ISSN 2395-1621

Use of DSA model by considering two sides of one single review

[#1]Mrunmayi Chaudhari, [#2]Vishal Mogal

[1]mrunmayi.chaudhari26@gmail.com
[2]vishalmogal.rmdssoe@sinhgad.edu

[1]Computer engineering, RMD Sinhgad School of Engineering,
[2]Computer engineering, RMD Sinhgad School of Engineering

ABSTRACT

Many a time the Bag-of-words (BOW) is measured as important part to mold text in arithmetical machine learning way in the sentiment analysis. Earlier BOW was used for text representation. Whereas on the other hand, sometimes the presentation of BOW remains limited due to some basic deficiencies in arranging the polarity shift problem. In this, firstly we propose a model called dual sentiment analysis (DSA), to deal with this problem for sentiment classification. In this we first propose a new data extension technique by creating a sentiment reversed review for each training samples and test reviews. Based on this, we propose a dual training algorithm in which we make use of original and reversed training reviews in pairs for training classifier, and a dual prediction algorithm for ordering the test reviews by considering two sides of one single review. And in this we also widen the DSA structure from polarity (positive-negative) classification to 3-class (positive-negative- neutral) classification, by considering the neutral reviews in addition. And then we finally develop a corpus-based method to construct a pseudo-antonym dictionary, which eliminates DSAs dependency on an outer antonym dictionary for review improvement The results express the utility of DSA in showing polarity shift in sentiment classification.

Keywords: Machine learning, sentiment analysis, opinion mining, natural language processing.

ARTICLE INFO

Article History

Received: 24th August 2016

Received in revised form :

24th August 2016

Accepted: 28th August 2016

Published online :

28th August 2016

I. INTRODUCTION

SENTIMENT analysis is one of the basic task of discovering the opinions of people and perpetuity of people towards explicit topics of interest. Whether it is product, thing or a movie, opinions of people matters a lot, and it somehow affects the important decision-making process of people. The rest another chief thing a person does is to see the kind of reviews and opinions that people have written or given when he or she wants to buy a product online. Social media such as Face book, blogs, twitter have become a distinct place where people give their opinions on certain topics. The sentiments of the tweets of a particular subject has numerous usage, including stock market analysis of a company, movie reviews, in psychology to analyze the frame of mind of people that has a variety of applications, and so on. Sentiments of tweets can be divided into many categories like positive, negative, neutral, extremely positive, extremely negative, and so on. The two types of sentiments considered here in this classification experiment are positive and negative sentiments. The data which is being labeled by humans has a lot of noise, and it is difficult many a times to achieve good accuracy**.**

II. LITERATURE SURVEY

Lei Zhang, Bing Liu, Suk Hwan Lim, Eamonn OBrien-Strain, Extracting and Ranking Product Features in Opinion Documents , proposed an important task of opinion mining for extracting People's opinion on features of an element. For example, the line, I love the GPS function of Motorola Droid expresses a positive opinion on the GPS function of the Motorola phone. GPS function is the feature here. This paper focuses on mining the features. Double propagation is a situation-of-the-art technique for solving the problem. It works very well for medium-size corpora. However, for large and small corpora, it can result in low accuracy and low evoke. To deal with these two problems, two improvements based on part-whole and no patterns are introduced to increase the recall. Then ranking feature is then applied to the extracted ,sorted feature candidates to improve the accuracy of the top-ranked candidates. We then rank feature

candidates by feature importance which is dogged by two factors: feature relevance and feature frequency. The problem is formulated as a bipartite graph and the well-known important web page ranking algorithm HITS is then used to discover important features and rank them high accordingly. Experiments on diverse real-life datasets give capable results.

Muhammad Zubair Asghar , Aurangzeb Khan, Shakeel Ahmad , Fazal Masud Kundi, A Review of Feature Extraction in Sentiment Analysis this paper proposed speedy increase in internet users along with growing influence of online review sites and social media has given origin to Sentiment analysis or Opinion mining, which aims at finding what other people think and comment on particular product. Sentiments or Opinions contain public generated content about products, services, policies and politics. People are many a times interested to get positive and negative opinions which contains mostly likes and dislikes being shared by users as features of particular product or service. Therefore product features or aspects have been given significant and distinct role in sentiment analysis. In addition to sufficient work being performed in text analytics, feature extraction in sentiment analysis is now flattering an energetic area of research. This review paper discusses various existing techniques and approaches for feature extraction in sentiment analysis and opinion mining.

