Sentiment Analysis of Twitter Data Using Logistic Regression

Raju Shrestha

Roll No. 48/2079

INTRODUCTION

- Sentiment is defined as a view or opinion that is expressed. It is a feeling of someone that he/she expresses either in textual or verbal form.
- A sentiment can be defined as a personal positive or negative feeling.

PROBLEM STATEMENT

- Every day millions of data is being collected on social media like twitter which contains people opinion about many things like the product and services they use, political and religious views etc.
- The collected data is unstructured and not organized in a pre-defined manner.

OBJECTIVES

- The main objective of this report is to classify the tweets into positive or negative using logistic regression.

ALGORITHM USED

- Logistic Regression is used for classifying the tweets into two classes.
- It is predictive analysis model based on binary classification. It classify the tweets based on the probability given to tweets belong to that particular class.
- To predict the tweets into positive or negative. I have used label dataset with probability value 0 for negative and 1 for the positive tweets.

LITERATURE REVIEW

- Sentiment analysis have become the growing area in the field of natural language processing. Supervised machine learning algorithm like Logistic Regression algorithm plays vital role in the sentiment analysis. There are many researched carried out for sentiment analysis.
- Pang, Lee, & Vaithyanathan, 2002) Studied various technique for sentiment analysis for the movies review. They compare the different classification algorithm like Naive Bayes classification, Maximum Entropy classification and Support Vector Machine. They also consider the different factor affecting the sentiment like unigrams, bigrams, Part of speech (POS) etc. They achieved accuracy of above 80% for all three algorithm using unigrams + bigrams.

LITERATURE REVIEW…

- (Waykar, Wadhwani, More, & Kollu, 2016) have focused mainly on the Naive Bayes classifier. They take the baseline for their research as (Pang, Lee, & Vaithyanathan, 2002). They display the result on pie chart for positive, negative and neutral for the specific keyword.
- (S.T, Wikarsa, MComp, Turang, & MKom, 2016) have focused on topic based classification based on the Logistic Regression. They also have used the confusion matrix as a classifier model. They achieved the accuracy of 92% for the tweets classification into selected topics.

LITERATURE REVIEW…

- (Tyagi & Sharma, 2018) have proposed research based on Logistic Regression. They have used Logistic Regression as classifier and unigram as a features vectors. For increasing the accuracy K-fold cross validation and tweet subjectivity is used. To further speed up the classification process they also use the idea of effective word score heuristics that find out the polarity score of the words which are frequently used.

METHODOLOGY

- For sentiment analysis supervised machine learning algorithm i.e logistic regression is used. Logistic Regression requires labeled data for training the classifier.

Data Collection:

- Data used for sentiment analysis was collected from the publically available source i.e from kaggle.com.
- The dataset consists of 1,00,000 training and 15,034 testing data in csv format and is labeled 0 for negative and 1 for positive.

METHODOLOGY…

Data Preprocessing:

- The twitter data consist of different properties in which most of it is not useful for sentiment analysis.

Data preprocessing includes various step:

- Usernames: Twitter consists of username which consist of symbol @ at the beginning.eg @sparkingroshan. It is replaced by the word AT_USER in data sets which is started by @ in the datasets.
- Usages Link: User includes the link in the tweets for the more detail information which is not useful for sentiment analysis. The link is replaced by the word ‘URL’.

METHODOLOGY…

- Stop Words: Stop word are those filler words which are not useful for sentiment analysis. These words includes most repeated word like a, an, the, for, etc. These words does not give any sentiment hence they are filtered out form the datasets.
- Removing Hash-tags: Hash-tag in a tweet is used to emphasize a particular word or sentence, for example #thisisgood. Removal of these hash-tags is important because these hash-tags do not define any sentiment.

METHODOLOGY…

- Repeated Letters: Tweets contain the very causal language. So the word such as hurrayyy is replaced with actual word hurray. The letter repeated more time is reduced to one.
- Stemming: Change a word in the text into its base term or root term. Example, happiness to happy

METHODOLOGY…

Feature Extraction:

- After preprocessing the tweets, tweets is converted into feature vector. Feature vector is used for building the model and is used to train the model which is further used to classify the unseen data.
- Feature vector is the n-dimensional vector of numerical value that represent the some object.
- The tweets in training data is split into words and each words into feature words.

METHODOLOGY…

- The feature words may consist of words unigram or bigrams. This report consider unigram as feature words. For eg. This is the ball is represented as this, is, the, ball as unigrams.
- The entire feature vector will be the combination of each of this feature words.

IMPLEMENTATION & TESTING

- The algorithm is implemented using Python along with NLTK, Scikit-learn library.
- Among the total data 85% of the data is used for training and 15% is used for testing.

CONCLUSION

- Tweets Sentiment Analysis Using Logistic Regression Algorithm was successfully implemented using python programming language.
- The accuracy is of the model is 77.2% which is quite low and can be improve by providing more datasets.

REFERENCES

- [1] Go, A., Bhayani, R., & Huang, L. (n.d.). Twitter Sentiment Classification using Distant Supervision.
- [2] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning on Empirical Methods in Natural Language Processing (EMNLP).
- [3] Ravikiranj. (2012, May 8). how to build a twitter sentiment analyzer ? Retrieved from https://ravikiranj.net/posts/2012/code/how-build-twitter-sentimentanalyzer/
- [4] S.T, I., Wikarsa, L., MComp, B., Turang, R., & MKom, S. (2016). Using Logistic Regression Method to Classify Tweets into the Selected Topics. ICACSIS.
- [5] Tyagi, A., & Sharma, N. (2018). Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic. International Journal of Engineering & Technology.
- [6] Waykar, P., Wadhwani, K., More, P., & Kollu, A. (2016). Sentiment Analysis of Twitter tweets using supervised classification technique. Int. Journal of Engineering Research and Applications.

Thank You