ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Developing a Neural Network Based Approach for Sentiment Classification

[1] M.Karthi, [2] R.Kirubhakaran,[3] S.Vijay,[4] K.K.Yuvaraj kumar, [5] D.Suganya
[1][2][3][4] Final Year Student, [5] Assistant Professor
[1][2][3][4][5]Department of IT, Sengunthar College of Engineering, Tiruchengode

Abstract – **The study of sentimental analysis and opinion mining deals with attitude and emotions. Opinion mining has several challenges. The first challenge is that a word is either positive in one situation or negative in another situation. Therefore, sentiment can be performed using social media messages.. To overcome this problem, social media messages are used which are free of cost and produced in world wide. In this the public concern can be measured using two-step word alignment approach. In the first step, raw reviews are separated into personal reviews and news reviews. In the second step, personal reviews are further classified into personal negative and personal non-negative. In both steps, the trained data is generated automatically using an emotional-oriented, clue-based method and the trained dataset can be tested using machine learning model such as Naïve Bayes. The proposed algorithm will increase the accuracy for epidemic domain.**

Keywords—classification, emotion, emotion analysis, Naïve Bayes, SVM and IBK.

1. INTRODUCTION

Data mining is a process of searching large data to discover patterns for simple analysis. Data mining is a technology to help companies focus on their data warehouse. So it is called as Knowledge Discovery in Data (KDD) [l2]. KDD decisions are allowed by data mining tools for businesses. Data mining tools can answer business questions that traditionally were time consuming to resolve. Figure l.l shows the data mining process an important task of public health officials is to keep track of health issues, such as spreading epidemics. In sentiment classification, it addresses the issue of spreading public concern about epidemics. Public concern about a communicable disease can be seen as a problem of its own. Keeping track of trends in concern about public health and identifying peaks of public concern are therefore crucial tasks. However monitoring public health concerns is not only expensive with traditional surveillance systems, but also suffers from limited coverage and significant delays. To address these problems, Sentiment classification explains social media messages, which are available free of cost, are generated world-wide, and are posted in real time. It involves with the process of measuring public concern using a two-step sentiment word alignment approach. In the first step, it can distinguish Personal reviews from News (i.e.,Non-Personal)reviews. In the second step, further separate Personal Negative from Personal Non-Negative reviews. Both these steps consist themselves of two sub steps. In the first sub-step (of both steps), our programs automatically generate training data using an emotion oriented, clue-based method. In the second sub-step, training and testing three different Machine Learning (ML) models with the training data from the first sub-step; this allows us to determine the best ML model for different datasets. Furthermore, it tests the already trained ML models with a human annotated, disjoint dataset. Based on the number of reviews classified as Personal Negative, it computes a Measure of concern and a timeline of the MOC. It attempts to correlate peaks of the MOC timeline to peaks of the News (Non-personal) timeline. Our best accuracy results are achieved using the two-step method with a Naive Bayes classifier for the epidemic domain (six datasets) and the Mental Health domain (three datasets).

2. LITERATURE SURVEY

The automatic analysis of user generated contents such as online new , reviews, blogs and reviews are extremely valuable for tasks such as mass opinion estimation, corporate reputation measurement, political orientation categorization, stock market prediction, customer preference and public opinion study.

Public health surveillance is critical to monitoring the spread of infectious diseases using traditional surveillance systems. To address this problem social media message has been used. Clue based method is the first step for auto labeling the raw reviews into personal reviews or news reviews. Riloff and Wiebe(2003) labeled the reviews using corpus such as MPQA corpus. For sentiment word alignment, the standard Naive Bayes algorithm is used. The classifiers have been reported to achieve maximum accuracy.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Pang and Lee L (2008) focused on the methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact based analysis. The proposed work includes a material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact .To facilitates future work, a discussion of available resources , benchmark datasets, an d evaluation campaigns is also provided.

Zhang et al (2011) proposed sentiment analysis on entities in reviews thus becomes a rapid and effective way o f gauging public opinion for business marketing or social studies. It proposed a new entity-level sentiment analysis method for Social media with lexicon based approach. This method produced high precision, low recall and to improve recall additional reviews that are likely to be opinionated arc identified automatically by exploiting the information in the result of the lexicon-based method. A classifier is then trained to assign polarities to the entities in the newly identified reviews.

Khan et al (2013) focused to predict the polarity of words and classify into positive and negative feelings that were expressed in any form or language. Sentiment analysis over social media offers a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc. The proposed algorithm Naive Bayes is implemented to overcome the issues of classification accuracy, data sparsely and sarcasm.

