EasyChair Preprint

№ 5978

Sensational News And News Maker Using Data Mining And Opinion Extraction

Aswathy Dhanapalan

July 1, 2021

Sensational News And News Maker Using Data Mining And Opinion Extraction

Aswathy Dhanapalan
School of SCIS
UOH
Hyderabad, India
20MCMT21
aswathydhanapalan2017@gmail.com

***Abstract*—The current outbreak and widespread of microblogging and social networking websites are reshaping many aspects of the modern day social interaction. Social Media today has exceeded the limits of entertainment or simple social interaction contexts. Social media now is better described as a living organism, that has a structure and a soul. The Social Media now reacts the pulse of the people; it reacts to their emotions, and interacts with their opinions. Analyzing and monitoring the content of Social Media can bring about some valuable insights, of which the conventional media means weren't able to convey. Social Networking is the medium which allows million users to share their opinion, ideas, views and huge amount of data is presented for internet users and a lot of data generated too. This project proposes the paradigm to extract the sentiment from famous micro blogging service, called twitter. Twitter is one of the most popular portals, where people post their opinions, views for everything. In this project, data mining techniques are used to automatically classify the sentiments of Tweets taken from Twitter dataset. This project tackles the concept of sensational news and newsmaker detection within Twitter. The project identifies the dimensions of similarity and interaction between news and tweets that provides the tools to calculate sensational news maker and news**

***Index Terms*—Data Mining, Sentimental Analysis, Microblogging**

I. Introduction

This project gives more beneficial solution by providing in depth detailed information of data. In this context this implementation serves processed information of tweets accessed from Twitter Server. This project proposes a new system that delivers large database of Social Networking Site (SNS) called Twitter. Many Third party application are building based on SNS like Twitter, they need to have processed data from their operational purpose. The main stream of the applications is visualization applications. The purpose of this document is to present a detailed description of data mining and opinion extraction used in finding the sensational news and newsmaker for a particular period along with the sentiments of people towards that particular news. It explains the purpose and feature of the system, the interfaces of the system what the system will do, the constraints under which it must operate on. This is mainly intended for uses in news medias.

The purpose of this project is to develop an automated way of gauging the sensational news for a particular time period, news icon for that particular time and the mood of people on social media towards that particular news, specifically Twitter. The software should extract the data from a large database and determine the sentiment of the Twitter community with respect to a given topic. The main scope of data mining are:

- Automated detection of news maker.
- Detection of controversial or tweets with maximum retweets and shares.
- Detecting the statements and policies taken by these newsmakers that has become the topic of discussion.
- Orientation of the people towards these statements.

II. SYSTEM ANALYSIS

A. Existing System

The existing system used only structured data analysis. It used phrases and grammar information only to a limited extent. The user has to search a lot in order to find the relevant information from various on-line sites. Analysis were done on small amount of data in these techniques. It does not work in real time environment since it uses hadoop platform for analysis. The existing methods of opinion mining are:

- Product review mining.
- Sentiment analysis of news articles.
- Tracking sentiments toward topics over time.
- Orientation of the people towards these statements.

B. Proposed System

Proposed system finds the sensational news and news-maker for a particular period along with the sentiments of people towards that particular news.It explains the purpose and feature of the system ,the interfaces of the system what the system will do, the constraints under which it must operate on. This is mainly intended for uses in news medias.

III. SYSTEM DESIGN

Design of the system includes mainly two steps:

- System design
- Detailed design

In System design a structural framework for the entire system is created. It is done in such a way that related part

come under particular groups. Thus after the system design, a network of different groups is obtained. It is the high-level strategy for solving the problem and building a solution. It includes the decision about the organization of the system into subsystems, the allocation of subsystems to hardware and software components, and major conceptual and policy decisions that form the framework for the detailed design.

In detailed design, each group is studied in detail and the internal operations are decided. Based on this, the data structures and the programming language to be used are decided. Apart from detailed design, the system design can be grouped into physical design and structural design. The physical design maps out the details of the physical system and plans the system implementation and specifies the hardware and software requirements.

Structured design is an attempt to minimize the complexity and make a problem manageable by subdividing into smaller segments, which is called modularization or decomposition. In this way structuring minimizes intuitive reasoning and promotes maintainable provable of systems. The structured design partitions a program into small, independent modules. They are arranged in a hierarchy that approximates a model of the business are and is organized in a top-down manner

Logical design proceeds in a top-down manner. General features, such as reports and inputs are identified first. Then each is studied individually and in more detail. Hence the structured design is an attempt to minimize the complexity and make a problem.

IV. System Working

Based upon the functions, the project had been divided to 6 modules which are mentioned as follows :

- Tweet Retrieval
- Text Preprocessing
- Named Entity Recognition
- Text Summarization
- Sentiment Calculation
- Data Visualization

A. Tweet Retrieval

In this first part, we will see different options to collect data from Twitter. In order to have access to Twitter data programmatically, we need to create an app that interacts with the Twitter API. Tweets are retrieved from twitter by using this API. Twitter provides REST APIs you can use to interact with their service. There is also a bunch of Python-based clients out there that we can use without re-inventing the wheel. In particular, Tweepy in one of the most interesting and straightforward to use.

B. Text Preprocessing

I start our analysis by breaking the text down into words. Tokenisation is one of the most basic, yet most important, steps in text analysis. The purpose of tokenisation is to split a stream of text into smaller units called tokens, usually words or phrases. While this is a well understood problem with several out-of-the-box solutions from popular libraries, Twitter data pose some challenges because of the nature of the language. The tokenisation is based on regular expressions. To improve the richness of your pre-processing pipeline, we can improve the regular expressions, or even employ more sophisticated techniques like Named Entity Recognition.

C. Named Entity Recognition

Named Entity Recognition (NER) is a subproblem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations and companies. NER is a fundamental task and it is the core of natural language processing (NLP) system. NER involves two tasks, which is firstly the identification of proper names in text, and secondly the classification of these names into a set of predefined categories of interest, such as person names, organizations (companies, government organisations, committees, etc), locations (cities, countries, rivers, etc), date and time expressions.

For humans, NER is intuitively simple, because many named entities are proper names and most of them have initial capital letters and can easily be recognized by that way, but for machine, it is so hard. One might think the named entities can be classified easily using dictionaries, because most of named entities are proper nouns, but this is a wrong opinion. As time passes, new proper nouns are created continuously.

Therefore, it is impossible to add all those proper nouns to a dictionary. Even though named entities are registered in the dictionary, it is not easy to decide their senses. Most problems in NER are that they have semantic ambiguity, on the other hand, a proper noun has Different senses according to the context.

D. Text Summarization

Summarization is a process of generating summaries by a computer program. Summarization process involves interpretation, transformation and generation.It is very difficult for human beings to manually summarize large documents of text. There is an abundance of text material available on the internet. However, usually the Internet provides more information than is needed. Therefore, a twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning.

E. Sentiment Calculation

Sentiment Calculation is a Natural Language Processing and Information Extraction task that aims to obtain writers feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of tweets. Generally speaking, sentiment calculation aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. In recent years, the exponential

increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment calculation today. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract latent public opinion and sentiment is a challenging task.

F. Data Visualization

While there are some options to create plots in Python using libraries like matplotlib or ggplot, one of the coolest libraries for data visualization is probably D3.js which is, as the name suggests, based on Javascript. D3 plays well with web standards like CSS and SVG, and allows to create some wonderful interactive visualizations. Data visualization is an important discipline in the bigger context of data analysis. By supporting visual representations of our data, we can provide interesting insights. We have discussed a relatively simple option to support data visualization with Python using Vincent.

V. IMPLEMENTATION

Implementation is the stage in the project where the theoretical design is turned into a working system and is giving confidence on the new system for users that it will work effectively and effectively. It involves careful planning, investigation of the current system and it constraints on implementation, design of achieve the changeover, an evolution of change over method. A part from planning major task of preparing the implementation are education and training of users .The more complex system being implemented, the more involved will be the system analysis and the design effort required just for implementation.

The product is developed in java environment in windows platform. Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale. Python supports multiple programming paradigms including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library.

VI. FUTURE SCOPE

In the proposed system all details about the system is designed into an application. It is designed as a web based application, which will work smoothly on any computer system with internet connection. For using/operating the application, there will be no need of additional staff. Any person with basic computer knowledge can use this application. The applications own operational staff is found to be sufficient and efficient to handle the system. Once the system is hosted on the web, the system will be automatically operated. Nowadays people are familiar with internet and e-mail, and hence, they can easily use the options from the home page of the new system. So the proposed system is operationally feasible.

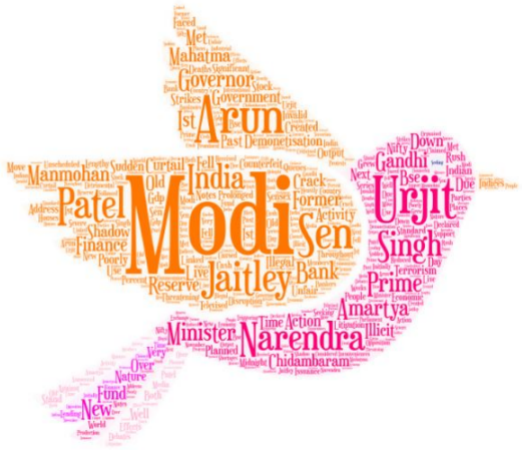

Fig. 1. News Maker detection

VII. CONCLUSION

In this Paper I have successfully processed the tweets to third party visualization applications. Text analysis focused on processing the tweets to extract information from the raw data of tweet, which can benefit the frontend application in projecting more information to the user, in terms of usability and exploring text-engineered data. In delivering a quality solution we conducted early tests on the application prototype, which is very useful to learn from mistakes and resolved the issues. The paper delivers a mechanism to facilitate NER (Name Entity Recognition) in extracting the annotation, Sentiment analysis over the text, significant phrase identification in a text and converting the physical address into Geographic coordinates (latitude and longitude) on tweets accessed from twitter in building a large twitter database, which is Readily available to use for various visualization tools and application back-end data. Am here to conclude we fulfilled the methodologies of possible information extraction by processing the data, to provide in-depth details of social networking data.

VIII. ACKNOWLEDGEMENT

This work would not have been possible without the support of many people. First and the foremost, we give thanks to Almighty God who gave us the inner strength, resource and ability to complete our project successfully.

I take this opportunity to express my sincere gratitude to the people who have been instrumental in the successful completion of the Paper.

I would like to express my sincere thanks to **Dr. Nagamani Molakatala, SCIS, University of Hyderabad**, who gave me the golden opportunity to do this wonderful project on the topic (Sensational News And News Maker Using Data Mining And Opinion Extraction), which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to her.

Many people, especially my friends have made valuable comment suggestions on this proposal which gave us an

inspiration to improve my work. I thank all the people for their help directly and indirectly to complete my project.

Lastly, I would like to thank each and every person who directly or indirectly helped me in the completion of the project especially my Parents and Peers who supported me throughout my project.

REFERENCES

[1] http://programminghistorian.org/lessons/output-data-as-html-file Accessed on 25/4/2021
[2] [2] https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/ Accessed on 04/3/2021
[3] [3] http://hadoopindepth.blogspot.in/ Accessed on 23/4/2021
[4] http://glowingpython.blogspot.in/2014/09/text-summarization-with-nltk.html Accessed on 24/05/2021