S.J.Veeraselvi, M.Deepa, Survey on Sentiment Analysis and Sentiment Classification proposed Opinions are the fundamental aspect to almost all decision making performance activities. The more and more usage of internet and the trade of user opinions through social media and public forums on the web has become the inspiration for sentiment analysis. Due to the huge or infinite amount of user opinions available on the web, it has become necessary to automatically examine and order sentiment expressed in opinions for making the decision making process a simple task. Opinion Mining or Sentiment Analysis is the Natural Language Processing method that allows the system to automatically recognize and extract sentiments which are expressed in user reviews. The fundamental task of sentiment analysis is sentiment classification which classifies a user review as positive, negative, neutral. This survey gives a general idea of sentiment analysis, sentiment classification, methods used for sentiment classification.

Andrea Esuli and Fabrizio Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining propose Opinion mining (OM) is a recent sub discipline at the crossroads of information retrieval and computational linguistics which does not deal with the topic a document is about, but with the opinion it gives. OM has a prosperous set of applications, ranging from tracking users opinions about products or about political candidates as uttered in online forums, to customer relationship management. In order to help the extraction of opinions from content, recent research has tried well to automatically resolve the PN-polarity of subjective terms, i.e. identify whether a term has a positive or a negative connotation. In this work we describe SENTIWORDNET, a lexical resource in which each and every WORDNET called synsets is and Neg(s), describing how objective, positive, and negative the terms in the synset are. The method is used to develop SENTIWORDNET is based on the quantitative analysis related to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification. The three scores are derived by combining the results produced by a committee of eight ternary classifiers, all characterized by similar accuracy levels but with different classification behavior. SENTIWORDNET is freely available for most research purposes, and is gifted with aWeb-based graphical user interface.

III.SYSTEM ARCHITECTURE

1. Dual Training:

In this Firstly in the training stage, all the original training samples that we have are reversed to their opposites. In this we call them as original training set and reversed training set respectively.

In data expansion technique we observe one-to-one correspondence between the original and reversed reviews. Means whenever we have two sets then a single element in one set is exactly matched with one single element in another set. And in this technique the classifier is taught by a large and huge combination of the likelihoods of the original and reversed training samples. And we call this process as dual training (DT). And here logistic regression is used .Logistic regression is mostly and widelyused statistical model for the binary classification problem. Logistic regression makes use of the logistic function for the prediction of the probability of a feature vector x which belongs to the positive class.

2. Dual Prediction:

And then in the prediction stage, we create a reversed test sample –x for each test sample x. But instead, here we make use of –x for helping in the prediction of x. This process is called dual prediction (DP). In DP, the predictions are made When we want to determine how positive a test review x is, in this we not only take into consideration how positive the original test review is, but also take into consideration how negative the reversed
test review is .

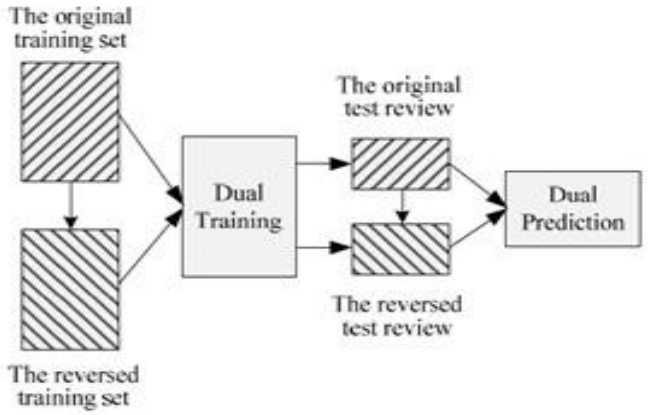


Fig. 1: The process of dual sentiment analysis. The rectangle filled with slash denotes the original data, and the rectangle filled with backslash denotes the reversed data.

Fig. 1: Dual Sentiment Analysis Process

On the contrary, when we measure how negative a test review x is, we take into consideration the probability of x being negative, as well as the probability of –x being positive and so on. In dual prediction, the use of reversed review is for reducing the prediction of original review.

3. The Antonym Dictionary for Review Reversion:

We have noticed that DSA highly and mainly depends on an outer antonym dictionary for the review reversion.

3.3.1 The Lexicon-based Antonym Dictionary

An Example of lexicon based Antonym dictionary is wordNet.WorldNet is a lexical database which keeps English words into sets of synonyms called synsets, which gives general definitions, and gives the various semantic relations between the sets. With the help of this it is possible to obtain the words and their opposites. The WordNet antonym dictionary is easy to use and direct. However, in many languages other than English, such an antonym dictionary is not readily available. Even if we try to get an antonym dictionary for the same, it is still hard to promise whether the vocabularies in the dictionary are domain consistent with our tasks. To solve this problem, we furthermore expand a corpus- based method to construct a pseudo-antonym dictionary.

