NATIONAL JOURNAL OF ARTS, COMMERCE & SCIENTIFIC RESEARCH REVIEW

SENTIMENTAL CLASSIFICATION USING NEAREST NEIGHBOR CLASSIFICATION TECHNIQUE

BHARATH S. N.*[1]
Dr. Praveen kumar katigar[2]

[1]Research Scholar
[2]Supervisor, Assistant Professor, Dept. of Comp. Science and Engg , Proudhadevaraya Institute of Technology

Abstract: *Sentimental analysis or Opining mining is one of the hot research theme in the domain of text processing due to availability of huge amount user reviews in the internet. Due to the large availability of user's reviews in the internet, developing a huge database for the task of sentimental analysis is a tedious task. In reality, maintaining the uncompressed database is not practically advisable, on the other hand development of algorithms for sentimental analysis on compressed data is a challenging task. This paper addresses the sentimental analysis problem on the compressed text data. We employed a popular classifier like nearest neighbor classifier for classification of the user reviews. To test the effectiveness of our propose method we conducted experimentation on publically available polarity review dataset and also our own datasets. An extensive experimentation is carried out to demonstrate the effectiveness of the proposed method.*

Keyword : Nearest Neighbor Classification, Sentimental Classification

1 INTRODUCTION

Sentiment classification is one of the distinctive applications of text classification, which aims to evaluate the mood of the public about a particular product or topic. Sentiment classification, which is also called as opinion classification, involves in development of a system which automatically classify the polarity of opinions present in the internet portals like Amazon.com, IMDB, Blogs, Discussion forums, peer-to-peer networks and various types of social network sites.

Opinion classification, being a special application of text classification finds several applications in business issues, political issues, entertainment issues, for example, in business to understand the voice of customer as expressed in everyday communications, in politics to understand the opinions of voters about political candidates, in shopping to purchase the products, in entertainment, advertisement, government, research and development, education for e-Learning and Blog analysis. Review summarization and filtering flames for newsgroups.

In this paper, we are addressing a novel method for classifying the review texts as positive or negative. Text documents are the most common type of information store house especially with the increased popularity of the internet. User opinions present in the internet portals like Amazon.com, IMDB, Blogs, Discussion forums, peer-to-peer networks and various types of social network sites have millions or billions of text documents. The web pages that are available in the internet are stored in the compressed format. Opining mining activities on these data are carried out by decompressing the data and taking it back to the standard format. These processes of decompressing and performing mining activities consume more computational time. However to the best of our knowledge, nowhere in the literature we can find any works on opinion classification compressed text data.

Though the opinion mining seems to be very simple, it has several challenging aspects in opinion classification. The first is to determine whether a document or portion is subjective. Second challenge is that the difficulty lies in the richness of human language used i.e. people don't always express opinion in a same way. In order to arrive at sensible conclusions, analysis of the sentiment context has to understand. However, "The movie was great" is very different from "The movie was not great". In the more informal medium like twitter or blogs the more likely people are to combine different opinions in the same sentences which is easy for a human to understand but, more difficult for a computer to parse. In this paper, we are addressing a novel method for classifying the user reviews as positive or negative at the compression level. We have conducted experimentation on movie and product datasets and also publically available polarity movie review dataset. We examine the quantitative comparative analysis between document level and sentence level opinion classification.

The rest of the paper is organized as follows. In section 2 a brief literature survey on the text classification is presented. In section 3 we present the model based on compression technique. In section 4 we discuss about experimentation details and comparative analysis. The paper will be concluded in section 5.

2 LITERATURE REVIEW

In recent years many researchers have proposed many machine learning approaches for classification of the polarity of user opinions. Most of approaches can be classified into two major branches: document level and sentence level. In (Pang et al., 2002), authors have examined the effectiveness of applying machine learning techniques (Naïve Bayes, Maximum Entropy and Support Vector Machines) to the sentiment classification problem. In (Turney 2002), author presents an unsupervised learning algorithm (Semantic orientation) for classifying a review as recommended or not recommended. In (Pang et al., 2004), they proposed a novel machine learning method that applies text-categorization techniques based on subjective portions of the document. Extracting subjective portions can be implemented using efficient techniques for finding minimum cuts in graphs. In (Pang et al., 2005), they address the rating-inference problem: rather than just determine whether a review is "thumbs up" or not. In (Read 2005), author demonstrates that using emotion reduces the dependency of domain, topic and time for sentiment classification. In (Mishne 2005), author addresses the task of classifying blog posts by mood using SVM. In (Zhang et al., 2008), they proposed machine learning approach based on string kernel for sentiment classification for Chinese reviews. In (Bhuiyan et al., 2009), they present a state of art review of opinion mining from online customer feedback. In (Chen et al., 2009), they proposed NN based index which combines the advantages of machine learning and information

