ISSN 2395-1621

Natural Language Processing on Hospitals: Sentimental Analysis and Feature Extraction

[#1]Atul Kamat, [#2]Snehal Chavan, [#3]Neil Bamb, [#4]Hiral Athwani,
[#5]Prof. Shital A. Hande

[2]chavansnehal247@gmail.com

[#12345]Department of Computer Engineering,
Sinhgad Academy of Engineering, Kondhwa, Pune.

ABSTRACT

Reviews have been very valuable for many organizations to know about their performance. Efforts have been made to automate the task of analyzing the sentiments using different techniques. Machine learning is a type of artificial intelligence that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. [1]Natural Language Processing uses Sentimental Analysis to detect whether the particular text is positive, negative or neutral. Sentimental Analysis uses logistic regression for creating the model. Understanding the genuine opinions of user-generated content automatically is of great help for commercial and marketing purpose. The task can be conducted on different levels, classifying the polarity of words, sentences or entire documents. Generally, the frameworks consists of extracting massive amount of genuine text reviews for training our model. [2] This captures the general sentiment clustering of sentences .It is usually achieved through web scraping. Different machine learning techniques are used to generate a model. Recently, logistic regression has been proposed as an effective means for solving real time classification problems. The purpose of this paper is to provide detailed analysis of reviews on collective basis.

ARTICLE INFO

Article History

Received: 11th June 2018

Received in revised form :

11th June 2018

Accepted: 13th June 2018

Published online :

13th June 2018

Keywords- Machine Learning, Sentimental Analysis, Django

I. INTRODUCTION

Natural Language Processing extracts information, predict patterns and determines the tone of the text. It has many applications in determining the reviews, getting the responses of online and social media. This kind of analysis has a huge importance in monitoring platforms as it allows the organizations to get the genuine opinion of the public. The applications of sentiment analysis are vast and dominant in many fields of work. By using Sentimental analysis, it quickly understands the positive or negative behavior of the reviews and provides the necessary information to the users. There are various approaches to sentiment analysis one of which is machine learning. Machine learning is a statistical approach that includes many algorithms which may be supervised or unsupervised. Logistic Regression has been used in sentiment analysis of texts. Lexical approach is also one of many approaches which has been made. This includes maintaining of a dictionary of lexicons which are pre-tagged. Apart from using Machine Learning in Sentimental Analysis, front-end applications for representation of overall analysis are used. One such platform is Django. Django has been used by many others for creating applications for web. It is a software, which uses a database. It includes some kind of user interactivity, and operates through a web browser. A Framework provides a structure and common methods for making such a software.

CREATING AND TRAINING A MACHINE LEARNING MODEL

Logistic Regression finds out the tables of probabilities that are used to estimate the likelihood that the new data belongs to various classes. The probabilities are calculated using Bayes theorem, which specifies how these events are related.

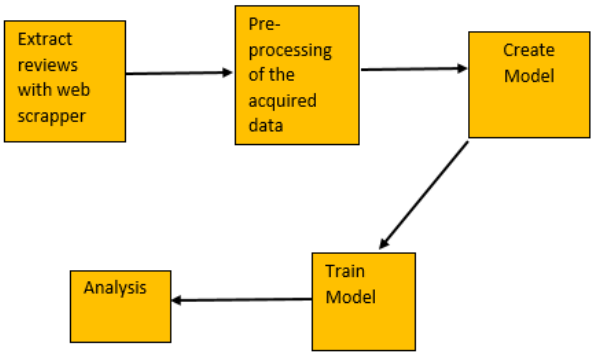

Figure 1: Flowchart of the Proposed Work

A. Bag of Words

Bag of words consist of unordered collection of words. Based on this, the Text classification can be done. The occurrence of each word is expressed as a feature for training a classifier .This occurrence is often referred to as the frequency of words. It also consists of stop words. Stop words are the list of words which are not required for training the classifier. These words are often used in the sentences and have no importance to determine the sentiment.

B. Obtaining Large Datasets for Training

Web Scraping can provide us with a data set which can be used as the training task. Web scraping is a process of getting information from multiple web pages automatically.

