**SUBJECT DETECTION IN TEXT**

A

MAJOR PROJECT REPORT

*Submitted for the partial fulfillment of*

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

***Submitted by***

Abhimanyu Srivastava (9913103634)

Digendra Sharma (9913103472)

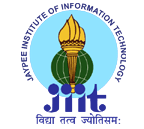
Yash Gupta (9913103480)

**Under the Guidance**

**of**

Prof. PulkitMehndiuatta

(Deptt of CSE/IT)



**Department of Computer Science and Engineering/IT**

**Jaypee Institute of Information Technology, Noida**

OCTOBER 2016

**TABLE OF CONTENTS**

Chapter 1: INTRODUCTION

Chapter 2: LITERATURE SURVEY

2.1: Sources

2.2: Summary of research paper

2.3: Summary of field survey

Chapter 3: RESULTS OF LITERATURE SURVEY

3.1: List of relevant problems

3.2: Problem Statement

3.3: Overview of proposed solution approach

Chapter 4: ANALYSIS, DESIGN AND MODELING

4.1: Functional & Non Functional Requirements

4.2: Overall Architecture

4.3: Proposed Algorithm

4.4: Test Plan

APPENDIX

A. Details of practice with new tool/technology

B. References

**1. Introduction**

In the world of computer science twitter has become one of the biggest web destinations for users to express their opinion and thoughts. Throughout the last few years twitter content continued to increase hundreds of millions of tweets are sent everyday by more than 285 million active users.

However, due to informal language used in twitter understanding opinions of users and performing such analysis become quite difficult. Companies leverage this unique ecosystem to tap into public opinion on their products and services and to provide real time costumer assistance. Traditionally companies rely on tools such as HootSuite sites. But these tools lack the efficiency and accuracy to determine opinion and thoughts on general public. So, in order to solve this problem we started with sarcasm detection as a major project but the major problem in sarcasm detection is to find context about which text is talking about and in order to find context we have to first determine subject so we have decided to first take up this issue relating to the efficient detection of subject. So we have done our major work on finding and proposing an efficient and optimized algorithm in subject detection. The data sets we are using are sarcastic tweets. So the base research papers we have used are on sarcasm detection.

**What is a Subject?**

A subject or a theme is what discourse, a discourse fragment or a sentence is about. It is the shortest summary of a discourse, the main proposition of paragraph or what is commented on in a sentence. The term subject is usually defined as the abruptness of a unit of discourse.

**How?**

Subject structure within a text can be tackled from two different points of view: from a Document Level (discourse subject) or a Sentence Level perspective (sentence subject). If a text is considered as a whole, it usually talks about single subject, but in a deeper analysis, several sub-subjects can be identified, giving additional information about the main subject. On the other hand, taking into account sentence structure, we can find that every sentence has a subject (the part of the structure that is being presenting) and a comment (what is being asserted about the subject). Next, these two-way perspectives are going to be explained in more detail.

**Positive**

In subject detection it is an application of basic language grammar that are used to detect a subject in a text. If you see we can make a lot of grammar rules for detecting a subject like before a noun there must be a/an/the so if it is a noun we can make some handset of rules for subject detection. It is challenging that we have to make our own grammar rules that are derived from our day to day English conversation.

So we are currently researching on how to make our own regular expression or basically a context free grammar which we have to use for subject detection. We are using Norman Lewis’s Better English and Wren and Martin’s High School Grammar as a reference books for the grammar part. It is a hard part to read and design regular expressions for subject detection as there are many set of rules defined in grammar therefore we have decided to take up this exciting challenge. Subject detection part is helping us to hone ourwriting skill and improved our spoken part also. Made us aware of all the tenses which are used in day to day life. This project is helpful in technical aspect as well as non-technical aspects. If this project is successful then it is also going to make us have a sound knowledge in English. On technical aspect it has given us basis of research work in natural language processing. Natural language processing is a new developing area and vast to explore and big organizations are running after it. So there is a huge scope of research development here This project has made us familiar with python programming which we extensively used for the web data crawling using tweepy libraries which uses the customers api and secret keys. We are now familiar with NLTK library which are used in the world of natural language processing. We are able to tokenize our data and then passed it through several filters like stemming, chunking, chinking, on our raw data mainly related to name entity reorganization in a sentence. So finally, we are able to move a step ahead with good knowledge in natural language processing research area which will help us to successfully complete this project.

