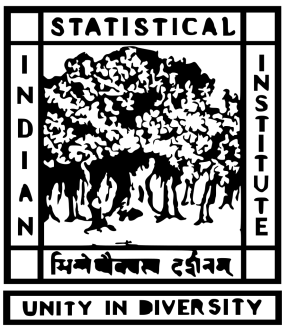
Summer Training Report on

Text Mining and Sentiment Analysis

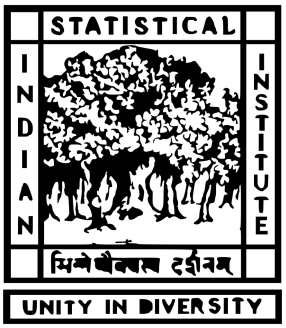


Submitted by:
Yash Katiyar
Department of Computer Science and Engineering
Indian Institute of Technology, Jammu
Email: *2017ucs0060@iitjammu.ac.in*

Under the guidance of: **Prof. S.K. Neogy**
Head, Indian Statistical Institute, Delhi Centre

July 17, 2019

Organisation Profile



Indian Statistical Institute, Delhi Centre

The Indian Statistical Institute was founded by Professor P.C. Mahalanobis in Kolkata on 17th December, 1931. The institute gained the status of an Institution of National Importance by an act of the Indian Parliament in 1959. The Delhi Centre of the Indian Statistical Institute was started in 1974, and shifted to its present campus in 1975. Some of India's leading statisticians, mathematicians, and economists have been on its faculty, and its training programs enjoy worldwide repute.The Delhi campus offers two master level courses Master of Statistics (M. Stat) and Master of Science (M. S.) in Quantitative Economics, and doctoral programs.The current head of the Indian Statistical Institute, Delhi Center is **Professor Samir Kumar Neogy**.

Acknowledgement

I would like to express my special thanks of gratitude to my **Professor Dr. Samir Kumar Neogy** (Head, ISI Delhi Center) who gave me the golden opportunity to do this wonderful project on the topic ***Text Mining and Sentiment Analysis***, which also helped me in doing a lot of Research and I came to know about so many new things. He has been very supportive and patient while guiding me and I would like to appreciate his cooperation and thank him for providing me this internship opportunity. These were two wonderful months where I learned not only about the topic of this project but also the importance of time management and guidance in life.
Secondly i would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Contents

1 Abstract

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down.

Sentiment Analysis is a method for judging somebody's sentiment or feeling with respect to a specific thing. It is utilized to recognize and arrange the sentiments communicated in writings. The web-based social networking sites like twitter draws in a huge number of clients that are online for imparting their insights in the form of tweets or comments. The tweets can be then classified into positive, negative, or neutral. In the proposed work, logistic regression classification is used as a classifier and unigram as a feature vector. For accuracy, k fold cross validation data mining technique is used. For choosing precise training sample, tweet subjectivity is utilized.

2 Introduction

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. For the first time in human history, we now have a huge volume of opinionated data recorded in digital form for analysis. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations. Large organizations employ sentiment analysis to get an overall review of their services or products and hence decide their strategy ahead.

2.1 Problem Statement

We have certain reviews as input data. The reviews are divided into 3 categories :
1. Positive reviews
2. Negative reviews
3. Unlabeled reviews

We have to design an algorithm and then implement it to a program so that the program can predict the sentiment of an unlabeled review with maximum accuracy. The reviews are of electronic devices from Amazon. The data has been taken from John Hopkins University website :
(https://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html)
and it is commonly called **Multi-Domain Sentiment Dataset**.

2.2 Objectives

The main objective of our algorithm is to clean and pre-process the raw data (which is in xml format) from the dataset to get meaningful data. Then, we need to employ data visualization techniques to convert data into mathematically presentable form.

Once we have finished processing the data, we have to design and implement an algorithm which employs machine learning techniques and natural language processing to build a sentiment analyzer

that learns on the processed labeled data and hence can now predict the sentiment of new unlabeled data/reviews. This algorithm once tested can be exported to various websites or apps and can thus give real-time analysis of reviews to give the sentiments of newly written reviews

3 Sentiment Analysis Techniques

The following techniques are the general approaches for analyzing sentiments:

3.1 Manuscript based

Manuscript level sentiment analysis can use any machine learning approach (supervised or unsupervised) to analyze any sentiment. Another intriguing strategy decides the PMI of expressions as positive or negative just to register the polarity of an expression.

3.2 Sentence based

Sentence based analysis divides the sentence into smaller similar phrases. This type of analysis is useful to deal with special type of sentences like conditional, negation sentences.

3.3 Aspect-based

Aspect-based analysis of sentiment defines that a product has many aspects or features or properties which in result have differ sentiments.

3.4 Lexicon based sentiment analysis

Lexicon based sentiment analysis is an attractive research area that uses methods such as WordNet distance to label the sentiment as fine and dire.

4 Classifiers

Following basic classifiers are considered.

4.1 Naive Bayes:

It is a family of algorithm that is used to construct classifier that assigns labels to the instances of problems.

4.2 Support Vector Machines (SVM):

SVM is used for classification purpose. It can also be used in regression analysis and outlier detection. It constructs a hyper-lane in high dimensionality space.

4.3 Logistic Regression:

It is used to determine the output or result when there are one or more than one independent variables. The output value can be in form of 0 or 1 i.e. in binary form.

