Multi –Domain Sentiment Classification Approach using Supervised Learning

Ms. Teena Batham[1]
M.Tech. Research Scholar, Department of CSE,
BITS (Bhopal)

Prof. Nidhi Dubey[2]
Assistant Prof., Department of CSE,
BITS (Bhopal)

---***---

Abstract— The amount of digital material available on the Internet is growing every day. As a result, the demand for tools to assist people in accessing and analyzing all of these information is increasing. Text classification, in particular, has shown to be extremely beneficial in terms of information management. The practice of classifying natural language text into one or more groups based on its content is known as text classification. It entails classifying textual opinions into categories such as "good" and "negative." The polarity (positive or negative) of a statement was the focus of previous Sentiment Analysis research. This is classified as binary classification, which divides a set of components into two categories. The goal of this study is to look into Multi Class Sentiment Classification, which is a novel method to sentiment analysis. Different classification models like Bayesian Classifier, Random Forest and SGD classifier are taken into consideration for classifying the data and their results are compared. Frameworks like Weka, Apache Mahout and Scikit are used for building the classifiers

1. Introduction

With new technology trends, binary valued information, also known as digital information, is growing at a rapid rate, and human involvement is also increasing for resource utilization. However, it is under the technical and effort resistivity of humans to analyze and fetch meaningful information from these techniques, necessitating various automation methods that can work quickly and errorless and extract meaningful data for analysis and decision making for any system. Text categorization is also a method for assessing data and generating useful information from any given or gathered survey pile. It is a machine-learning procedure, in which the text is collected into multiple categories by different classification algorithms. Every classifier has its own way of gathering the features and using them in order to classify the text. Text classification has big range of applications like spam filtering, genre categorization, language identification, routing the emails, sentiment analysis and many such applications. All these applications use wide range of text classification techniques. This project's data set is a corpus of online e-commerce websites. To divide the data into train and test sets, some pre-processing operations are completed. For the classification job, many characteristics such as stop-words, stemmers, n-grams, and parts of speech tagging were utilized on this dataset. On this dataset, Bayesian, Random Forest, and Stochastic Gradient Descent models were used to classify the reviews into different labels.

The following is a breakdown of the work. Section 2 describes the background study that was done to investigate various methodologies and models. Section 3 discusses the problem and the dataset structure. The recommended approach is outlined in Section 4, and the paper is eventually finished in Section 5.

2. Literature Review

In 2013 ,Liu.B et al [1] to develop a Naive Bayes Classifier on top of the Hadoop framework for sentiment classification allows them fine-grained control over the algorithm to accommodate large scalability data.

In 2014, Rachna Mishra et al [2] explored a variety of classifiers and classification algorithms. Also specifies a number of measurements and factors that are important in spam categorization. Shows an analysis of various supervised classifiers utilizing various data mining tools such as Weak and Rapid Miner.

In 2013, Taifeng et al [3]], proposed a new perspective of looking at click prediction. They primarily focused on the factors that influence a user's decision to click on a particular. Word Clouds are the most effective technique of visualizing text since the frequency of the words in the text is related to the font size. This is usually done in a static manner, which means that the summarizing is done for static text

In 2014, Florian Heimerl et al [4], The word cloud explorer, a prototype system that incorporates natural language processing techniques, was developed. And it completely relies on word clouds as a visualization tool. This system also includes search, click-based filtering, part-of-speech filtering, a stop word editor, and a co-occurrence cloud.

In 2013 , Akaichi et al [5] , The existing research works tend to identify the state of mind of users but are insufficient because of the ambiguity in the conveyed text. For their research, the writers. . A method that depends on Naive Bayes and SVM is proposed by them and several lexicons related to emoticons, interjections have been built to determine the sentiment of status updates.

In 2011, Haruna Isah , et al [6] , were involved in developing a framework for gathering and analyzing the user views using machine learning, text mining and sentiment analysis. The proposed framework was evaluated on Facebook comments and data from Twitter

In 2013, Abdullah et al [7] , is to extract fundamental information about stocks from news sources and use it in stock market analysis. They reviewed previous business studies and devised a system that includes a text parser and an analyst. The system is evaluated, and this equipped framework is able to analyze and forecast the decisions from any data source.

In 2013, Neethu et al [8], proposed a classification system that uses the natural language processing techniques and k- means clustering algorithm for categorizing the research papers.

In 2010, Jain et al [16], The authors first determine if the sentence is polar or neutral in this method. In the next level disambiguation is removed from polar sentences by classifying them into positive, negative or both.

In 2012, Esmin et al [17], followed three level hierarchical strategies for finding the sentiment of twitter data. In each and every level, the emotion is found and it is further fine-tuned in its lower level.

In 2014, Li et al [18], Training the data in multiple layers with different set of features in every layer is discussed. The performances at every layer are discussed and compared

In 2008, Fan et al [19], discusses a new strategy which finds the sentiment of text based on the results from multiple levels. Text is initially divided into parts and the sentiment is found for every part. Integrating the sentiment results of every part determines the sentiment of the actual text.

3. PROPOSED WORK

The process flow that has been followed to solve this problem of multi class sentiment classification is shown in the below picture.

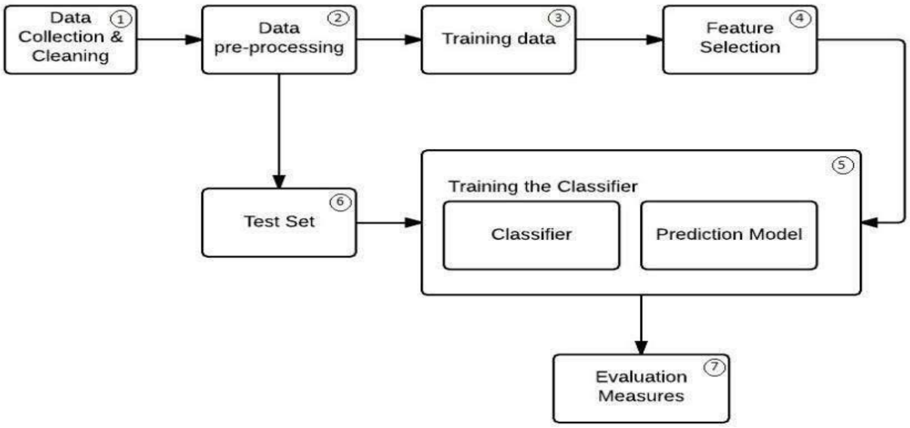

Figure 1: Flowchart of Proposed Work

In the first phase, data is collected and checked for any data outliers. During the initial phase, all such noise is removed. The data is transformed during the pre-processing phase so that the machine learning algorithms can work directly on the processed data. During this phase of the project, the train and test sets are separated. Following the acquisition of the train set, features are chosen from it during the Feature Selection phase. Words are the features of this dataset, and several techniques such as stop-word removal, stemming, and n-grams are used during this phase. After that, the classifier is implemented in such a way that it adapts the Predictive Model, which has multiple phases. The classifier is then trained using the train set, and the test set is used to evaluate the classifier.

4. RESULT

5. CONCLUSION

This study proposes a novel method for dealing with multi-class sentiment classification problems. In general, a classification model performs better when the number of labels in the dataset is limited to two or three. This paper suggests a classification model for dealing with multiple labels. This model operates in phases, with the sentiment of the data instance fine-tuned at each stage. This work can be used for other domains that involve classification problems by making some adjustments to the multi-tier predictive model and by using of various contexts specific lexicons.

6. REFERENCES

1) Liu, B., Blasch, E., Chen, Y., Shen, D. and Chen, G. (2013) ‘Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier’, IEEE International Conference on Big Data, 2013.

2) Rachana Mishra., Ramjeevan Singh Thakur. (2014), 'An Efficient Approach for Supervised Learning Algorithms using Different Data Mining Tools for Spam Categorization', Fourth International Conference on Communication Systems and Network Technologies, 2014.

3) Taifeng Wang., Jiang Bian., Shusen Liu., Yuyu Zhang., Tie-Yan Liu. (2013), ' Psychological Advertising: Exploring User Psychology for Click Prediction in Sponsored Search', 2013.

4) Florian Heimerl, Steffen Lohmann, Simon Lange, Thomas Ertl. (2014), ' Word Cloud Explorer: Text Analytics based on Word Clouds', 47th Hawaii International Conference on System Science, 2014.

5) Akaichi, J. (2013) ‘Social Networks’ Facebook' Statutes Updates Mining for Sentiment Classification’, International Conference on Social Computing, 2013.

6) Haruna Isah, Paul Trundle, Daniel Neagu, ' Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis', 2011.

7) Abdullah, S. S., Rahaman, M. S. and Rahman, M. S. (2013) ‘Analysis of stock market using text mining and natural language processes processing’, International Conference on Informatics, Electronics and Vision (ICIEV), 2013.

8) Neethu, M. and Rajasree (2013) ‘Sentiment analysis in twitter using machine learning techniques’, Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013.

9) Patel, B. and Shah, D. (2013) ‘Significance of stop word elimination in meta search engine’, International Conference on Intelligent Systems and Signal Processing (ISSP), 2013.

10) Silva, C. and Ribeiro, B. (2003) ‘The importance of stop word removal on recall values in text categorization’, Proceedings of the International Joint Conference on Neural Networks, 2003.

11) Harrag, F., El-Qawasmah, E. and Salman, A. M. S. Al- (2011) ‘Stemming as a feature reduction technique for Arabic Text Categorization’, 10th International Symposium on Programming and Systems, 2011.

12) Warintarawej, P., Laurent, A., Pompidor, P. and Laurent, B. (2010) ‘Classification of brand names based on n-grams’, International Conference of Soft Computing and Pattern Recognition, 2010.

13) Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K. and Caro, J. (2013) ‘Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning’, IISA 2013.

14) Xue, D. and Li, F. (2015) 'Research of Text Categorization Model based on Random Forests', IEEE International Conference on Computational Intelligence & Communication Technology, 2015.

15) Zhang, Y., Wang, C., Xiao, B. and Shi, C. (2012) 'A New Method for Text Verification Based on Random Forests', International Conference on Frontiers in Handwriting Recognition, 2012

16) Jain, T. I. and Nemade, D. (2010) 'Recognizing Contextual Polarity in Phrase- Level Sentiment Analysis', International Journal of Computer Applications, 7(5), pp. 12–21, 2010.

17) Esmin, A. A. A., Roberto L. De Oliveira Jr. and Matwin, S. (2012) 'Hierarchical Classification Approach to Emotion Recognition in Twitter', 11th International Conference on Machine Learning and Applications, 2012.

18) Li, J., Fong, S., Zhuang, Y. and Khoury, R. (2014) 'Hierarchical Classification in Text Mining for Sentiment Analysis', International Conference on Soft Computing and Machine Intelligence, 2014.

19) Fan, N., Cai, W. and Zhao, Y. (2008) 'Research on the Model of Multiple Levels for Determining Sentiment of Text', IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008.