A web scraper is basically a program that uses a Hyper-Text Transfer Protocol to send request and copying data from Web servers to local databases. Since a web scraper fetches and downloads each page its main task is known as web crawling. Web scraping is used in sentiment analysis to collect training sets and a web scraping program can be modified or built to do so. A web scrapper named Beautiful Soup is used. It is easy and intuitive to use. It scrapes data from the websites and converts it into a structured html format.. This training data collected is then arranged in separate databases according to their class. Work has been done on python with the famous Beautiful Soup library to scrape data which uses request function to access websites through HTTP. The HTTP page is then stored in a source code form. Specific tags from the source code can then be chosen to be stored on the database.

C. Creation of Machine Learning Model

After extracting datasets from a web scrapper it is now necessary to create a model. This model is considered as the machine learning model.

1. Formatting of the training data

The words of the datasets are broken down into tokens. The tokens are the simplest form of sentences which consists of one word. Let's consider a sentence "The ambience is excellent". This sentence gets broken down into tokens. The token form of the sentence would be "The", "ambience", "is" and "excellent". Words are them Lemmatized. This involves breaking down the tokens into their base form. For example, "speedy" becomes "speed" and "better" becomes "good". After Lemmatizing of tokens, the Stop words (unwanted words) are removed. Stop words are the most frequently used words and show no significance in determining the sentiment. This leads to simplified set of tokens.

B. Word Index Map

The token generated are added into a dictionary which consists of unique words only. For each of these words, the frequency is computed. Thus, we form a dictionary which consists of all the token words and their frequency.

2. Tokens to vector

Now comes the part where we create the model. Tokens are converted to vectors and labels are defined. The positive tokenized words have label 1. The negative tokenized words have label 0. The word index map is used to create a "data" variable. This data variable is a 2D array. The rows consists of reviews. The columns consists of words that are present in the word index map. For example: Consider a review that consists of words "good" &"bad". The system counts the frequency of each words and divides it with the total number of words in the particular review. Figure 2 and 3shows the internal representation of the data variable.

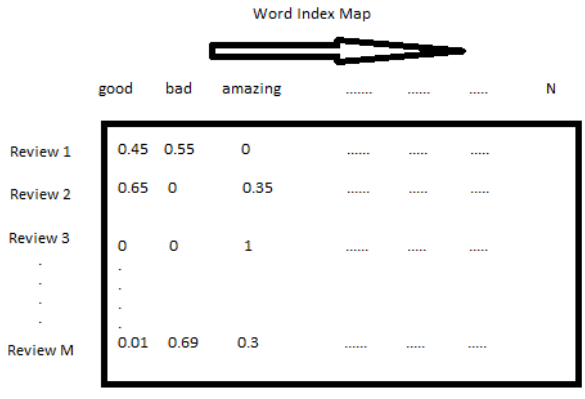

Figure 2: Internal Structure of Data Variable

```
0.00479616 |  0.0119904 |  0.00483092 |  0.0120773 | 1
0.51213136 |  0.0113404 |  0.02383092 |  -0.012073 | 0
0.10479096 |  -0.011704 |  -0.0453692 |  0.0350773 | 0
```

Figure 3: Model required for training

3. Logistic Regression

Logistic Regression is a statistical method, popularly used for analyzing the dataset and determining the accuracy of the model. Logistic Regression defines the relationship between dependent and one or more independent variables. They are classified according to their sentiment (positive and negative). Considering the frequency i.e. the number of times they have occurred in a training set named as positive or negative class, each word is given some weight according to the model generated. This weight is considered as the polarity. The polarity consists of positive and negative values having a range from +∞ to - ∞. Logistic Regression consists of binary values: 0 and 1.Any token or word is more inclined towards the positive class if its value is greater than the some other tokens value. The Logistic regression equation:

y=y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

where y is the output that is predicted,b0 is the expected value and b1 is used to represent the single input value (x). This equation runs internally in the logisticregression() function in python. We just have to provide the dependent and independent values.