5 Algorithm and Implementation

The code for this report has been made on **Python**.

5.1 Dataset Description

The reviews have been collected by John Hopkins University. These are reviews from a very famous website **AMAZON**. The reviews are from categories: **books, dvds, electronics, kitchen & housewares**. The reviews are in the form of xml files since they have been collected from the website. There are 3 files in the dataset :

1. positive : review file with 4000 positive reviews from respective customers
2. negative : review file with 4000 negative reviews from respective customers
3. unlabeled : review file containing unlabeled data for testing purpose

Additional data : We also have a text file named ***stopwords.txt***. It contains the basic words that don't contribute much to the sentiment of a review. Such words can be removed from the review for better predictions.

5.2 Approach

5.2.1 Data extraction

Since we have xml files, therefore we need a tool to extract only the relevant information from it. We use ***BeautifulSoup*** library which helps us to extract text within certain xml tags. Hence we extract the text from <review_text>tag to continue with further process.

5.2.2 Data preparation

Now we have review text in the form of list of sentences. Now the data is prepared for finally giving it to the machine learning model. Firstly the review is converted into lowercase letters. Then, the preparation takes place in the following steps:

1. Tokenization : Each review is broken into its individual words. This is achieved with the help of ***tokenize*** function of ***nltk***(natural language toolkit) library. These words are often referred to as tokens.

2. Lemmatization : Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. All the tokens are reduced to their root words after morphological analysis thereby reducing the total set of words. This is achieved with the help of ***WordNetLemmatizer*** from ***nltk.stem***

3. Removal of unnecessary stopwords : The unnecessary words such as the articles, pronouns, etc. which do not contribute heavily to the sentiment of a review are removed in order to clean the data even more.

4. Count Vectorization : We count the number of occurrences of each token in a certain review and then normalize it with the total number of words in that review thereby giving a value between 0 and 1. We convert this into a vector of numbers each corresponding to certain token in a set of tokens with each unique occurrence.

5. Final data formulation : We finally construct a 2-D matrix where each row contains a unique review and each column represents the count-vectorization of a unique word in the set of tokens/words. The last column contains the label : **0** for *negative review*, **1** for *positive review*. This final matrix is fed into the machine learning model for training purpose.

5.2.3 Training and Testing

Since we did not have separate test dataset, hence we split the data into train set of 3500 reviews and test set of 500 reviews after randomly shuffling the data. Then we fed this into the classifiers as discussed above. We trained each of the models/classifiers with the train set and tested their accuracy. ***sklearn***(scikit-learn) library was used to achieve this. It contains all the 3 classifiers and many others. This saves time since we don't have to make our own models for these classifiers.

5.3 Conclusion

The following classifiers were used for sentiment analysis :

1. Naive Bayes : The multinomial naive bayes classifier was the fastest and marginally the most accurate classifier of the three. It took the least time to train the model and after several shuffles gave accuracy approximately around 80%. This proved to be the most effective method here.

2. Logistic Regression : The Logistic Regression model was fast enough but slower than the Naive Bayes model. It gave accuracy varying between 65%-75%. However, this model was better for visualization as we can see how a bias weight is associated with each token,i.e., how much a word/token is affecting the sentiment of a review and in what sense (positive or negative).

3. Support Vector Machines(SVM) : This proved to be the slowest classifier of all. Where the other two classifiers took less than 30 seconds to train on a relatively small dataset, SVM took about 20-30 minutes in training the model for the same train set. The accuracy was also disappointing at rate of about 55%-60%

5.4 Inference

As it is observed these days, that many individuals' posts surveys with respect to any item, movie, game or occasion via web-based networking media stages. For this, it is essential for the organizations to define particular sentiments of such surveys keeping in mind the end goal to realize that what individuals think about the item. The projected method utilizes one such stage called twitter to play out the sentiment categorization. The info is taken as tweets in the wake of verifying the client. The current framework has utilized Naive Bayes and Logistic Regression classifiers to classify the sentiments. For highlighting features Logistic Regression is used and for defining labels Effective score of a word is utilized.

6 Experience and Future

A lot of experience was gained by this internship on the topic of Sentiment Analysis and Text Mining. Both are kind of analogous with respect to this report. The enthusiasm to learn something new motivated me to learn new things in the field of machine learning and natural language processing.I learned how with the help of a handful of data one can change the world. As a complete project, I wish to take this project further. I wish to employ newer, faster and more accurate techniques to improve upon the project. I would also make a finished product out of it in the form of a user-interface where one can just copy paste a review and know the sentiment of it without reading whole of it. This can also be developed further to give a star rating out of 5 to a review to help the customers more. Currently I am working on **LSTM(Long Short-Term Memory)** algorithm for the sentiment analysis which is a much better model for the same.

7 References and Citations

References

[1] Abhilasha Tyagi, Naresh Sharma : *Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic* in **International journal of Engineering and Technology**, 7 (2.24) (2018) 20-23

[2] Multi-Domain Sentiment Dataset
`https://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html`

[3] *What is Text Mining?* by **Marti Hearst** from **SIMS, UC Berkeley**

[4] `https://en.wikipedia.org/wiki/Lemmatisation`

[5] ***Sentiment Analysis and Opinion Mining*** by Morgan & Claypool
`https://ieeexplore.ieee.org/document/6812968`

[6] `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`