3.3.2 The Corpus-based Pseudo-Antonym Dictionary

In this corpus based pseudo antonym dictionary, we make use of Mutual information.

The mutual information (MI) of two casual variables is a quantity that gives or measures the mutual dependence of the two random variables. MI is mostly used as a feature assortment method in sentiment classification. In this firstly, we select all adjectives, adverbs and verbs in the training corpus as candidate features, and use the MI metric to estimate the significance of each candidate feature to the Positive (+) and Negative (-) class, respectively: in this we have two groups of words. And Then, we rank two groups of features that we have in a decreasing order of mutual information MI. And the words that have the same ranking positions are considered or taken to be as a pair of antonyms.

Building sentiment lexicon: A huge amount of the work which is done in sentiment analysis uses a pre-constructed lexicon of sentiment manner words for examining texts. These lexicons can be built using thesauri, taking a small set of sentiment-behavior words such as "excellent "and "horrible "and then expanding on each of the two polar sides of sentiment that is positive and negative. First we add the directly related words in the thesaurus, which is followed by adding words depending on how repeatedly they co-occur with words which are already in one of the two polar sides.

Weighting schemes and normalization: Based on to a study done by Paltoglou and Thelwall in 2010, increase in sentiment analysis accuracy can be achieved by implementing weighting schemes that are not purely dual in nature. They show that by implementing or making use of a modification on weights, for example through the use of TF/IDF5, on the whole accuracy can be greatly achieved:

Different Modules considered here are:

1) **Pre-processing:** During the growth of the LLA, one dataset was used, the Large Movie Review Dataset provided by Stanford for use in testing and increasing in the field of sentiment analysis (Maas et.al, 2011). The LMRD mainly consists of an impressive 50.000 classified movie reviews, 25.000 positive and 25.000 negative.

2) **Stemming:** Stemming, is the process of cutting down or reducing the words making them short (to their stems), in an attempt to minimize or reduce the number of unique words which are available for analysis (Porter, 1980).

During the stemming process one tries to remove postfixes of words, such as removing "ing "or "ly "from the end of words for ease, so that similar words can more effectively and efficiently be grouped together and used whenever required.

3) **Stopwords:** Stopwords can be stated as words which holds (little to) no value in terms of relevance and Savoy, 2010). And these consist of words such as the, and, a and so on. These words are too common and frequent, and taking place in most or much of every review that they hold no value with respect to the sentiment.

4) **Tokenization and removal of punctuation:**

Tokenization is the process of separating or sorting out a string of characters into many shorter strings or characters, or words (The Stanford NLP Group n.d.).

5) **The non-existence of pre-processing:**

Even though the earlier sections clarify the parts of pre-processing which can be used for the analysis, the LLA can function without some or all of these steps. The point is not to exclude useful tools when they are available, but instead to still be able to function when they are not.

III.IMPLEMNTATION DETAILS

In this project we are making use of Naïve bayes technique for classification of the sentiments and it is best suited for the sentiment classification.

4.1 Naive Bayes

Naive Bayesian classification is dependent on Bayesian theorem. It is mostly suited or applicable when the dimensions of inputs that we are giving are high. Parameter estimation for this mainly makes use of method known as maximum likelihood. And its advantage is that it requires or makes use of small amount of training data to make estimations regarding parameters.

Naive Bayes classifiers are family of short and simple probabilistic classifiers which is mostly based on the theorem of Bayes with the very strong independence assumptions between the given features. Naive bayes remains the most popular and famous technique for text categorization. It is competitive or active in this domain with one or more advanced methods.

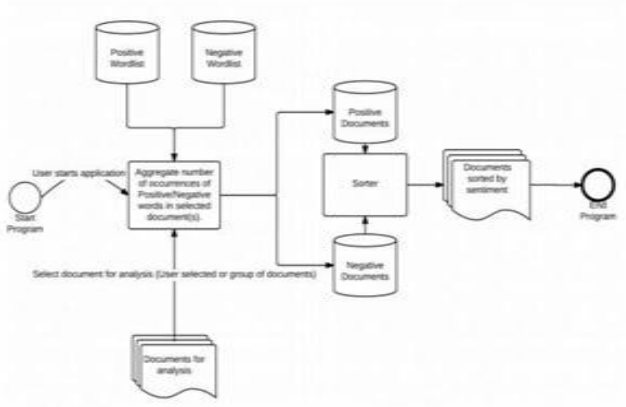


Fig. 2: System Architecture for sentiment classification

Mathematical model:

1.Dual Training and Dual prediction :

Input set:

I/p={D ,D', X,X',W}

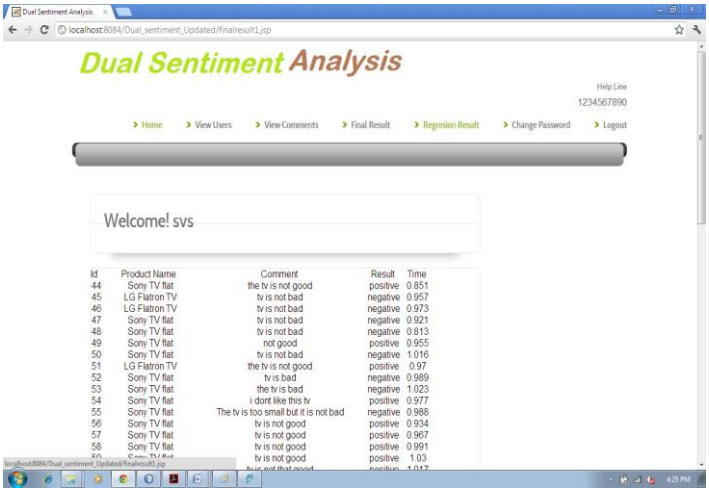
D-original training set

D'-reversed training set
X-original sample
X'-reversed sample
W- weights of samples
Output set:
O/p={p(.|x,x')}
P(.|x,x')- Dual prediction of original and reversed sample

IV. RESULTS

In our project we are taking reviews of products as input where we will find the class of reviews.Then we reverse the review by performing data mining operations such as= tokenizing, removing stop words.Then we find the class of reversed review if both class are opposite then our prediction of original review is true.Thus we are considering both side of review which gives us exact class of review. In the result, we can see both original and opposite review.

And the steps for the same are given as follows:
1) Get the comments or reviews regarding a product or service from the users.
Eg.I like this book, it is interesting.(Total words=7)
2) Remove the Phrases from the comments or reviews. After removing phrases or stop words we will get (Like, book, interesting)
3) Calculate or find out the probability of the given words.(Positive-negativeneutral). Now compare each word with the data set and calculate count. So "like "and "interesting "are two positive words.(Probability(positive=(2/7)))
4) Create reverse comments of the corresponding original comments. Now replace the verbs with opposite words we will get the reverse comment.
5) Calculate the probability for the word. And repeat the process for reversed comment.
6) And then match the result that we get for both.



V. CONCLUSION

This survey often discusses and gives details on various approaches and ways to Opinion Mining and analysis of Sentiment. It gives a detailed vision of different applications and all possible challenges for Opinion Mining making it a tricky task. Many of the machine learning methods like Maximum Entropy and Support Vector Machines also has been discussed. Many of the applications of Opinion Mining are based on bag-of-words(BOW), which do not confine context which is actually essential for Sentiment Analysis. The recent developments in Sentiment Analysis and its related sub-tasks are also presented in this. The state of the art of existing approaches has been given with the focal point on the following tasks: Subjectivity detection, Word Sense Disambiguation, Feature Extraction and Sentiment Classification using various Machine learning techniques.

VI. REFERENCES

[1] Alexander Pak and Patrick Paroubek, Twitter as a corpus for Sentiment Analysis and Opinion Mining, Proceedings of the Seventh conference on International Language Resources and Evaluation, pp. 1320-1326, 2010.

[2] Bo Pang, Lillian Lee and Vaithyananthan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp. 79-86, 2002.

[3] Luis Cabral and Ali Hortacsu, The dynamics of seller reputation: Theory and evidence from eBay, The Journal of Industrial Economics, Vol. 58, No. 1, pp. 54-78, 2010.

[4] Ellen Spertus, Smokey: Automatic recognition of hostile messages, Proceedings of Innovative Applications of Artificial Intelligence, pp. 10581065, 1997.

[5] Sanjiv Das, Asis Martinez-Jerez and Peter Tufano, eInformation: A clinical study of investor discussion and sentiment, Financial Management, Vol. 34, No. 3, pp.103137, 2005.

[6] Yun Niu, Xiaodan Zhu, Jianhua Li and Graeme Hirst, Analysis of polarity information in medical text, Proceedings of the American Medical Informatics Association, Annual Symposium, pp. 570-574, 2005.

[7] Shitanshu Verma and Pushpak Bhattacharyya, IncorporatingSemantic Knowledge for Sentiment Analysis, Proceedingsof International Conference on Natural Language Processing, 2009.ng ,november 2,2011.