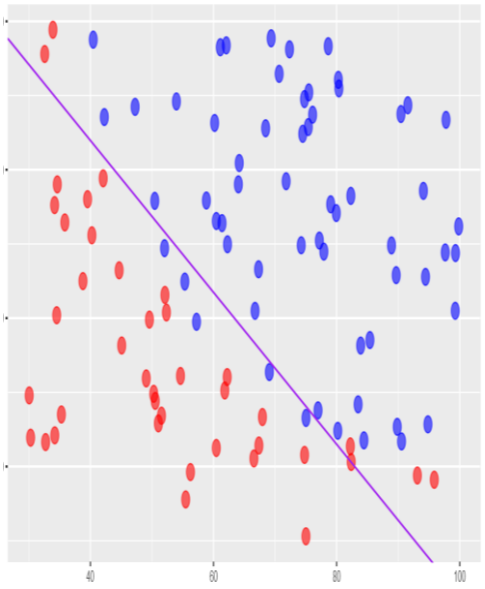

Figure 3: Logistic Regression Structure

4. Training and Testing Data

The first 100 rows from the created model are considered for training. The last 100 rows are considered for testing the model. The classification rate is computed.

5. The Dataset

This weighed model is considered as the training data. The training data in logistic regression must be balanced. If the dataset is not balanced, then active learning approach is used for the text classification, which helps the system to overcome the problem of training data size. When the dataset is small, there is a possibility that the training data and testing data are not compatible and may give wrong results. Cross validation gives an average result but it doesn't help in training data selection. There is a definite statistical approach for building representative training dataset (training data selection). Once the training data is created, a real time text is given. This real time text is the review from the user. The polarity is then computed. It can update the service accordingly. There is no need of a management to look into each review to know about their performance. A system using machine learning can automatically find the review highlights from a set of reviews and update the information to the service after learning from this dataset.

6. No star ratings

Sentiment analysis not only focuses on the user (customer), but also the service provider. The data to be reviewed or analyzed could be from any platform which may not be having a star rating associated with it. This makes the sentiment analyzer very useful to reduce a lot of human efforts. Feature extraction along with sentiment analysis can prove more useful than star based reviews as what feature or attribute in the natural language causing the opinion can be extracted at the same time.

5. JavaScript Representation for Faster Analysis

Highcharts is a JavaScript library that is used to create an interactive environment. Highcharts consists of many kinds of charts. Pie charts have been used for better representation of sentiments.

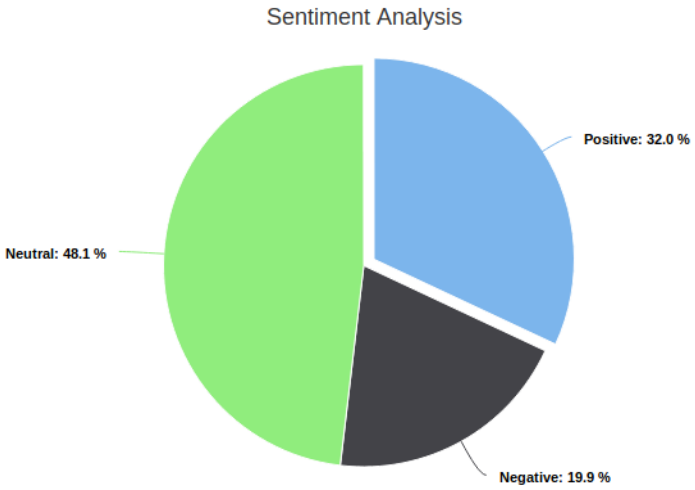

Figure 3: Highcharts for sentiment analysis representation

II. FEATURE EXTRACTION

After finding the polarity of reviews, it is important to perform certain analysis. Natural Language Processing makes sure that this is done automatically within

milliseconds. It takes millions of reviews as input. It then uses certain machine learning algorithms to analyze the entire data. This analysis is benefitial for the marketing team.

A. Why Feature Extraction?

Suppose a hospital website has 5000 reviews. By using sentiment analysis algorithms, the overall polarity can be estimated. But this might not be enough for the analysis of their performance. For example If a particular hospital finds out that it has 60% positive reviews and 40% negative, it is benefitial for the hospital to find out what exactly is good in the 60% and what exactly is bad in the 40%. This can be done by Feature Extraction.