retrieval techniques. In (Alec et al., 2009), they have introduced a novel approach for automatically classifying the sentiments of Twitter messages using distant supervision. In (Li et al., 2010), they proposed a machine learning approach to incorporate polarity shifting information into a document-level sentiment classification system. In (Claster et al., 2010), they proposed a multi-knowledge based approach using Self Organizing Maps (SOM) and movie knowledge in order to model opinion across a multi-dimensional sentiment space. In (Zizka et al., 2011), they present machine learning approach to classifying the customer reviews. In (Jebaseeli et al., 2012), they provide an overall survey about sentiment analysis related to product reviews. In (Vinodhini et al., 2012), they presents a survey that covering the techniques and methods in sentiment analysis and challenges appear in the field. In (Mullen et al., 2004), they introduce an approach to sentiment analysis which uses SVM to bring together diverse sources of potentially pertinent information. In (Wiegand et al., 2010), they present a survey on role of negation in sentiment analysis. In (Das et al., 2009), they proposed a sentence level emotion identification using word level emotion classification. In (Zhai et al., 2011), they proposed semi supervised learning method to cluster product features for opinion mining. In (Pak et al., 2010), they focus on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. In (Wang et al.,2012), they present a simple model variant where an SVM is built over NB log-count ratios as feature values and get good results.

Most of the approaches present in the literature works on uncompressed user reviews. Whereas the challenging and required is to classify documents at compression level. In literature we can find many compression techniques which are used for the effective representation of data in compressed format. In this paper we consider only the lossless compression schemes. Run Length Encoding (RLE) is a simple and popular data compression algorithm. It is based on the idea to replace a long sequence of the same symbol by a shorter sequence. Huffman compression (Huffman 1952) it is a statistical lossless compression method that converts characters into variable length bit strings. Huffman compression technique works on frequency of individual symbol. The Huffman algorithm is a so-called "greedy" approach to solving this problem in the sense that at each step, the algorithm chooses the best available option. Lempel–Ziv– Welch (LZW) is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch. The LZW compression algorithm organized around a translation table or string table, that maps input characters into the fixed length codes [Ziv and Lempel 1977]. Among different compression techniques, we have used LZW compression technique. LZW compression is used as the foremost technique, mainly because of its versatility and simplicity. Typically, the LZW compression can compress executable code, text, and similar data files to almost onehalf of their original size. It usually uses single codes to replace strings of characters, thereby compressing the data. LZW also gives a good performance when extremely redundant data files are presented to it like computer source code, tabulated numbers and acquired signals. The common compression ratio for these cases is almost in the range of 5:1.Though RLE and Huffman compression techniques are also very simple; they are not suitable for text documents and also these two methods does not provide good compression ratio like LZW method.

3 PROPOSED METHODOLOGY

In this paper we are presenting a novel method used for sentimental classification of compressed user reviews. Normally text documents are available in several formats such as html, xhtml, pdf, plain text etc. The first step is to preprocess the text document, hence to bring them to a common format before processing the text. In the literature we have stop word elimination, stemming, pruning etc as pre-processing steps. In this work we have used only stop word elimination technique. Once the pre-processing is done on training data, the text documents are compressed using LZW compression scheme and a compressed training document library is created. The working principle of LZW compression technique is given as follows.LZW is a universal lossless compression algorithm which is organized around string table. String table contains strings that have been encountered previously in the text being compressed. It consists of a running sample of strings in the text, so the available strings reflect the statistics of the text. It uses greedy parsing algorithm, where the input string is examined character-serially on one pass, and the longest recognized input string is parsed off each time. A recognized string is one that exists in the string table. Each such added string is assigned a uniquely identified by code value. The proposed model is of two stages, in which stage one is of creation of database in which all pre-processed text data are compressed and stored in the database, stage two is classification stage in which given unknown sample is classified into its corresponding class label using compression technique.

Algorithm: LZW text compression.
Input: Pool of text data
Output: Pool of compressed text data, String table.
Method:

1. Initialize table to contain single character strings.
2. Prefix string ω ← Read first input character.
3. K← Read next input character
   If no such K (input exhausted) : code (ω) – output; EXIT
4. If ωK exists in string table : ωK- ω; repeat 3;
5. else ωK not in string table : code (ω) – output;
6. ωK – string table;
7. K – ω; repeat Step.

Algorithm end.

At each execution of the basic step an acceptable input string ω has been parsed off. The next character K is read and the extended string ωK is tested to see if it exists in the string table. For each training document we obtain a string table which is referred as dictionary representation and stored in the library. Further, given a test document we obtain dictionary representation and during classification we used string matching based neural networks classification technique. We classify the test document by adopting neural network classifier and class label will be assigned to the test document. The block diagram of the proposed model is as shown in fig 1.

Algorithm : Compression based sentimental analysis
Input : Compressed User Reviews
Method :

Step1 : Input collection of compressed user reviews.
Step 2: Apply text representation on compressed user reviews.
Step 3: Develop a database of compressed text documents (dictionary).
Step 4: Input a compressed user review for evaluation.
Step 5: Apply text representation on compressed user reviews.
Step 6: Apply nearest neighbor for classification of user reviews.
Step 8: Give a class label to the testing document.

Algorithm ends.

4 EXPERIMENTATION

To show the effectiveness of the proposed compression level sentimental classification approach, we have conducted two sets of experiments. In first set of experiments, we have used 40% of the database for training and remaining 60 % is used for testing. In second set of experiments, we have used 60 % training and 40 % for testing. For the purpose of evaluation of results, we have calculated precision, recall and f measure for each set of experimentation. The details of the experiments are shown in the following table 1.

4.1 DATASETS

Any system has to be tested to find its effectiveness and efficiency for which it has been designed. In this section we describe the dataset collection and data preparation stages that we have followed for creating our own datasets. Here is a detailed introduction to the datasets.

DATASET 1 Movie Reviews: This dataset contains two sets of movie reviews. The reviews are in .txt format and which are collected from IMDB review site.

DATASET 2 Product Reviews: This Product review contains reviews of four different products such as Mobile, Home appliances, Watch and Camera, which are collected from different review sites. Ex: Amazon, Epions etc. All reviews of this product are in .txt format.

DATASET 3 Polarity review movie dataset: This dataset is created by Bo Pang et al., in 2004. All reviews of this dataset were written before 2002, with a cap of 20 reviews per author (312 authors' total) per category. We collected this dataset through referring the Bo Pang et al., in 2004 paper. The reviews are in .txt format.

Table 1: Sentimental classification result table on different dataset using proposed model

DATASETS	40 : 60	60:40
Movie Review 1	0.7846	0.8083
Mobile Review	0.7973	0.8100
Product Reviews	0.7943	0.8022
Polarity Review Movie dataset	0.8022	0.8200

5. CONCLUSION

We have proposed a novel method for classifying user reviews at compression level. The proposed method uses compression scheme, string matching and nearest neighbor classification algorithm for sentimental classification technique. To check the efficiency and the robustness of the proposed models, an extensive experimentation is carried out on all the four dataset. The performance evaluation of the proposed method is carried out by performance measures such as precision, recall and f-measure. Even though, the results are not better than other uncompressed based techniques, they are comparatively equal to them, i.e., approximately 83% of classification accuracy. In this paper we have put forward a new representation model for text data for classification of user reviews using compression technique, which is first of its kind. Further, we explore novel proximity measures for comparing text in compressed format which may improve the classification accuracy.

References

A. NishaJebaseeli and E. Kirubakaran.:A Survey on Sentiment Analysis of (Product) Reviews.International Journal of Computer Applications,(2012) Vol. 47.

Alec Go, RichaBhavani, and Lei Huang.:Twitter Sentiment Classification using Distant Supervision.Processing, (2009) pp. 1-6.

