NAÏVE BAYES CLASSIFIER FOR SENTIMENT ANALYSIS

***Phyu Thwe*[1], *Cho Cho Lwin*[2], *Yi Yi Aung*[3]**

[1,2,] *Faculty of Information Science, Myanmar Institute of Information Technology*
[3]*Faculty of Computer Science, University of Computer Studies, Mandalay*

Abstract

Sentiment analysis is an approach to explore whether collected content is in a positive, negative, or neutral state. When thinking varied prospects, for instance, celebrities, politicians, food, places, or another topic, sentiment analysis is employed to examine people's opinions, likes, and interests. Social media are additionally used to communicate our feelings on items and administrations. The remarks and evaluations of millions of clients of the social site can be gathered to extricate their perspectives and emotions towards any item or administration and utilize this data for future business and business enhancements or for domain investigation. This paper presents an open source approach where we gather tweets from the Twitter API, at that point preprocessing, investigate, and visualize those tweets. This sentiment analysis depends on retrieving text information from the communicated web, at that point characterizing individuals' perspectives into three remarkable notions, for example, positive, negative, and neutral.

Keyword: sentiment analysis, naïve Bayes classifier, natural language processing.

1. INTRODUCTION

Sentiment analysis or opinion mining is the study of people's opinions, feelings, emotions and attitudes towards entities such as products, services, organizations, individuals, subjects, events and attributes that Fast. Sentiment analysis has gained popularity among a wide range of people with different interests and motivations. Individual opinion is very important in arriving at a result or when making a decision, because people's opinion is made up of their past experience. Internet users express their opinions, views and ideas on many topics on various social networking sites and other sites, these sites can be used to extract data, which serve as a source of sentiment analysis [1].

Sentiment analysis is generally classified by aspect level, document level, and sentence level. The convergence of the sentence is seen at the sentence level. These phrases were combined to form a document. The document level is to determine the opinion of the document. At the aspect level, the system is classified and is required natural language processing according to the associated aspect of this entity.

In this paper, four sections will be studied: the related work in sentiment analysis is shown in section 2 using the tweeter dataset, section 3 explains the process of data cleaning, preprocessing, tokenization for sentiment classification and the proposed system is described in section 3 and section 4 finally presents the conclusion.

2. RELATED WORK

In this paper [2], the official Weibo API is used to track Weibo data, preprocess Weibo text, use the SVM algorithm to initially filter the text, use Naive Bayes to analyze Weibo sentiment, and split Weibo positively. The three types of negative and objective, while using the Adaboost algorithm to strengthen Bayes' naïve algorithm.

In this paper [3], the optimal parameter settings are first identified separately for each CNN and LSTM component. Then, all of the optimal parameter settings are identified for the system integration recognition framework around the optimum for each component. The experimental results establish that the precision of sentiment analysis with CNNs constructed on the LSTM model has better-quality by 3.13% and 1.71% respectively, compared to a single CNN or LSTM model.

In this paper [4], an attempt was made to propose a sentiment analysis method for the Twitter dataset. The polarity of each tweet is calculated to distinguish whether the tweet is positive or negative in the proposed method. A feeling polarity corresponds to the user's emotions, such as anger, sadness, happiness, and joy. The proposed system has been implemented using Python.

This paper [5] describes the automatic extraction of user preferences in smart tourism. User comments posted on social media are used as a rich source of data that implicitly includes their preferences. The proposed method extracts the user's preferences through the semantic grouping of comments and sentiment analysis. The results of the evaluation indicate high values of precision, recovery and parameters of measurement in the preferences of tourists to take away.

3. PROPOSED SYSTEM

Sentiment analysis is performed to rank the sentence in tweets using machine learning algorithms like naive bayes. The purpose of the system is to provide positive, negative and neutral for the data. In the proposed system, there are two phases which are the training phase and the testing phase. In the training phase, the system performs preprocessing steps such as data cleaning, tokenization, detection of stop word. And then the features extraction is performed and obtained, the system trains these feature vectors with the Naive Bayes classifier. Naive Bayes runs a classification model to classify the positive, negative, and neutral labels of data. In the testing process, the system tracks actual Twitter data. Finally, the system ranks the positive, negative, and neutral percentages for each field in this actual test data. Figure 1 shows the general flow of the proposed system.

3.1. Preprocessing

Data cleaning is an important procedure for the data preprocessing tool. It is a method of data mining techniques. Remove bad error data and reduce unnecessary data information and also lack of data is included in data cleansing techniques. The presence of noise data can affect the intrinsic characteristic of a classification problem.

The first step in text analysis is tokenization. Tokenization is the process of separating a part of text into smaller chunks, such as words or sentences. The token is a single entity that makes up the building blocks of a sentence or paragraph.

Tokenization is the first step in text analysis. The process of dividing a paragraph of text into smaller pieces, such as words or sentences, is called tokenization. The token is a single entity that makes up the building blocks of a sentence or paragraph.

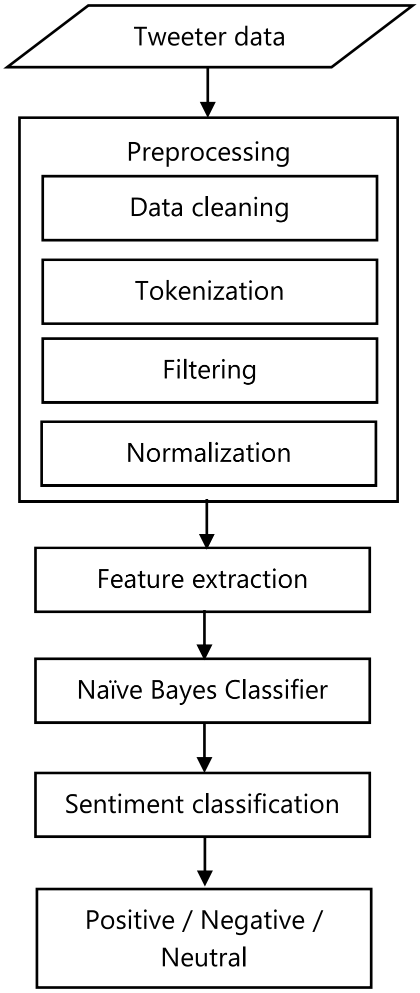

Figure 1 Proposed system flow

Stop word removal can be thought of as an entity selection routine, in which entities that do not contribute to correct ranking decisions are considered spurious words and are removed from the entity space accordingly. A supervised method called the mutual information (MI) method is that works by calculating mutual information in the interval of a given term and a class of documents (e.g., positive, negative). An indication of the term amount of information can inform about a certain class. It was suggests that the term has low discriminatory power and it deleted by mutual information.

Normalization is known as the process of eliminating irrelevant data from a huge collection of draw out data in sentiment analysis. Tags, URLs, and links that contain a lot of noise in the extracted data. In data preprocessing, the system remove the noise from the extracted text and make it clearer and more consistent. Before moving on to the analysis phase, the data must be pre-processed in each text mining process. Therefore, the extracted data was preprocessed for further processing. The URLs that are not generally useful for the sentiment analysis process are removed from the data.

Stemming is the process of reducing inflection to its root forms. Even though the root does not have a proper meaning, it occurs in a group of related words under the same root.

Lemmatization is the process of converting a word to its base form. It considers the contexts and converts the word to its meaning form by using python NLTK Lemmatizer that uses the WordNet database to find lemmas for inflected terms. The canonical form of the word is known as motto.

3.2. Feature Extraction

The transformation of the input data into the feature set is called feature extraction. If the extracted features are chosen correctly, the feature set is expected to achieve the desired mission by using a reduced version instead of a full size input. Feature extraction includes dropping the amount of assets required to describe a large data set.

The simple TF-IDF model works well and emphasizes rare words rather than treating all words the same in the case of the binary bag of words model. However, this pattern does not work correctly when it encounters a sentence containing negations. This negation is a very common linguistic construct that affects the word / sentence polarity. Therefore, the model should be framed in such a way that if the presence of negations is considered, a better result can be obtained.

There are three types of feelings, namely positive, negative and neutral feelings. Positive feelings which refer to the speaker's positive attitude towards the text. Emotions with positive feelings reflect well, increase, grow, etc. Negative feelings refer to the speaker's negative attitude towards the text. Emotions with negative feelings reflect evil, fall, underdevelopment, etc. If the negative feelings are more, it means that these areas are underdeveloped. Neutral feelings that the emotions are not reflected in the text. It is neither preferred nor overlooked. Although this class does not imply anything, it is very important for a better distinction between positive and negative classes.

There are two techniques such as supervised and unsupervised techniques for analyzing feelings. The classification is carried out by means of a function which associates the characteristics of a given text with lexicons of discriminating words whose polarization is determined in the unsupervised technique. The main task is to build a classifier in supervised technique. The training samples can be tagged manually or obtained from a user-generated tagged online source. That is needed by the classifier. Support Vector Machine (SVM), Naïve Bayes Classifier and Decision Tree Classifier are the most widely used supervised algorithm. The system performs sentiment analysis using the supervised technique, the Naive Bayes classifier. The Naive Bayes classifier trains the features to build the classifier model. The classification model realizes that the output features of the feature extraction step classify positive features, negative features and neutral features with their weights. In this way, the test phase works well easily and accurately predicts the class.

4. CONCULATION

Twitter is a great starting point for social media analysis. People share their opinions on Twitter with the general public. Sentiment analysis is one of the most common analyzes that can be performed on a large number of tweets. In the proposed system, tweets are tracked using Twitter's Twitter streaming API. The collected tweets are pre-processed using the techniques of the Natural Language Toolkit. The characteristics of the tweets are selected based on TF-IDF and the Naive Bayes classifier is used to classify the tweets as positive, negative, or neutral. The proposed system is implemented using Python.

REFERENCES

[1] Baidya Nath Saha, Apurbalal Senapati, “Long Short Term Memory (LSTM) based Deep Learning for

Sentiment Analysis of English and Spanish Data", 2020 International Conference on Computational Performance Evaluation (ComPE), 2020.

[2] Xiafei Feng, "Research of Sentiment Analysis based on Adaboost Algorithm", 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019.

[3] Jinfeng Gao, Ruxian Yao, Han Lai, Ting-Cheng Chang, "Sentiment Analysis with CNNs Built on LSTM on Tourists Comments", IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (IEEE ECBIOS 2019).

[4] Sanjeev Dhawan, Kulvinder Singh, Priyanka Chauhan, "Sentiment Analysis of Twitter Data in Online Social Network", 5th IEEE International Conference on Signal Processing, Computing and Control (ISPCC 2k19), Oct 10-12, 2019, JUIT, Solan, India.

[5] Zahra Abbasi-Moud, Hamed Vahdat-Nejad, Wathiq Mansoor, "Detecting Tourist's Preferences by Sentiment Analysis in Smart Cities", 2019 IEEE Global Conference on Internet of Things (GCIoT), 2019.