Tribhuvan University

Institute of Science and Technology

Seminar Report

On

"Sentiment Analysis of Twitter Data Using Logistic Regression"

Submitted to

Central Department of Computer Science and Information Technology

Tribhuvan University, Kirtipur

Kathmandu, Nepal

In the partial fulfilment of the requirement for Master's Degree in Computer Science and Information Technology (M. Sc. CSIT) First Semester

Submitted by

Raju Shrestha

Roll No. 48/2079

June, 2023

Tribhuvan University

Institute of Science and Technology

Supervisor's Recommendation

This is to certify that Mr. Raju Shrestha has submitted the seminar report on the topic **"Sentiment Analysis of Twitter Data Using Logistic Regression"** for the partial fulfilment of Master's of Science in Computer Science and Information Technology, first semester. I hereby, declare that this seminar report has been approved.

_____

Supervisor

Asst. Prof. Mr. Bikash Balami

Central Department of Computer Science and Information Technology

Letter of Approval

This is to certify that the seminar report prepared by Mr. Raju Shrestha entitled **"Sentiment Analysis of Twitter Data Using Logistic Regression"** in partial fulfilment of the requirements for the degree of Master's of Science in Computer Science and Information Technology has been well studied. In our opinion, it is satisfactory in the scope and quality as a project for the required degree.

Evaluation Committee

……..……..	……..……
Asst. Prof. Sarbin Sayami	Asst. Prof. Bikash Balami
(H.O.D)	(Supervisor)
Central Department of Computer Science and Information Technology	Central Department of Computer Science and Information Technology

……

(Internal)

Acknowledgement

The success and final outcome of this report required a lot of guidance and assistance from many people and I am very fortunate to have got this all along the completion. I am very glad to express my deepest sense of gratitude and sincere thanks to my highly respected and esteemed supervisor **Asst. Prof. Bikash Balami,** Central Department of Computer Science and Information Technology for his valuable supervision, guidance, encouragement, and support for completing this paper.

I am also thankful to **Asst. Prof. Sarbin Sayami**, HOD of Central Department of Computer Science and Information Technology for his constant support throughout the period. Furthermore, with immense pleasure, I submit by deepest gratitude to the Central Department of Computer Science and Information Technology, Tribhuvan University, and all the faculty members of CDCSIT for providing the platform to explore the knowledge of interest. At the end I would like to express my sincere thanks to all my friends and others who helped me directly or indirectly.

Raju Shrestha(48/2079)

Abstract

Sentiment Analysis (also known as opinion mining or emotion AI) is a method for judging somebody's sentiment or feeling with respect to a specific thing written in a piece of text. It is used 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.

Keywords: Sentiment analysis, Opinion mining, Text classification, Unigrams, Polarity, Machine learning, Logistic regression, Natural Language Processing.

Table of Contents

List of Figures

List of Abbreviations

API: Application Programming Interface

CSS: Cascading Style sheet

CSV: Comma Separated Values

HTML: Hypertext Markup Language

NLTK: Natural Language Toolkit

POS: Part of Speech

URL: Uniform Resource Locator

Chapter 1: Introduction

1.1 Overview

In the present days, Micro blogging has become a very popular communication tool among Internet users. Many users share tweets or messages everyday on prevalent sites, for example,

Twitter and Facebook. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy

accessibility of micro blogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to micro blogging services. As more and more users post about products and services they use, or express their political and religious views, micro blogging websites become valuable sources of people's opinions and sentiments. Such data can be efficiently used in research, business or social science.

Sentiment is positive or negative reviews about product or on any topics. We people can identify tweets by reading whether it is positive or negative. But if there is huge data to be read then it would be tedious and time consuming. So, if all this process could be done with the help of automated program then it would be easier and above manual process could be eliminated.

Sentiment Analysis of Twitter Data Using Logistic Regression is a web-based application which takes tweets as an input and gives sentiment value as an output.

1.1.1 Defining Sentiment

A 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. For example: This is the best budget smartphone. This is positive sentiment. And, this phone have bad resolution is consider as negative sentiment.

1.1.2. Characteristic of Tweets

Twitter message have many unique attributes (Go, Bhayani, & Huang) which are as follows: Tweets and Length: Tweets are the status posted by user which is of 140 length.