Hassan et al (2015) dealt with sentiment analysis on social media has attracted much attention on commercial and public sectors. A lexicon-based approach and senticircle was introduced for sentiment analysis on Social media. It allowed for the detection of sentiment at both entity-level and review-level. In this method, social media datasets were used to derive word prior sentiments and increase the accuracy of sentiment analysis.

Xiang Ji et al (2011) developed Epidemics Outbreak and Spread Detection System (EOSDS) as a prototype that makes use of the rich information retrievable in real time from Social media. EOSDS provides three different visualization methods of spreading epidemics, static map, distribution map, and filter map, to investigate public health threats in the space and time dimensions.

3.EXISTING SYSTEM

In Existing system Keeping track of trends in concern about public health and identifying peaks of public concern are therefore crucial tasks. However, monitoring public health concerns is not only expensive with traditional surveillance systems, but also suffers from limited coverage and significant delays. To address these problems, it uses Social media messages, which are available, free of cost, are generated world-wide, and are posted in real time. It measures public concern using a two-step sentiment word alignment approach. In the first step, it distinguishes Personal reviews from News (i.e., Non-Personal) reviews.

3.1 Dataset Collection The dataset was obtained from [10]. For creating the dataset, first seven different emotion words for seven emotion categories were collected from existing psychology theory, then Twitter Streaming API was used for collecting tweets which had at least one of the emotion word in the form a hashtag[8]. The collected dataset was then divided into testing and training sets. The training sets were used to train the classifiers so that the test data is correctly labeled.

3.2 Preprocessing Preprocessing of the collected data is of utmost importance because the tweets are only 140 words long in length and hence there is a presence of slang, URL¨s, user-mentions which do not contribute in any manner to the classification process, in fact the presence of such elements can mislead the classifier. The preprocessing steps include the following[11]:

•Lower casing all the words.
•Replacing user mention with @user.
•Replacing letters and punctuations that are repeated more than twice with two same letters (Eg.happpy□happy).
•Removing hashtags.

3.3 Feature Extraction

After data preprocessing, the stopwords are removed from the tweets by using a stopword file, this is done because they usually constitute large components of sentences and they do not provide useful information. After stopword removal the feature extraction process is done. In this paper bag of words is considered for training both the classifiers that are used.

3.4 Classifier

Naïve Bayes and Support Vector Machine(SVM) are used for classification purpose because SVM is found to be very useful in matters of handling sparse data and Naïve

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Bayes works effectively for text classification. pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

The results were taken by experimenting with the size of the dataset used for training. The training data with 600 records approximately, with equal records belonging to the five classes of emotion that is happy, sad, angry, fear and surprise is used for taking results given in the TABLE 1 and for TABLE 2 a training set of 1000 records is considered.

	Accury	F1score	Recall	Precision
SVM	0.94	0.9395	0.94	0.9484
Naïve ayes	0.94	0.9407	0.94	0.9484

TABLE 1: Results for SVM and Naïve Bayes for accuracy, F1 score, Recall and Precision with an input of 600 training.

	Accury	F1score	Recall	Precision
SVM	0.89	0.8967	0.8932	0.91
Naïve ayes	0.86	0.867	0.861	0.92

TABLE 2: Results for SVM and Naïve Bayes for accuracy, F1 score, Recall and Precision with an input of 1000 training.

4. DRAWBACKS OF EXISTING SYSTEM

• identifying reviews in the free text reviews, a straightforward solution is to employ an existing aspect identification approach.
• The network lifetime may even be reduced.
• Less accuracy prediction on opinion analysis.
• User review based word alignment is cumbersome.
• High in latency to analyze the datasets.

5. PROPOSED SYSTEM

In proposed work a reviews aspect ranking framework to automatically identify the important reviews of reviews from online user reviews. NB Naive Bayes based opinion review analysis reviews possess the following characteristics: (a) they are frequently commented in user reviews; and (b) users' opinions on these reviews greatly influence their overall opinions on the reviews. A straight forward frequency based solution is to regard the reviews that are frequently commented in user reviews as important. However, users' opinions on the frequent reviews may not influence their overall opinions on the reviews, and would not influence their purchasing decisions. It measures public concern using a two-step sentiment word alignment approach.
In the first step, distinguish Personal reviews from News (i.e., Non-Personal) reviews. In the second step, it further separate Personal Negative from Personal Non- Negative reviews. Both these steps consist themselves of two sub steps. In the first sub-step (of both steps), our programs automatically generate training data using an emotion oriented, clue-based method. In the second sub-step, it trains and testing three different Machine Learning (ML) (Unigram, bigram, trigram)models with the training data from the first sub-step; this allows us to determine the best ML model for different datasets. Furthermore, it tests the already trained ML models with a human annotated, disjoint dataset.

