growing exponentially in numbers as well as in volumes as sites are dedicated to specific types of products and they specialize in collecting users' reviews from various sites such as Amazon etc. Even Twitter is an area where the tweets convey opinions, but trying to obtain the overall understanding of these unstructured data (opinions) can be very time consuming. These unstructured data (opinions) on a particular site are seen by the users and thus creating an image about the products or services and hence finally generating a certain judgment. These opinions are then being generalized to gather feedbacks for different purposes to provide useful opinions where we use sentiment analysis. Sentiment analysis is a process where the dataset consists of emotions, attitudes or assessment which takes into account the way a human thinks [1]. In a sentence, trying to understand the positive and the negative aspect is a very difficult task. The features used to classify the sentences should have a very strong adjective in order to summarize the review. These contents are even written in different approaches which are not easily deduced by the users or the firms making it difficult to classify them. Sentiment analysis

influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand about their products or services in such a way that it can be offered as per the user's needs. There are two types of machine learning techniques which are generally used for sentiment analysis, one is unsupervised and the other is supervised [2]. Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering. Supervised learning is based on labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained to produce reasonable outputs when encountered during decision- making. To help us to understand the sentiment analysis in a better way, this research paper is based on the supervised machine learning. The rest of the paper is organized as follows. Second section discusses in brief about the work carried out for sentiment analysis in different domain by various researchers. Third section is about the approach we followed for sentiment analysis. Section four is about implementation details and results followed by conclusion and future work discussion in the last section..



In our approach we used the twitter dataset and analyzed it. This analyses labeled datasets using the unigram feature extraction technique. We used the framework where the preprocessor is applied to the raw sentences which make it more appropriate to understand. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content. The complete description of the approach has been described in next sub sections and the block diagram of the same is graphically represented in Fig. 1 Fig.1. Diagram of the Approach to Problem

A. Pre-processing of the datasets

The tweets contain a lot of opinions about the data which are expressed in different ways by individuals .The twitters dataset used in this work is already labeled. Labeled dataset has a negative and positive polarity and thus the analysis of the data becomes easy. The raw data having polarity is highly susceptible to inconsistency and redundancy. The quality of the data affects the results and therefore in order to improve the quality, the raw data is pre-processed. It deals with the preparation that removes the repeated words and punctuations and improves the efficiency the data. For example, "that painting is Beauuuutifull #" after preprocessing converts to "painting Beautiful." Similarly, "@Geet is Now Hardworking" converts to "Geet now hardworking".

B. Feature Extraction

The improved dataset after pre- processing has a lot of distinctive properties. The feature extraction method, extracts the aspect (adjective) from the dataset. Later this adjective is used to show the

positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using unigram model [15]. Unigram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences. For above example, i.e. "painting Beautiful" through unigram model, only Beautiful is extracted from the sentence.

C. Training and classification

Supervised learning is an important technique for solving classification problems. In this work too, we applied various supervised techniques to get the desired result for sentiment analysis. In next few paragraphs we have briefly discussed about the three supervised techniques i.e. naïve bayes, maximum entropy and support vector machine followed by the semantic analysis which was used along with all three techniques to compute the similarity.

• Naive Bayes

It has been used because of its simplicity in both during training and classifying stage. It is a probabilistic classifier and can learn the pattern of examining a set of documents that has been categorized. It compare the contents with the list of words to classify the documents to their right category [16].

C* = argmaccPNB(c|d)

Class c* is assigned to tweet d, where, f represents a feature and ni(d) represents the count of feature fi found in tweet. There are a total of m features. Parameters P(c) and P (f|c) are obtained through maximum likelihood estimates which are incremented by one for smoothing. Pre-processed data along with extracted feature is provided as input for training the classifier using naïve bayes. Once the training is complete, during classification it provides the polarity of the sentiments. For example for the review comment "I am happy' it provide Positive polarity as result.

• Maximum entropy

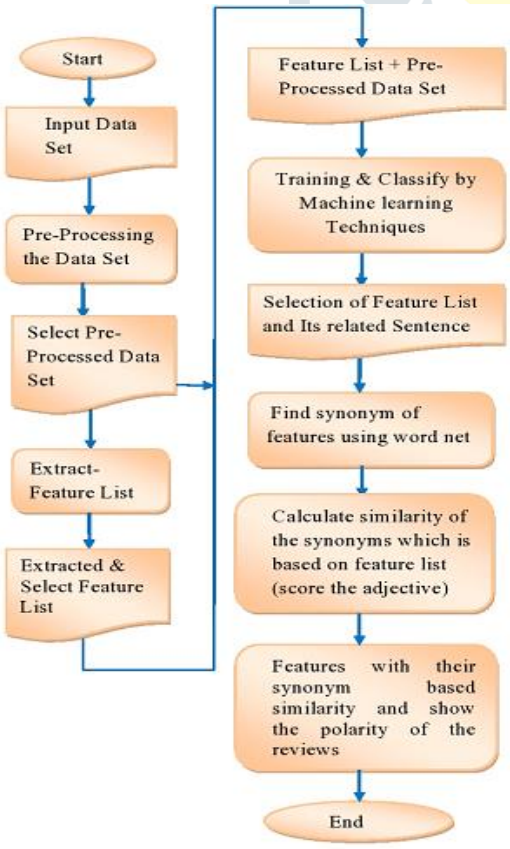
Maximum entropy maximizes the entropy defined on the conditional probability distribution. It even handles overlap feature and is same as logistic regression which finds distribution over classes. It also follows certain feature exception constraints [17]. Where, c is the class, d is the tweet, and is a weight vector. The weight vectors decide the significance of a feature in classification. It follows the similar processes as naïve bayes, discussed above and provides the polarity of the sentiments.