Username: Username in twitter is started with @followed by text and number. Eg @barackobama

1.2 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. And, the data is unstructured and not organized in a pre-defined manner. These text are usually difficult, time-consuming and expensive to analyze, understand, and sort through.

Many company wants to know how positive or negative peoples are about their product and services. People want to know other people how much positive or negatives tweets he/she tweet. Tweets Sentiment Analysis Using Logistic Regression Algorithm will provide the positive or negative sentiment on tweets that people have tweeted.

1.3 Objective

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

1.4 Logistic Regression

The algorithm used is Logistic Regression. Logistic Regression 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 and negative. I have used label dataset with probability value 0 for negative and 1 for the positive tweets.

Logistic regression, use a Logistic function, for instance, a sigmoid function to estimate probabilities between positive or negative label and data features.

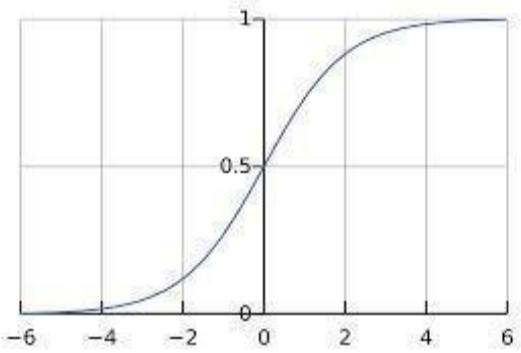

Figure 1 - Sigmoid Function

Logistic Regression is a discriminative model which means computing P(y|x) by discriminating among the different possible values of the class y based the given input x. The equation for this is as shown below:

$$P(c|x) = \sum_{i=1}^{N} w_i . f_i$$

To generate a value of P(y|x) of an output that is in between value 0 and 1, the following exp function is used:

$$P(c|x) = \frac{1}{z} \exp \sum_i w_i . f_i$$

To change the normalization factor Z and specify the number of features as N is as follows:

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i . f_i)}{\sum_c \exp(\sum_{i=1}^{N} w_i . f_i)}$$

The final equation for computing the probability of y being of class c given x is:

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i . f_i(c,x))}{\sum_{c' \epsilon C} \exp(\sum_{i=1}^{N} w_i . f_{i(c',x)})}$$

Chapter 2: Literature Review

Sentiment analysis have become the growing area in the 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.

(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.

(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.

Supervised machine learning classifier required the trained data set to work. For this I have used publicly available labeled dataset.

Chapter 3: Methodology

This report uses supervised machine learning algorithm which is Logistic Regression. Logistic Regression requires labeled data for training the classifier.

3.1 Data Collection

Data used by the Sentiment Analysis of Twitter Data Using Logistic Regression Algorithm was collected from the publically available data which is already being placed by other researchers. It is collected from the (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. So this project uses only the positive and negative datasets.

3.2 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.

1. 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.

2. 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'.

3. Stop Words: Stop word are those filler words which are not useful in for sentiment. 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.

4. 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. Thus pre-processing is done and hash-tags before any word are removed.

5. Repeated Letters: Tweets contain the very causal language (Waykar, Wadhwani, Pooja, & Kollu, 2016) so the word such as hurrayyy is replaced with actual word hurray. The letter repeated more time reduced to the one.

6. Stemming: Change a word in the text into its base term or root term. Example, happiness to happy.

3.3 Feature Extraction

After preprocessing the tweets, tweets is converted into feature vector. Feature vector are the most important concept in implementing classifier (ravikiranj, 2012). 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 features that represent the some object. In tweets we can consider the presence or absence of words that appear in the tweets. The tweets in training data is split into words and each words into feature words. The feature words may consist of words unigram or bigrams. This project 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.