5.1 NAIVE BAYESIAN CLASSIFIER

The Naive Bayes method is a set of supervised learning algorithms based on Bayesian theorem and it is suited when the dimensionality of the inputs is high. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods.

5.2 SUPPORT VECTOR MACHINE

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze [l7] data and recognize patterns, used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper Plane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

5.3 IBK

Following four steps are used in proposed work.
1) Reviews extraction and Preprocessing.
2) Aspect Identification of the product

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

3) Classify the positive and negative reviews of product by sentiment classifier.
4) The probabilistic ranking algorithm used for product ranking.

The IBK Based review classification purpose of the work it are going to find import aspect of the product and its rank this aspect by using numerous consumer reviews. The consumer reviews contain rich and valuable knowledge about the product. And this knowledge is very useful for both consumer and firm. Consumer can make wise purchasing decision by paying more attention towards important aspect or feature. And firm will concentrate on important features or aspect while improving the quality of the aspect. So in this proposed framework, this will identify the important aspect of product from online consumer reviews. The important aspects are commented again and again in consumer review and the consumers opinions on the important aspects are greatly influence their overall opinions on the product. From the consumer reviews the important aspect are identified by using NPL tool, and will classify the sentiment on that aspect, and finally it are going apply the ranking algorithm to determine the particular product rating.

The results were taken by experimenting with the size of the dataset used for training. The training data with 600 records approximately, with equal records belonging to the five classes of emotion that is happy, sad, angry, fear and surprise is used for taking results given in the TABLE 1 and for TABLE 2 a training set of 1000 records is considered.

	Accuracy	F1score	Recall	Precision
IBK	0.93	0.9295	0.943	0.9384
SVM	0.94	0.9395	0.94	0.9484
Naïve Bayes	0.94	0.9407	0.94	0.9484

TABLE 1: Results for IBK,SVM and Naïve Bayes for accuracy, F1 score, Recall and Precision with an input of 600 training.

	Accuracy	F1score	Recall	Precision
IBK	0.92	0.9295	0.943	0.9384
SVM	0.89	0.8967	0.8932	0.91
Naïve Bayes	0.86	0.867	0.861	0.92

6. MODULE DESCRIPTION

• Preprocessing of features
• principal component analysis
• machine learnlng classifiers for personal review classification

6.1 PRE-PROCESSING OF FEATURES

In cases of disease surveillance on Social media, the classical division of sentiments into positive and negative is inappropriate, because diseases are generally classified as negative. Positive emotions could arise as a result of relief about an epidemic subsiding, but ignore this possibility. Thus, a two-point ‘‘Like rt scale" with the points positive and negative would not cover this spectrum well. Rather, it started with an asymmetric four-point Like scale of "strongly negative", "negative", "neutral", and "positive " then combined "strongly negative" and "negative" into one category, and "neutral" and "positive" into another. It use "Negative" as the name of the first category and "Non-Negative" for the second one. Thus, the problem reduces to a two-class word alignment problem, and a Personal review can either be a Negative review or a Non-Negative review. Some features need to be removed or replaced. It first deleted the reviews starting with "RT", which indicates that they are re-reviews without comments to avoid duplications. For the remaining reviews, the special characters were removed. The URLs in Social media were replaced by the string "url". Social media’s special character "@" was replaced by “tag”. For punctuations, “!” and “?” were substituted by “excl” and “ques”, respectively, and any of “.-|? =/” were replaced by “symb”. Social media messages were transformed into vectors of words, such that every word was used as one feature, and only unigrams were utilized for simplicity.

6.2 PRINICPAL COMPONENT ANALYSIS

The principal component analysis parses each review into a set of tokens and matches them with a corpus of Personal clues. There is no available corpus of clues for Personal versus News classification .The MPQA corpus contains a total of 8221 words, including 3250 adjectives, 329 adverbs, 1146 any-position words, 2167 nouns, and 1322 verbs. As for the sentiment polarity, among all 8221 words, 4912 are negatives, 570 are neutrals, 2718 are positives, and 21 can be both negative and positive. In terms of strength of subjectivity, among all words, 5569 are strongly subjective words, and the other 2652 are weakly subjective words. Social media users tend to express their personal opinions in a more casual

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

way compared with other documents, such as News, online reviews, and article comments. It is expected that the existence of any profanity might lead to the conclusion that the review is a Personal review.