Alexander Pak and Patrick Paroubek.:Twitter as a Corpus for Sentiment Analysis and Opinion Mining.In Proceedings of LREC, (2010).

Bo Pang and Lillian Lee. : Seeing Stars : Exploiting Class Relationships for Sentiment Categorization with respect to Rating Scale.In Proceedings of ACL, (2005).

Bo Pang and Lillian Lee.:A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum cuts.In Proceedings of the ACL,(2004) pp. 271-278. .

Bo Pang, Lillian Lee, and ShivakumarVaithyanathan.:Thumbs Up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),(2002), pp. 79-86.

Changli Zhang, WanliZuo, Tao Peng and Fengling He.: Sentiment Classification for ChineseReviews Using Machine Learning Methods Based on String Kernal. Third InternationalConference on Convergence and Hybrid Information Technology(ICCIT), IEEE, (2008).

ChangqinQuan and Fuji Ren,:Sentence Emotion Analysis and Recognition Based on Emotion Words Using Ren-CECps . In Proceedings of the International Journal of Advanced Intelligence, ,(2010),Volume 2, Number 1,pp. 105-117.

D. A. Huffman., 1952. A method for construction of minimum-redundancy codes. Proceeding of the Institute of Electrical and Radio Engineers, 40(9) pp. 1090 – 1101.

Dipankar Das and SivajiBandyopadhyay.: Sentence Level Emotion tagging. In Affective Computing and Intelligent Interaction and Workshops, 3rd International Conference on .IEEE,(2009) pp. 1-6,.

Fung, G.P.C, Yu, J.X, Lu. H, and Yu, P. S.: Text Classification without Negative ExampleRevisit. IEEE Transactions on Knowledge and Data Engineering.(2006)Volume 18, pp.23-47.

GiladMishne.:Experiments with Mood Classification in Blog Posts. In 1st Workshop on Stylistic Analysis Of Text For Information Access, (2005).

G Vinodhini and R. M. Chandrasekaran.: Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering.(2012) Vol. 2.

Jan Zizka and VadimRukavitsyn.:Automatic categorization of reviews and opinions of Internet e-shopping customers.Annual Conference on Innovations in Business andManagement, (2011).

Jin, Wei, Hung Hay Ho, and Rohini K. Srihari.: OpinionMiner: a novel machine learning system for web opinion mining and extraction.In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, (2009) pp. 1195-1204.

Jonathon Read.: Using Emoticons to reduce dependency in Machine Learning Techniquesfor Sentiment Classification.In proceedings of ACL, (2005).

Long-Sheng Chen and Hui-Ju Chiu.: Developing a Neural Network based Index for Sentiment Classification. In Proceedings of the International MultiConference ofEngineers andComputer Scientists(IMECS), (2009) Vol. I.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow and Andres Montoyo.: A Survey on the Role of Negation in Sentiment Analysis. In Proceedings of theWorkshop on Negation and Speculation in Natural Language Processing, (2010) pp. 60-68.

Peter D. Turney.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the ACL,(2002)pp. 417-424.

Rigutini, L.: Automatic Text Processing: Machine Learning Techniques. Ph.D. Thesis, University of Siena, (2004).

Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang and Guodang Zhou.: Sentiment Classification and Polarity Shifting. In Proceedings of the 23rdInternational conference on Computational Linguistics,(2010)pp. 635-643.

Tony Mullen and Nigel Collier.: Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In Proceedings of EMNLP,(2004) Vol. 4, pp. 412-418.

TouhidBhuiyan, YueXu and AudunJosang.:State-of-the-Art Review on Opinion Mining fromOnline Customers' Feedback. In Proceedings of the 9th Asia-Pacific Complex Systems Conference,(2009).

Wang, Sida, and Christopher D. Manning.: Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: (2012)Short Papers-Volume 2,pp. 90-94.

Wiebe, Janyce, and Ellen Riloff,.:Creating subjective and objective sentence classifiers fromunannotated texts. Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg,(2005)pp486-497.

William B. Claster, DINH QuocHung and SubanaShanmuganathan.: Unsupervised ArtificalNeural Nets for Modeling Sentiment. Second International Conference on Computational Intelligence, Communication systems and Networks (IEEE), (2010).