B. N-Grams for Feature Extraction

N-Grams are widely used in text mining and natural language processing. [3] It consists of Bigrams and Trigrams.
Bigrams deals with extracting a sequence of two most frequently used words. A Trigram, on the other hand, deals with extracting a sequence of three most frequently used words. N-grams can also be used for sequences of words or almost any type of data. These N-Grams have proved to be very effective and reliable in the field of natural language processing. This can be used as review highlights.

III. DJANGO

Django is a web framework which is written completely in Python. It observes the model view architecture pattern. Django is very secure. The security in accessing data is much better than PHP. It hides the site's official source code from direct viewing by users. The Django components have lots of dependencies amongst them. Popular websites like Instagram, Pinterest and NASA use Django.

Django consists of the following modules:

I. The Model Layer: It consists of the entire database that is used. Our model layer consists of tables for creating a user, creating a hospital, searching the hospital via name and searching the hospital via speciality. Admin handles the entire database.

II. The view layer: The view layer consists of all the functions that are executed after opening a particular URL. Urls.py consists of the path and the associated function name in views.py.

III. The template layer: It is the presentation layer. It consists of all the html pages that are accessed by the users as well as the hospitals.

Our project uses the Django Framework to provide a convenient way of accessing and updating the medical history. Our project uses the Django Framework to provide a convenient way of accessing and updating the medical history. It consists of a user friendly search engine optimization. Hospitals can be accessed via name as well as their speciality. Hospitals with better reviews are tend to be shown first. It consists of user profiles and hospital profiles.

IV. FUTURE SCOPE

Research on classifying sarcastic words which cannot be classified based on their weights is an ongoing topic of interest for many scientists. [4] Recognizing sarcasm is very much important for understanding people's genuine sentiment and beliefs. For example, if we consider a statement "The journey was so good that everyone felt sick", will lead to a conclusion that the sentiment is positive towards the event of "felt sick". [5] Word Sense Disambiguation deals with finding out the correct meaning of the word. This is associated with the meaning of the word in the text. Here the task of checking if the sense of a word is literal or sarcastic can be termed as Literal or Sarcastic Sense disambiguation (LSSD). To tackle sarcasm authors have proposed a crowd-sourced task that relates to the task of creating a parallel database that groups words that can have a sarcastic meaning. Considering the above text, good in this context will be considered as bad or terrible and thus will be listed in the database.
Word embedding in a modified SVM achieve the best results among others.

The authors also put forward unsupervised techniques to detect semantically opposite words or phrases. One of which is a co-training algorithm proposed by Barzilay and McKeown (2001). The co-training algorithm is used to remove paraphrases from a sentence. If we have a data base having sentences with tags named IM and SM where IM is the Intended meaning and SM Sarcastic meaning we can extract the opinion describing words as para-phrases. Suppose,

SM1: The cycle was so good that it broke.
IM1: The cycle was so bad that it broke.
Here, the anchor words can be discarded and para-phrases can be found.

V. CONCLUSION

The proposed work shows a way of analyzing millions of reviews. The analysis includes finding the overall sentiment, extracting rich features, providing an interactive and secure environment for the users as well as the hospital team. Natural Language processing eases the entire approach of marketing. The model shows many capabilities of analyzing the data. Different algorithms can be applied in order to make the approach more and more efficient.

REFERENCES

[1] Oskar Ahlgren, "Research on Sentiment Analysis: The first Decade" in *Machine Learning*: IEEE..

[2] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques". In Data Mining, 2003. ICDM 2003. Third IEEE International Conference, pp. 427-434, 2003.H. Poor, *An Introduction to*

Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4

[3] M. Trupthi,Suresh Pabboju,"Improved Feature Extraction and Classification - Sentiment Analysis." International Conference on Advances in Human Machine Interaction (HMI - 2016)

[4] Shalini Raghav, Ela Kumar, "Review of automatic Sarcasm Detection".2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)

[5] Himdweep Walia1 , Ajay Rana , Vineet Kansal," A Naïve Bayes Approach for working on Word Sense Disambiguation ".IEEE Conference,2017.