6.3 MACHINE LEARNING CLASSIFIERS FOR PERSONAL REVIEW CLASSIFICATION

To overcome the drawback of low recall in the clue-based approach, combine the high precision of clue-based classification with Machine Learning-based classification in the Personal vs. News classification. After the Personal vs. News classifier is trained, the classifier is used to make predictions on each in TO, which is the preprocessed reviews dataset, The goal of Personal vs, News classification is obtain the Separate Labels.

7. ADVANTAGES OF PROPOSED SYSTEM:

• It first identifies the nouns and noun phrases in the documents. The occurrence frequencies of the nouns and noun phrases are counted, and only the frequent ones are kept as reviews.
• The language model was built on reviews, and used to predict the related scores of the candidate reviews. The candidates with low scores were then filtered out.
• The admin can easily identify related opinion reviews on that session.
• Easily determine reviews quality by using customer reviews.
• It can find Based on the number of reviews classified as Personal Negative; compute a Measure of Concern (MOC) and a timeline of the MOC. It attempt to correlate peaks of the Moc timeline to peaks of the News (Non-Personal) timeline.
• Best accuracy results are achieved using the two-step method with a Naïve Bayes classifier for the Epidemic domain (six datasets) and the Mental Health domain(three datasets).

8. CONCLUSION

The project is to monitor the public health concern from the reviews and them as positive and negative opinion. To find accuracy a two-step sentiment classification approach is implemented: In the first step, classify health reviews into Personal disease inference reviews versus News reviews. It uses a subjective clue-based lexicon and News stop words to automatically extract training datasets labeling Personal disease inference disease inference reviews and News reviews. These auto-generated training datasets are then used to train Machine Learning models to classify whether a review is Personal disease inference disease inference or News. In the second step, utilized an emotion-oriented clue-based method to automatically extract training datasets and generate another classifier to predict whether a Personal disease inference review is Negative or Non-Negative. In sentiment classification, by combining a clue-based method with a machine learning method, good accuracy can be achieved. This overcomes the drawbacks of the clue-based method and the Machine Learning methods when used separately.

Future work is to implement with irony concept. Ironic statement is used to express the opposite of what is being said. For the sentiment-classification task, the appearance of irony often indicates the opposite of the literal meaning of a statement and Irony is implemented using MapReduce in hadoop which is open source framework for running large number of dataset.

9. REFERENCES

[1]. Yoad Lewenberg, Yoram Bachrach, Svitlana Volkova," Using Emotions to Predict User Interest Areas in Online Social Networks", 2015 IEEE

[2]. James A. Russell, "Emotion in Human Consciousness Is Built in Core Affect", Journal of Consciousness Studies,12,No.8-10,pp 26-42, 2005.

[3]. DavidWatson; Auke Tellegan, "Towards a consensual structure of Mood", Psychological Bulletin, Vol. 98, No. 2. 219-235,1985.

[4]. Hugo Lövheim, A new three-dimensional model for emotions and monoamine neurotransmitters, Medical Hypotheses 78 (2012) 341–348. Processing and its Applications.Research in Computing Science 46, 2010, pp. 131-142.

[5]. Paul Ekman, "Universals and Cultural Differences in Facial Expressions of Emotion", Nebraska Symposium on Motivation, Vol 19,1971.

[6]. Liza Wikarsa, Sherly Novianti Thahir, "A text mining application of Emotion Classifications of Twitters Users using Nave Bayes Method", IEEE 2015.

[7]. Li Yu, Zhifan Yang, Peng Nie, Xue Zhao, Ying Zhang, "Multi-Source Emotion Tagging for Online News", 12th Web Information System and Application Conference 2015.

[8]. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth," Harnessing Twitter „Big Data‟ for Automatic Emotion Identification", 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.

[9]. K Dhanasekaran and R Rajeswari, "Text feature classification approach for effective information extraction via discriminative sequence analysis", International Journal of Applied Engineering Research, Vol. 10 (1), pp.2067-2079, 2015.

[10]. http://knoesis.org/projects/emotion.

[11]. Sanket Sahu, suraj Kumar Rout, Debasmit Mohanty,"Twitter Sentiment Analysis:A more enhanced way of classification and scoring", 2015 IEEE International Symposium on Nanoelectronic and Information Systems.