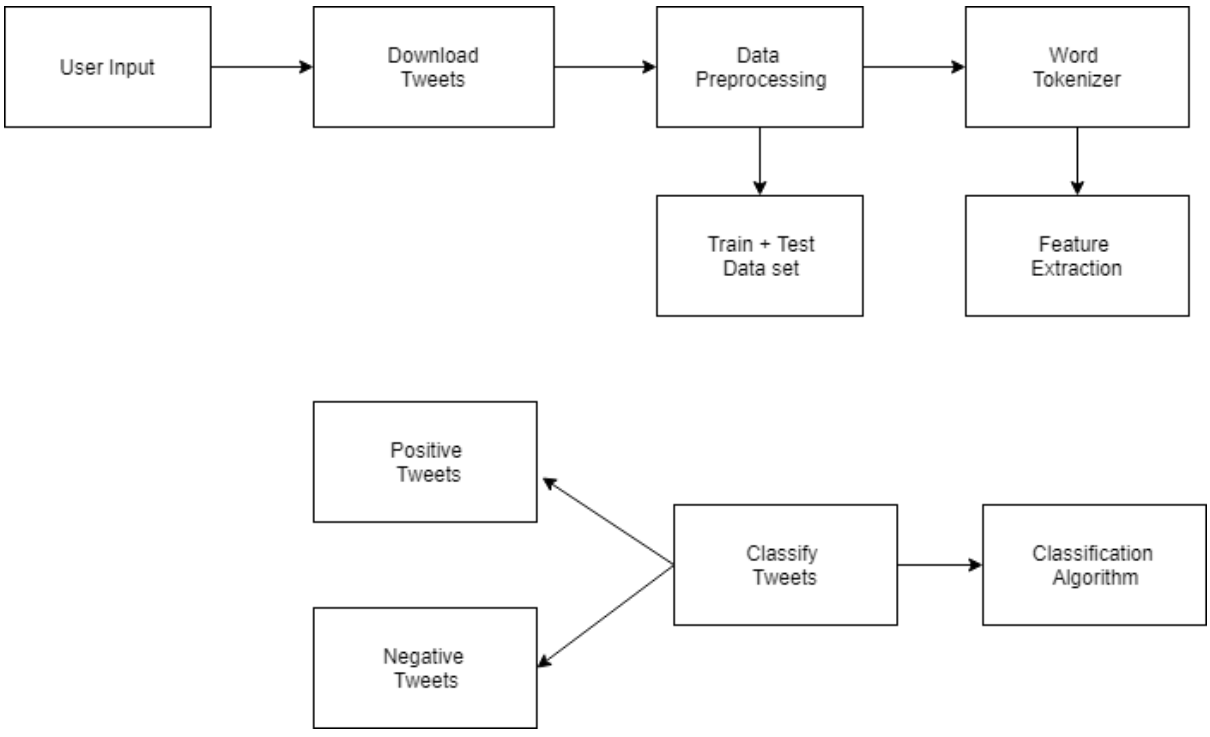

Figure 2 - System Architecture

Chapter 4: Implementation and Testing

This section describes the technologies used in Tweets Sentiment Analysis Using Logistic Regression Algorithm. Tweets Sentiment Analysis Using Logistic Regression Algorithm is a web application that uses flask python framework. HTML, Twitter Bootstrap CSS, JavaScript are used to develop front-end and python, NLTK, Scikit-learn are used to develop back-end. HTML is used for presentation technology. JavaScript are implemented to show the result of the application in a dynamic way.

All the algorithms for the application are written in Python. Algorithms used in Sentiment Analysis of Twitter Data Using Logistic Regression is Predictive analysis model. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables. The algorithm is implemented using python programming language.

4.1 Description of Major Function

The major function in the application are:

4.1.1 Preprocessing

This is the function which is run for processing the tweets.

Input: It takes the inputs as tweets.

Process: It call other function like remove URL, filter stop words, etc.

Output: It gives the list of the process tweets.

4.1.2 Feature extractor

This function is implemented after the preprocessing data.

Input: This takes pre-processed data as in input.

Process: It then uses the method, extract_feature() to process the taken input, process and extract the feature.

4.1.3 Classifier

This class implements the Logistic Regression algorithm which take feature from Feature extractor.

Input: It takes input from the feature extractor.

Process: It classify the tweets as positive or negative and return the list of tweets with its sentiment value.

Output: It gives classified tweets with positive and negative tweets value.

4.2 Testing

Among the total data 85% of the data is used for training and 15% is used for testing. For testing hit and trial method is followed and the testing module from Scikit-learn is used for testing.

4.3 Evaluation Metrics

Confusion Metrics: The confusion metrics provides a detailed breakdown of the model's predictions, showing the number of true positive, true negative, false positive, and false negatives.