• Support vector machine

Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors of size m each. Then every data represented as a vector is classified in a particular class. Now the task is to find a margin between two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully [18].

• Semantic Analysis

After the training and classification we used semantic analysis. Semantic analysis is derived from the WordNet database where each term is associated with each other. This database is of English words which are linked together. If two words are close to each other, they are semantically similar. More specifically, we are able to determine synonym like similarity. We map terms and examine their relationship in the ontology. The key task is to use the stored documents that contain terms and then check the similarity with the words that the user uses in their sentences. Thus it is helpful to show the polarity of the sentiment for the users. For example in the sentence"I am happy" the word ''happy'' being an adjective gets selected and is compared with the stored feature vector for synonyms. Let us assume 2 words; 'glad' and 'satisfied' tend to be very similar to the word 'happy'. Now after the semantic analysis, 'glad' replaces 'happy' which gives a positive polarity.



REFERENCES

[1] R. Feldman, " Techniques and Applications for Sentiment Analysis

," Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.

[2] Y. Singh, P. K. Bhatia, and O.P. Sangwan, "A Review of Studies on

Machine Learning Techniques," International Journal of Computer

Science and Security, Volume (1) : Issue (1), pp. 70-84, 2007.

[3] P.D. Turney," Thumbs Up or Thumbs Down? Semantic Orientation

Applied to Unsupervised Classification of Reviews," Proceedings of the

40th Annual Meeting of the Association for Computational Linguistics

(ACL), Philadelphia, pp. 417-424, July 2002.

[4] Ch.L.Liu, W.H. Hsaio, C.H. Lee,and G.C.Lu, and E. Jou," Movie Rating

and Review Summarization in Mobile Environment," IEEE Transactions

on Systems, Man, and Cybernetics, Part C 42(3):pp.397-407, 2012.

[5] Y.Luo,W.Huang," Product Review Information Extraction Based on

Adjective Opinion Words," Fourth International Joint Conference on

Computational Sciences and Optimization (CSO), pp.1309 – 1313,

2011.

[6] R.Liu,R.Xiong,and L.Song, "A Sentiment Classification Method for

Chinese Document," Processed of the 5th International Conference on

Computer Science and Education (ICCSE), pp. 918 – 922, 2010.

[7] A.khan,B.Baharudin, "Sentiment Classification Using Sentence-level

Semantic Orientation of Opinion Terms from Blogs," Processed on

National Postgraduate Conference (NPC), pp. 1 – 7, 2011.

[8] L.Ramachandran,E.F.Gehringer, "Automated Assessment of Review

Quality Using Latent Semantic Analysis," ICALT, IEEE Computer

Society, pp. 136-138, 2011.

[9] B.Agarwal,V.K.Sharma,andN.Mittal,"Sentiment Classification of

Review Documents using Phrase Patterns," International Conference on

Advances in Computing, Communications and Informatics (ICACCI),

pp. 1577-1580, . 2013.

[10] J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,," Aspect-Based Opinion

Polling from Customer Reviews," T. Affective Computing2(1):pp. 37-

49, 2011.

[11] M.Karamibekr,A.A.Ghorbani,"Verb Oriented Sentiment Classification,"

Processed of the IEEE/WIC/ACM International Conferences on Web

Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp.

327-331, 2012.

[12] A. Neviarouskaya, H.Prendinger, and M.Ishizuka," SentiFul: A Lexicon

for Sentiment Analysis," T. Affective Computing 2(1), pp.22-36, 2011.

[13] L.Liu, X.Nie,and H.Wang," Toward a Fuzzy Domain Sentiment

Ontology Tree for Sentiment Analysis," Processed of the 5th Image

International Congress on Signal Processing (CISP), pp. 1620 – 1624,

2012.

[14] R. Srivastava, M. P. S. Bhatia," Quantifying Modified Opinion Strength:

A Fuzzy Inference System for Sentiment Analysis," International

Conference on Advances in Computing, Communications and

Informatics (ICACCI), pp. 1512-1519, 2013.

[15] C. Tillmann , and F. Xia, "A phrase-based unigram model for statistical

machine translation," Proceedings of the 2003 Conference of the North

American Chapter of the Association for Computational Linguistics on

Human Language Technology: companion volume of the Proceedings of

HLT-NAACL, pp.106-108, 2003.

[16] B.Ren,L.Cheng," Research of Classification System based on Naive

Bayes and MetaClass," Second International Conference on Information

and Computing Science, ICIC '09, Vol(3), pp. 154 – 156, 2009.

[17] C.I.Tsatsoulis,M.Hofmann,"Focusing on Maximum Entropy

Classification of Lyrics by Tom Waits," IEEE International on Advance

Computing Conference (IACC), pp. 664 – 667, 2014.

[18] M.A. Hearst,"Support vector machines,"IEEE Intelligent Systems, pp.

18-28, 1998.