**Negative**

On the negative part this challenge seems farfetched as English grammar rules are changing according to the usage of people in their daily life. So, it seems we are not moving in one definite direction as the challenge is not stationary as grammar rules are changing day by day according to common usage and more so on social networking sites such as twitter. There is also one challenge that if a statement contains pronoun then it is a tough part to detect anything about subject detection as the information we got related was worthless and as of now no use. So, it requires to have a sound foundation in theory of computation. This was only the one part, our proposal is that we need to optimize and develop new algorithm for the subject detection part which seem farfetched as of now because the data sets are real time fetched and it is growing day by day so it is difficult to have an optimized algorithm which can run on our raw data which is roughly around 5Lacs tweets. Therefore, we are trying very hard and have to go through lot of research work to overcome such demotivation and difficulties which are a hurdle in our way of completing this project.

We have been going through research papers in which there are few proposed algorithm which are proficient. So, the biggest challenge in this project is efficiency. Without efficiency this project will not fulfil its requirements which we have currently undertaken. Currently there are very less algorithms which have efficiency greater than 50% and our challenge is to make it more efficient than 50% which seems tough according to the researches we have done so far. So, therefore it seems a good challenge to undertake such projects.

**2. Literature Survey**

**2.1 Sources of Formulation**

www.Pythonprogramming.net

[www.NLTK.org](http://www.Nltk.org)

www.stackoverflow.com

RESEARCH PAPER

Sarcasm Detection in Twitter (Published - 2016 2 23)

Cited by “Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis”

Cited by “Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter”

**2.2 Summary of relevant papers**

Title of paper

Sarcasm detection in Twitter

Authors

**Mondher Bouazizi**

Graduate School of Science and Technology

Keio University

Yokohama, Japan

Email: bouazizi@ohtsuki.ics.keio.jp

**Tomoaki Ohtsuki**

Department of Information and Computer Science

Faculty of Science and Technology, Keio University

Yokohama, Japan

Email: ohtsuki@ics.keio.ac.jp

Year of Publication

2016

Publishing details where this paper was published

2015 IEEE Global Communications Conference, GLOBECOM 2015

Institute of Electrical and Electronics Engineers Inc.

ISBN9781479959525

Published - 2016 2 23

58th IEEE Global Communications Conference, GLOBECOM 2015 - San Diego, United States

**Summary**

Twitter has become one of the biggest web destinations for users to express their opinions and thoughts. Throughout the last few years, Twitter content continued to increase: hundreds of millions of tweets are sent everyday by more than 285million active users1. Many companies and organizations have been interested in these data for the purpose of studying the opinion of people towards political events [1], popular products [2] or movies [3].However, due to the informal language used in Twitter and the limitation in terms of characters (i.e., 140 characters per tweet), understanding the opinions of users and performing such analysis is quite difficult. Furthermore, presence of sarcasm makes the task even more challenging: sarcasm is when a person says something different from what he means. Some people are more sarcastic than others, however, in general, sarcasm is very common, though, and difficult to recognize. Oxford dictionary2 defines sarcasm as “the use of irony to make or convey contempt”. However, in general, people employ sarcasm in their daily life not only to make jokes and be humorous but also to criticize or make remarks about ideas, persons or events. Therefore, it tends to be widely used in social networks, in particular microblogging websites such as Twitter. That being the case, the state of the art approaches of sentiment analysis and opinion mining tend to have lower performances when analyzing data collected from such websites. Maynard et al. [4] show that sentiment analysis performance might be highly enhanced when sarcasm within the sarcastic statements is identified. Therefore, the need for an efficient way to detect sarcasm arises. In this paper, we propose an efficient way to detect sarcastic tweet. Although it does not need an already-build user knowledge-base as in the work of Rajadesingan et al. [5], our approach considers the different types of sarcasm and detect the sarcastic tweets regardless of their owners or their temporal context. The remainder of this paper is structured as follows. Section II presents our motivations and describes some of the related work. Section III describes in details our proposed method ,the different features we used and the tuning procedure of the parameters we defined. Section IV illustrates our experiments and results, and Section V concludes this work and proposes possible directions for future work.

Web link

https://keio.pure.elsevier.com/ja/publications/sarcasm-detection-in-twitter-all-your-products-are-incredibly-ama

**2.3 Summary of field survey, experimental studies new tools**

In this project we are using Pycharm (python IDE) for executing codes and downloading new libraries, Tweepy libraries to do web data crawling from twitter using the customer secret key and api key. We are basically implementing all the NLTK libraries used for natural language Processing (NLP) i.e. tech machines to read, understand, and analyze human natural language.

Some functionalities includes - tokenizing words and sentences with NLTK, Stemming words with NLTK, part of speech tagging with NLTK. Chunking and Chinking with NLTK and name entity recognition with NLTK.

