International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
www.ijemr.net https://doi.org/10.31033/ijemr.10.2.11

OSEMN Approach for Real Time Data Analysis

Kajal Kumari[1], Mahima Bhardwaj[2] and Swati Sharma[3]
[1]B.Tech Student, Department of Computer Science, MIET, Meerut, INDIA
[2]B.Tech Student, Department of Computer Science, MIET, Meerut, INDIA
[3]Assistant Professor, Department of Information Technology, MIET, Meerut, INDIA

[1]Corresponding Author: kajal.kumari.cs.2016@miet.ac.in

ABSTRACT

Data analysis system is the study of people opinions, sentiments, attitudes, and emotions expressed through written language. Sentiment analysis is just a part of it in this system we compare the accuracy result of two languages sentiment analysis. If we saw than sentiment analysis is one of the most active research areas. It is popular because of two reasons. First, it has a big range of applications because opinions are center of almost all human activities and it shows our behaviours. Whenever we make a decision, we have to heard other's opinions as well. Second, it presents many challenges research problems, which had never been strive before the year 2000. Part of the reason for the lack of study before was that there was small dogmatic text in digital forms. There is no surprise that the establishment and the rapid growth in the field coincide with the social media on the Web. In fact, the research has also increase outside of computer science to manage science and social science due to its importance to business and society as a whole.

Information analysis system is the system in which we measure the accuracy rate of both languages chirps. The main thing is this that this project is a newly formed project. We can say that sentiment analysis is just a part of it for that we have to understand what is sentiment classification and analysis. So Sentiment classification is a way to inspect the personal data in the chirps or data and then extract the opinion. Chirps analysis the method by which information is withdraw from the opinions, and emotions of people in regards to things. During decision taking the opinion of other person shave a drastic effect on users or customers ease because they make choices regarding to e-shopping, choosing events, products, things. The approaches towards chirps analysis work according to a particular level, document level. This paper aims at analysing a solution for the sentiment classification at a powdery, mainly in the sentences in which the polar nature of the chirps or sentences given by three categorization name as positive ,negative and neutral.

Keywords— OSEMN, Real Time, Data Analysis, Data Science

I. INTRODUCTION

Bilingual real time information analysis system is the study of people opinion, sentiments, attitudes, and emotions expressed in written language. Bilingual real time information analysis system is the system in which we measure the accuracy rate of both languages chirps . The main thing is this that this project is a newly formed project. We can say that sentiment analysis is just a part of it for that we have to understand what is sentiment classification and examination. Sentiment analysis refers to the use of NLP, text investigation and estimation semantic to recognize and draw out personal data in source. Generally speaking, sentiment analysis aims to ascertain the attitude of a speaker or a writer with respect to some topic on the overall situation polarity of a document. The attitude may be his or her judgment or evaluation assessment state, or the deliberate emotional communication. Sentiment analysis is the process of detecting a piece of writing for positive, negative, or combo feelings bound to it. Humans have the inherit ability to determine sentiment; however, the process is time consuming, not staying the same throughout, and costly in a business conditions.

For example if we consider deman cloud based sentimental analysis software. deman cloud-based sentimental analysis software extract the sentiments of a document and its components through the following steps:

- A document is broken in its basic part of speech which identify the structural part of a document, paragraph, or sentence.
- Sentiment-bear phrases are identified through the use of specifically designed algo.
- Each sentiment beared phrase in a document is given a score based on a logarithmic scale that having a range between 10 and 10.
- Finally, the scores are combined to determine the overall sentiments of the document or sentence Document scores having a range between 2 and 2.

Deman cloud-based sentimental analysis software is based on the NLP technology and delivers you more consistent result than humans. Using automatic sentimental analysis, Deman analyses each document and its components based on algorithms developed to extract sentiments from the content in a similar manner as a human can do only 60,000 times much faster.

II. METHODOLOGY

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.11

There are mainly two types of approaches for Bilingual real time information analysis system classification

Machine Learning

The machine learning based words classifiers are a kind of supervised machine learning model, where the classifier needs to be trained on some fully labeled trained data before it will be applied to actual classified task. The training data is usually an substitute portion of the original data hand labeled by itself. After suitable training they can be used on the actual test data. The Naive Bayes is a classifier whereas SVM(Support Vector Machine) is a kind of vector space classifier.

(SVM) is a kind of classifier which requires that the text documents should be transformed to feature vectors before they are used for classification. Usually the text documents are transformed to multidimensional vectors or have multiple dimension. The entire problem of classification is then classifying every text document. It is a type of large margin classifier.

Here the goal is to find a decision boundary between two classes

This approach needs:

- A good classifier model such as Naive Byes
- A training set for each class

There are various training sets available on, twitter dataset, etc. Class can be Positive, negative. For both the classes we need training data sets.

Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and semantic concerned with the interactions between computers and human (natural) languages. This approach utilizes the library publicly available, which provides a sentiment polarity values for every term occurring in the document. In this lexical resource each term n occurring in document is associated to three numerical scores object(n), positive(n) and negative(n), describing the objective, positive and negative polarities of the term, respectively. These three scores are calculated by combining the results produced by eight ternary classifiers.

It groups words together based on their meanings.

Major tasks in NLP that helps in withdrawing sentiments from a text:

- Extracting part of the sentence that reflects the sentiment
- Understanding the structure of the sentence
- Different tools which help process the textual data

III. PRIOR APPROACH

Data Science Process

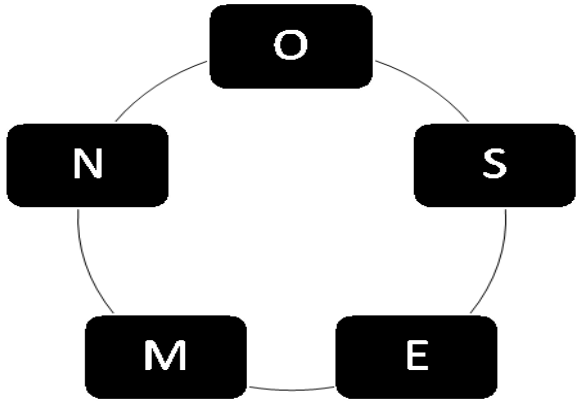

Figure 1: Data science process

[1] O(Obtain) – gather data from linked sources.
[2] S(scrub) – clean data into that formats that machine understands.
[3] E(explore) – EDA explore the data called exploratory data analysis.
[4] M(model) – construct the models to predict and forecast.
[5] N(interpret) – show the results into good formats.

IV. OUR APPROACH

We are working on OSEMN framework of data science in our project named as REAL TIME DATA ANALYSIS SYSTEM.

Gathering Data

The very first step of our is to collect and obtain the data. Gathering data means collect data from different sources, this first step is crucial because if we don't have data than how we move on in our project. For extracting data we use query to get data from database.

We take data in a format named as microsoft excel to extract or gather the data and then we used to convert it into a usable format or data. so we are moving from unsupervised to supervised data.

The different way we are using for collecting the data is from twitter using web APIs. The twitter which is a very famous social media allowed end-users to extract their data from their web servers.

That we are doing in our project. We are taking the chirps from the tweeter with the help of Web API. To obtain the chirps from twitter through Web API we are generating keys that are named as CONSUMER KEY, CONSUMER SECRET KEY, TOKEN KEY, TOKEN

SECRET KEY. And the code for obtaining the chirps from the tweeter is given below:

```
consumer_key = "Sh4rtCV78EummAG1yWaQaXmyw"
consumer_secret=
"h6BTyRKOKPol7ucXiZRNmiJuFTM4wvsu8adIU5rt5qV
y6N6TfX"
access_token="1046769984737665029-
y7HJ9s9wBEyDkJKHc5ik3CCyX083Q3"
access_token_secret=
"j51htadkbtPBkuiL46ZlWp1fB7tlXREia32UarRH5WKAv"
```

Data Scrubbing

In data scrubbing procedure we are cleaning the data and filtering the data. If we use unfiltered data then it affect our final result accuracy and inappropriate.

Scrubbing of data also includes the data extraction and exchanging the values. In our project for scrubbing the data we apply LEMMATIZATION, STEMINIZATION.

Data Exploration

Data exploration is a process to explore the data we don't move direct to MI and AI, we have to search the data.

Firstly, we have to explore the data and its all characteristics. There are many categories of data like categorical data, interval, ratio data and so on. But our data is categorical type. Chirps can be belong to any category and we have to just classify to which category they belong. In this step we explore our data through EDA process. EDA is a exploratory data analysis, there are many things we do in EDA process like we extract the chirps from twitter and analysis it that they are scrubbed or not and our data is ready for modelling process or not.

The next step will be to calculate illustrative statistics to extract attributes and test notable variables. Testing notable variables often times is done with connection association. For example, exploring the connection association. The word named as "Feature" used in Machine Learning or Modelling, is the data features to help you identify what are the feature that show this database.

```
d=pd.read_csv("sample tweets.csv")
```

Data Modelling

This is the most fascinating stage of the data science project life cycle. As many people would call it "Where the magic transpires".

Once again afore reaching to this stage bear in your mind that the scrubbing and exploring stage is critical in order for this process to make sense. So take your time on those stages in lieu of forward to this process.

One of the first thing you require to do in modelling data is to abbreviate the dimensions of your data. Not all your characteristics or values are essential to sooth saying your model. Hence what you require to do is to cull the pertinent once that will grant to the prognostication of results you are probing for.

We apply some algorithms in this step and on the basis of those algorithms we classify our data. The algorithm used in our project is naive byes. On the basis of category we choose two categories from two languages that one is English and French and extract the chirps of those categories in both the languages and check the accuracy rate of both categories and compare the accuracy rate of two languages.

Other than relegation and the prognostication of the result our purpose of this stage can include the grouped the data to understand the logic behind those clusters. For example you would relish to group your e- commerce website customers to understand their comportment on your website. So this would require you identify groups of data points with clustering algorithm utilizing methods like k-designates or make presage utilizing regression like linear or logistic regression. Lastly in this step we train models for relegation.

Data Interpreting

We are at the final and most paramount step that is interpretation data. This is rudimentally refers to the presentation of the data, delivering the result in such a way that is able to answer the business questions you asked when you started the project, together with the actionable perception that is found through data science.

In this step we are going to give or represent those languages chirps whose accuracy rate is higher than the other language chirp. That means we are going to give outcome of that language which gives us an effective result. And we will show this output through PIE CHART.

It is very consequential to present findings in such a way that is subsidiary to our organization or else it would be unavailing to our stakeholder.

Instead of technical skills we required one other skill as well that is to able to tell a clear and applicable story.

V. CONCLUSION

In today's scenario machine learning is growing topic and in that sentiment analysis is a hot topic. In this project we are trying to compare chirps of two different languages (English and french). Firstly we will classify chirps into three categories - positive, negative. For classification we are using classification algorithm (Naive byes algorithm) in our project. Thus through this process we will find the accuracy of both languages by comparing there results, in this procedure we are using real time data. Every chirp is extracted from the real time data.

We are using election data in our project. We are using thousands of chirps to improve the result accuracy.

REFERENCES

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.11

[1] B. Jansen, M. Zhang, K. Sobel, & A. Chowdury. (2009). *Micro-blogging as a online word of mouth branding.* Available at: https://faculty.ist.psu.edu/jjansen/academic/pubs/jansen_Micro_blogging_as_Online_Word_of_Mouth_Branding.pdf.

[2] C. Manning & H. Schuetze. (1999). *Foundations of analytical Natural Language Processing(NLP)*. Available at: https://www.cs.vassar.edu/~cs366/docs/Manning_Schuetze_StatisticalNLP.pdf.

[3] B. Pang, L. Lee, & S. Vaithyanathan. (2002). *Thumbs up sentiment classification using machine learning proficiency*. Available at: https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf.

[4] B. Pang & L. Lee. (2008). Opinion Mining and sentiment analysis. *Foundation and Trends in Information Revival*, *2*(1-2), 1-135.

[5] B. Pang & L. Lee. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* Available at: https://www.cs.cornell.edu/home/llee/papers/cutsent.pdf.

[6] J. Read. (2005). *Using emotions to reduce dependency in machine learning techniques for sentiment classification*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.3058&rep=rep1&type=pdf.

[7] P. Turney. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.* Available at: https://www.aclweb.org/anthology/P02-1053.pdf.