ZhongwuZhai, Bing Liu, HuaXu and PeifaJia,.:Clustering Product Features for Opiniong Mining. In Proceedings of the 4th ACM International Conference on Web Searchand Data Mining,(2011) pp. 347-354.

Ziv. J and Lempel A., 1977. A Universal Algorithm for Sequential Data Compression, IEEE Transactions on Information Theory 23 (3), pp. 337–342.

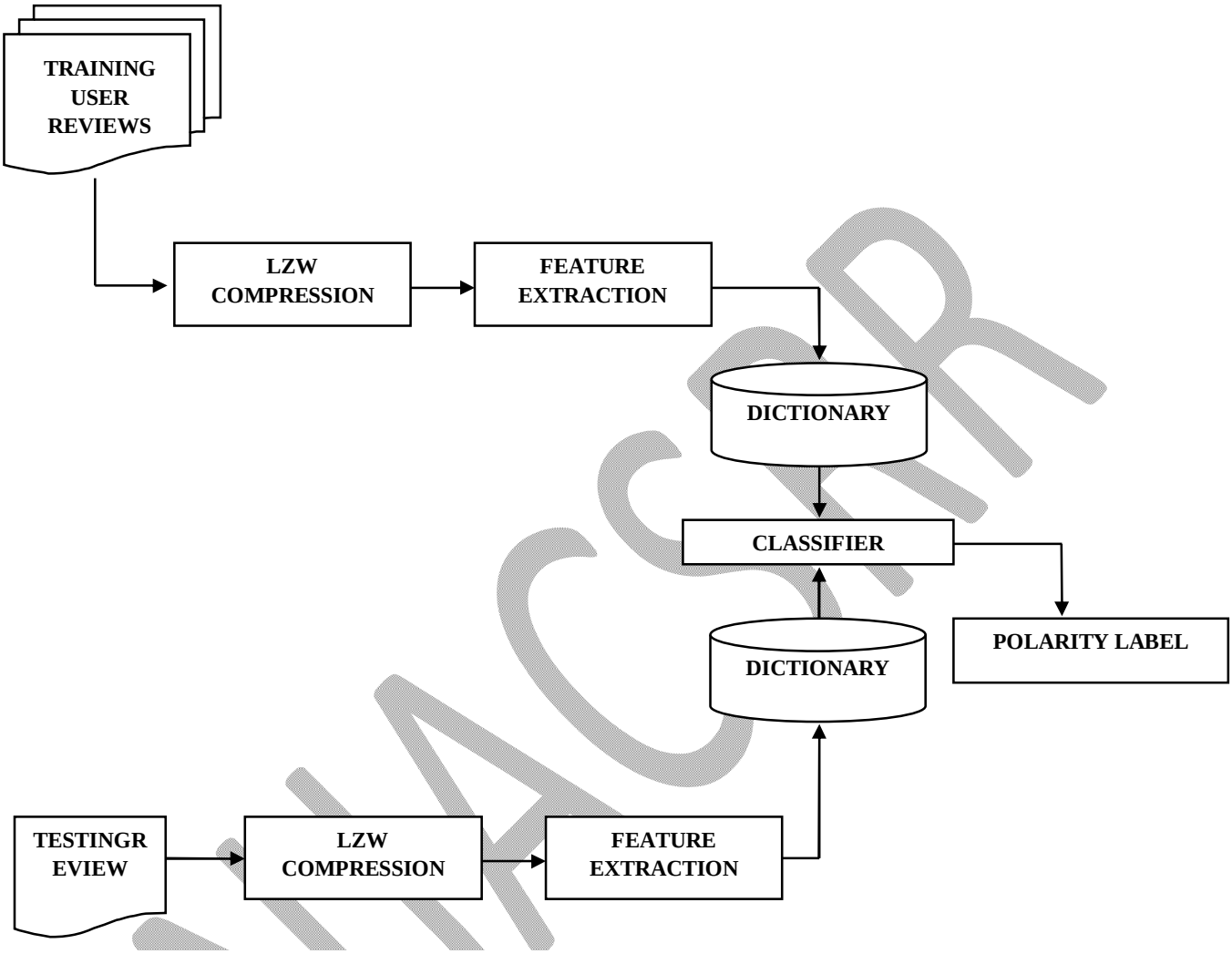

Fig 1: Block diagram of document level opinion classification