Accuracy: Since we are working on a classification problem of classifying the tweets (positive and negative). Accuracy will be a good evaluator for the models. This is most commonly use metric to evaluate how well the model predicts. Accuracy is the ratio of the number of correct predictions made against the total number of prediction made.

$$\text{Accuracy} = (TP + TN)/(\text{Total Number of Samples})$$

Chapter 5: Conclusion

Tweets Sentiment Analysis Using Logistic Regression Algorithm was successfully implemented using python programming language. Twitter is undoubtedly a place where most people express their thoughts and feelings. It is very important to have an appropriate model in prediction of positive or negative tweets. 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-sentiment-analyzer/

[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*.

Appendix

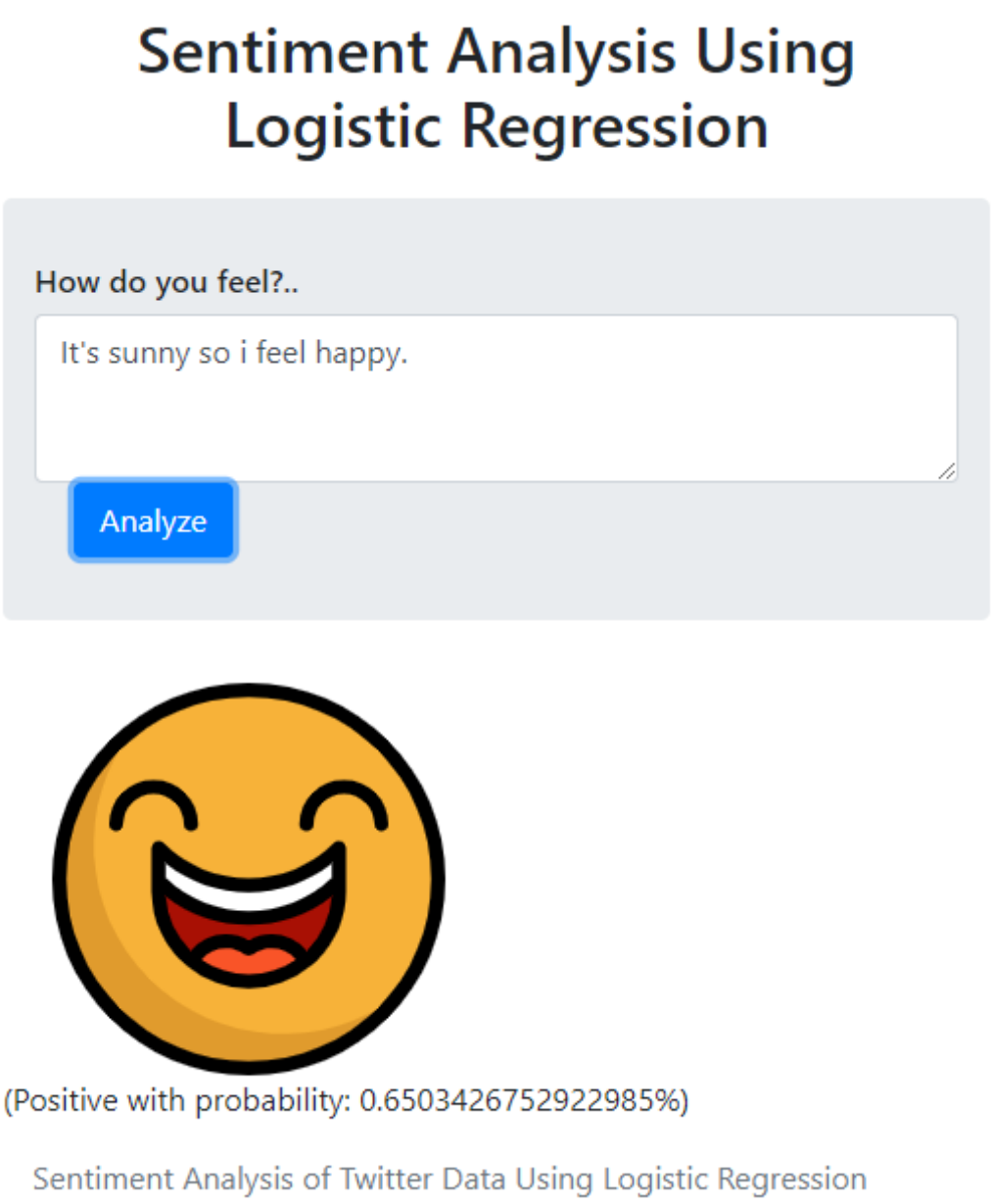

Figure 3 – Sentiment Analysis Result