**New Tools**

PyCharm

PyCharm is an Integrated Development Environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django.

PyCharm is cross-platform, with Windows, Mac OS X and Linux versions. The Community Edition is released under the Apache License and there is also Professional Edition released under a proprietary license - this has extra features.

**3. Results of Literature Survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NAME OF THE PAPER** | **PUBLICATION DATE** | **TECHNIQUE** | **FEATURE** | **ACCURACY** |
| Sarcasm detection in Twitter | 23 February 2016 | POS Tagging | Sarcasm Detection | **>**90 % |

**SUBJECT DETECTION IN TEXT**

**↓**

**DATA CRAWLING**

**↓**

**TOKENIZING**

**↓**

**CLEANING**

**↓**

**FILTERATION**

**↓**

**POS TAGGER**

**↓**

**CHUNKING↔CHINKING**

FLOW DIAGRAM OF COMPLETE PROJECT

**3.2** **Relevant current/open problems**

The known method to convert a sentence into Noun and Verb Phrases (or in other words – a tree..) is parsing. However, the problem with parsing algorithms is that their complexity is quite bad.

The second problem is that full-parsing was a bit of an overkill for what we wanted to achieve.

A noun or a noun phrase will always be subject in more general sense, but a sentence can have many nouns and noun phrases.

Syntax is not sufficient. One would have to use some form of shallow semantic analysis to identify what you call the "subject".

**3.3 Problem statement**

Filtering of the clustered raw data and proper clustering under POS Tagging. Removing worthless data such as repetitions and special symbols otherwise proper Tags cannot be made.

**3.4 Overview of proposed solution approach and Novelty/benefits**

We are using POS Tagging for categorizing all the relevant words in a sentence and assign different tags according to their categories and sub-categories of part of speech which helped us in finding the subject of the sentence.

We are using grammar rules with the help of regular expressions to implement all the rules of English grammar in finding the subject.

Benefits of proposed solution

Using POS Tagging helped us to categorize and tag all the words keeping in view the context in which they are used and categories and sub-categories of part of speech.

**3.5 Task division among group members**

Digendra Sharma

Has done a lot of researches on the subject detection part.

Has helped in designing grammar rules for the subject detection.

Has used the tweepy library extensively for web data crawling using customers secret and api key.

Has done POS Tagging to categorize words under different tags.

Abhimanyu Srivastava

Has done most of the coding in python using the NLTK library effectively.

Has form all his researches basis based on NLTK library of natural language processing.

Working on developing efficient algorithm for the subject detection part.

Yash Gupta

Has done most of the coding in python using the NLTK library effectively.

Has form all his researches bases based on NLTK library of natural language processing.

Working on developing efficient algorithm for the subject detection part.

**4. Analysis, Design and Modeling**

**4.1 Overall description of the project**

In order to solve this problem we started with sarcasm detection as a major project but the major problem in sarcasm detection is to find context about which text is talking about and in order to find context we have to first determine subject so we have decided to first take up this issue relating to the efficient detection of subject. So we have done our major work on finding and proposing an efficient and optimized algorithm in subject detection. The data sets we are using are sarcastic tweets. So the base research papers we have used are on sarcasm detection

A subject or a theme is what discourse, a discourse fragment or a sentence is about. It is the shortest summary of a discourse, the main proposition of a paragraph or what is commented on in a sentence. The term subject is usually defined as the about ness of a unit of discourse

Subject structure within a text can be tackled from two different points of view: from a Document Level (discourse subject) or a Sentence Level perspective (sentence subject). If a text is considered as a whole, it usually talks about a single subject, but in a deeper analysis, several sub-subjects can be identified, giving additional information about the main subject. On the other hand, taking into account sentence structure, we can find that every sentence has a subject (the part of the structure that is being presenting) and a comment (what is being asserted about the subject).

In subject detection it is an application of basic language grammar that are used to detect a subject in a text. If you see we can make a lot of grammar rules for detecting a subject like before a noun there must be a/an/the so if it is a noun we can make some handset of rules for subject detection. It is challenging that we have to make our own grammar rules that are derived from our day to day English conversation.

So we are currently researching on how to make our own regular expression or basically a context free grammar which we have to use for subject detection. We are using Norman Lewis’s Better English and Wren and Martin’s High School Grammar as a reference books for the grammar part. It is a hard part to read and design regular expressions for subject detection as there are many set of rules defined in grammar therefore we have decided to take up this exciting challenge. Subject detection part is helping us to hone our writing skill and improved our spoken part also. Made us aware of all the tenses which are used in day to day life. This project is helpful in technical aspect as well as non-technical aspects. If this project is successful then it is also going to make us have a sound knowledge in English. On technical aspect it has given us basis of research work in natural language processing. Natural language processing is a new developing area and vast to explore and big organizations are running after it. So there is a huge scope of research development here This project has made us familiar with python programming which we extensively used for the web data crawling using tweepy libraries which uses the customers api and secret keys. We are now familiar with NLTK library which are used in the world of natural language processing. We are able to tokenize our data and then passed it through several filters like stemming, chunking, chinking, on our raw data mainly related to name entity reorganization in a sentence. So finally, we are able to move a step ahead with good knowledge in natural language processing research area which will help us to successfully complete this project.

**4.2 Functional requirements and Non Functional requirements**

FUNCTIONAL REQUIREMENTS

1. POS Tagging

Categorizing and tokenizing the words of the sentence.

2. Chunking

Divide sentence into chunks

3. Chinking

Removing unwanted words from the chunked data

4. Name entity recognition

Helping in recognizing the subject.

5. Crawling

Extracting data (tweets) from twitter.

6. JSON Parsing

NON-FUNCTIONAL REQUIREMENTS

1. Internet Connection

Extracting the data from networking site i.e. Twitter.

2. Hardware

Malfunctioning of computer/laptop while running the code.

3. Browser

Speed of a particular browser affects crawling of data.

**4.3 Risk Analysis**

Risk in our proposed project is to find sarcasm in the tweets efficiently .Our aim is to achieve efficiency of above 80%, but if we failed to achieve our target we can try to achieve efficiency above 50% on the safer side we started our approach by detecting subject in the text and we are confident enough to achieve this.

**4.4 Test Plan**

In order to test the performance of the proposed code and implementation, we will use a single dataset in our program and compare the results with manually checked subject of the same text.

**4.5 Implementation**

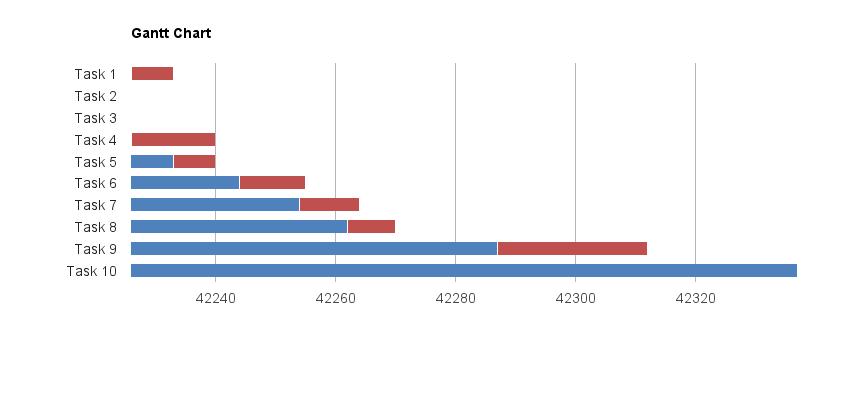
* First we have crawled twitter API data into a JSON file and then converted it into text file using json parsing in the code.
* Then we have used puncked sentence tokenizer to convert the raw data of the text file into separate sentences.
* Next we have tagged each sentence to its part of speech using (POS Tagging).
* We have then used chunking and chinking to determine subject in the text and remove worthless data.

**Appendix**

1. Project Plan as Gantt chart or WBS

**GANTT CHART**

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Start Date | Task Data | Days To Complete |
| Task 1 | 15/8/16 | Designing the Skeleton of the project | 7 |
| Task 2 | 22/8/16 | Learning basics of NLTK and Python | 7 |
| Task 3 | 29/8/16 | Learning the concept of POS Tagging and chunking and chinking | 7 |
| Task 4 | 5/9/16 | Crawling the data from the twitter | 7 |
| Task 5 | 12/9/16 | How to improve the existing algorithm compared to the rest in the market | 7 |
| Task 6 | 19/9/16 | Looking into resources of designing the algorithm | 7 |
| Task 7 | 29/9/13 | Applying POS Tagging, chunking and chinking | 10 |
| Task 8 | 7/10/16 | Implementing the code to find subject of the text | 8 |



B. References

www.Pythonprogramming.net

[www.NLTK.org](http://www.Nltk.org)

www.stackoverflow.com

RESEARCH PAPER

Sarcasm Detection in Twitter (Published - 2016 2 23)

Cited by “Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis”

Cited by “Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter”