**CHAPTER 1**

**INTRODUCTION**

## 1.1 Introduction

For our final project, I have chosen to analyze a movie reviews. What I will be extracting from the data set is the significance of attributes that result in a large gross revenue of a movie. The goal of analyzing this data set is to successfully figure out which attributes are the most significant when determining future success of a movie title before it is released. Critics and human instinct, when it comes to movies, is sometimes unreliable. I want to be able to accurately predict what attributes in fluence movie success based on several characteristics in specific areas such as the IMDB website.

**1.2 Motivation and Problem Statement**

An article published in a newspaper or magazine that describes and evaluates a movie. Reviews are typically written by journalists giving their opinion of the movie. For many of us, reviews are like one written by our friends on facebook, are important in making our decision to watch a movie.

Similarly, these reviews are available to movie production companies which helps them- To understand sentiment and check the popularity of their films ¬to figure out new marketing strategies and future directions. Human mind can read and understand whether a review is positive but for movie studios it is difficult to hire employees to simply read and judge movie opinions. So here comes Machine Learning to rescue - to process, reliably extract and classify the sentiment of unstructured movie reviews. Regarding this scenario, sentiment analysis is a brilliant term of artificial intelligence for understanding the user or audience’s sentiment in the basis of polarity and subjectivity.

IMDb Movie has a lot of users and the reviews of a movie users are increasing day by day. “Sentiment Analysis on IMDb Movie Reviews” scraps the review data as data set and completes the classification of review texts by subjectivity or objectivity and negative or positive attitude of users. This project presents and empirical study of efficacy of classifying movie review.

Thus this is a completely different approach by removing the unstructured data and then classifying reviews employing Natural Language Processing toolkit.

**1.3 Project Scope**

This section will describe the scope of the project, and briefly outline the preconditions and constraints that narrows down the focus area of this project.

Our first experiments, uses a dataset based on positive, negative and neutral movie reviews from IMDb. We gathered reviews from IMDb website and other available resources.Our dataset contains 150000 reviews. It is intended as a larger benchmark dataset for sentiment classification, and consists of movie reviews, gathered from IMDb.

Our second experiments, uses a dataset based on sarcastic and non-sarcastic movie reviews from IMDb. We gathered reviews from IMDb website and other available resources. Our dataset contains 100000 reviews.

The age of the datasets, does not concern us, since the theories tested in this thesis, can be applied to newer datasets without any problems. The trustworthiness of the documents and data sources will not be a focus of this project.

Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Researchers and Businesses are very interested to understand the thoughts of people and how they respond to everything happening around them. Movies Industries can use this project to evaluate their movies and services and to improve their movies and services. Sentiment analysis by this project has more possibility and accuracy rather than manual analysis of sentiment data virtually as we have included sarcasm detection along with sentiment analysis to improve the accuracy.

Sarcasm detection can help us understand the polarity of a review correctly. The users sometimes tends to express his/her sentiment in the form of sarcastic utterances. The sarcastic utterance usually shifts the polarity of text from negative to positive and likewise.

There is a lot of scope in analyzing the users review on the web. Now-a-days, there are lots of website which post the reviews of different movies provided by different reputed movies industries. Sentiment analysis will have to pace up with this change. “Sentiment Analysis on IMDb movie reviews” can help the movie industries to change strategies based on user feedback.

**1.4 Project Objective**

The objective of this project is to show how sentimental analysis can help improving the user experience over a system interface. The learning algorithm will learn what users’ emotions are from statistical data then determine the sentiment and if there is any utterances of sarcastic review. Reviews will be classified into positive, negative and neutral sentiments. “Sentiment Analysis on IMDb Movie Reviews” is the process of user review data mining, analyzing of users feedback and display a graph of the result which presents the polarity values basis of positivity and negativity.

**1.5 Impact, Significant and Contribution**

We’ve put impact, Significantand contribution in the center of our feedback cycle. The ability to understand positive feelings, or negative feelings has been difficult, for machines that lack feelings. Movie industries want to accommodate the sentiment analysis tools into areas of user feedback, marketing. Al though, by having this project, the industries will able to track the users and fulfill their parameter. The industries can easily detect the sentiment of the users by analyzing their review data.

**1.6 Background Analysis**

Creating a predictive model fort his data set is not vital to human existence, however it would be useful for some movie- goers. This analyzation pertains to the entertainment/movie industry. It can help producers, actors, actresses, directors, film investors, and movie-goers determine how successful the proposed movie will be. With out the predictive modeling, there would only be gut decisions/personal preferences about how a movie will turn out. Not everyone thinks that a certain actor or actress is amazing, therefore saying the entirety of the movie is amazing. Putting it in terms of analytical processing makes the prediction more stable and unbiased. This project would be deemed significant to this group of people mentioned previously because it will be an unbiased predictive data set that will be utilized to determine gross revenue. Every producer and director believe their movie will be one of the greatest, and they will do everything in their power to make it the greatest. However, majority of the time, this turns out to be false. They can take this data set and implement it into their thought process when planning their movie. On the flipside, I hear a lot of the time that people will go see a movie and say I just wasted x amount of money to see that horrible film!”. Movie-goers can use this data set to make the same predictions once the movie is announced with primary and supporting actors/actresses. It could possibly save movie-goers money when debating on whether to go see a movie or not.

**Goals:**

There are a couple of goals that I wish to achieve with this data set. The goals I wish to

achieve are:

* Assist directors and producers in maximizing their potential revenue of a proposed film
* Save money or spend money wisely when debating on seeing a new film
* Gain practice in using multiple linear regression
* Develop more skill in pre-processing techniques such as data partitioning and handling missing data
* Learn more about post processing technique sensitivity analysis

2

**CHAPTER 2**

**LITERATURE REVIEW**

## 2.1 Introduction

The literature review is an important part of a Project. It should be thorough and accurate. This project going to review on the article, book and internal resources to study about the “Sentiment Analysis on IMDb movie reviews”.

* An analysis of temporal multivariate networks derived from IMDB – Used methods such as (p, q)-core and 4-ring to identify subgraphs and short cycles
* An analysis of how an individuals movie preferences correlated with Oscar winning titles – Used Linear Regression
* An analysis a movie dataset using regression and k-nearest neighbor methods

Literature Review

* An analysis a movie dataset using regression and k-nearest neighbor methods

* + 1. **Natural Language Processing**

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI). The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

* + 1. **Natural Language Toolkit (NLTK)**

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). it provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language”. It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample data sets as well as accompanied by a cook book and a book which explains the principles behind the underlying language processing tasks that NLTK supports.

**The Good:**

* Makes NLProc easier and more accessible
* Python (great learning language)
* Lots of documentation (and 2 books!)
* Designed for training custom models
* Includes many training corpora
* Many algorithms to experiment with

**The Bad:**

* NLProc is hard
* Few out-of-the-box solutions (see Pattern)
* Not designed for big-data (see Mahout)
* Doesn’t have latest algorithms (see Scikits-Learn)
* No online or active learning algorithms
  + 1. **Stop Words**

Stop words are generally thought to be a “single set of words”. It really can mean different things to different applications. For example, in some applications removing all stop words right from determiners (e.g. the, a, an) to prepositions (e.g. above, across, before) to some adjectives (e.g. good, nice) can be an appropriate stop word list. To some applications however, this can be detrimental. For instance, in sentiment analysis removing adjective terms such as ‘good’ and ‘nice’ as well as negations such as ‘not’ can throw algorithms off their tracks. In such cases, one can choose to use a minimal stop list consisting of just determiners or determiners with prepositions or just coordinating conjunctions depending on the needs of the application. Examples of minimal stop word lists that can use:

* **Determiners** –Determiners tend to mark nouns where a determiner usually will be followed by a noun.  
  Examples: the, a, an, another
* **Coordinating conjunctions**– Coordinating conjunctions connect words, phrases, and clauses.  
  Examples: for, an, nor, but, or, yet, so
* **Prepositions** – Prepositions express temporal or spatial relations.  
  Examples: in, under, towards, before

In some domain specific cases, such as clinical texts, we may want a whole different set of stop words. For example, terms like “mcg” “dr” and “patient” may have less discriminating power in building intelligent applications compared to terms such as ‘heart’ ‘failure’ and ‘diabetes’. In such cases, we can also construct domain specific stop words as opposed to using a published stop word list.

* + 1. **NLTK Stop Words**

Natural Language Processing with Python Natural language processing (NLP) is a research field that presents many challenges such as natural language understanding. Text may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the text to be processed. There is no universal list of stop words in NLP research, however the NLTK module contains a list of stop words.

* + 1. **Lemmatization**

Lemmatization is an important aspect of natural language understanding ([NLU](https://searchenterpriseai.techtarget.com/definition/natural-language-understanding-NLU)) and natural language processing ([NLP](https://searchbusinessanalytics.techtarget.com/definition/natural-language-processing-NLP)) and plays an important role in [big data analytics](https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics) and artificial intelligence ([AI](https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence)). Complex algorithms use the rules of linguistic morphology, in context with a particular language's vocabulary, to group words used in speech and writing by inflected forms. [Deep learning](https://searchenterpriseai.techtarget.com/definition/deep-learning-deep-neural-network)is used to analyze and understand the grouping as a whole, so when any inflectional form of a word is mentioned, the base term's entire lemmatization is included.

In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. The goal of lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

For instance:

am, are, is $\Rightarrow$ be  
car, cars, car's, cars' $\Rightarrow$ car

The result of this mapping of text will be something like:

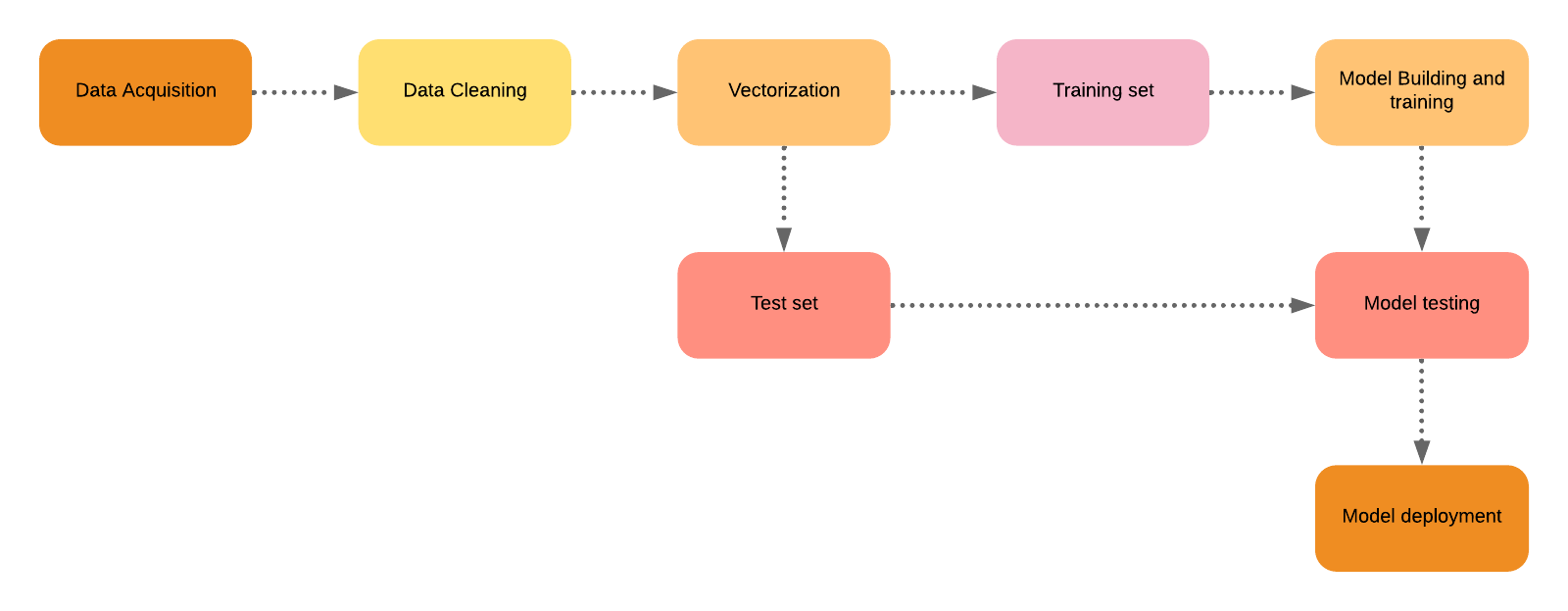
the boy's cars are different colors $\Rightarrow$the boy car be different color

* + 1. **sklearn**

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including naïve bayes, logistic regression, support vector machine, k-means etc, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

We have used naïve bayes and logistic regression from scikit-learn for classifying our sentiment and sarcasm data

* 1. **Methodology**
* **Data acquisition:** We will collect reviews from IMDb website and prepare a dataset.
* **Data pre-processing:** The data gathered will be in a raw format and this data won’t be feasible for the analysis. Therefore, certain steps will be executed to convert the data into a small clean data set.



* **Vectorization:** Pre-processed data will be converted into numbers using Vectorization model e.g. Bag-of-Words, TF-IDF vectorizer.
* **Splitting data into training set and testing set:** Data will be split into training and testing set (80-20%).
* **Model building, training and testing:** An important point to note is that during training the classifier only the training set is available. The test data set won’t be used during training the classifier. The test set will only be available during testing the classifier.

**CHAPTER 3**

**SYSTEM IMPLEMENT PROCESS**

## 3.1 Process

The steps of our system model.

Those are:

* The first step is data scrapping from “………….” using a script to scrape review as dataset.
* The second step is analyzing data to sort out the number of words, characters and other attributes to be needed for next steps.
* The third step is stop words count for natural language processing.
* The fourth step is data cleaning for reducing of unwanted data\symbols or special characters from the dataset.

The fifth step is lemmatization of the words for grouping together the inflected from of a word.



**Figure-3.1:** Process of system Implementation

## 3.2 Tools

The main tools used to develop this project is python (OOP Language), Python library, package and module: NumPy, matplotlib, pandas, re, nltk.corpus and nltk.

**Python:** Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

**NumPy:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

**Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.

**Pandas:** Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

**Regular Expression (RE):** Regularexpression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern. Usually such patterns are used by string searching algorithms for "find" or "find and replace" operations on strings, or for input validation.Regular expressions are widely used in UNIX world. The Python module re provides full support for Perl-like regular expressions in Python.

**NLTK:** The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). it provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

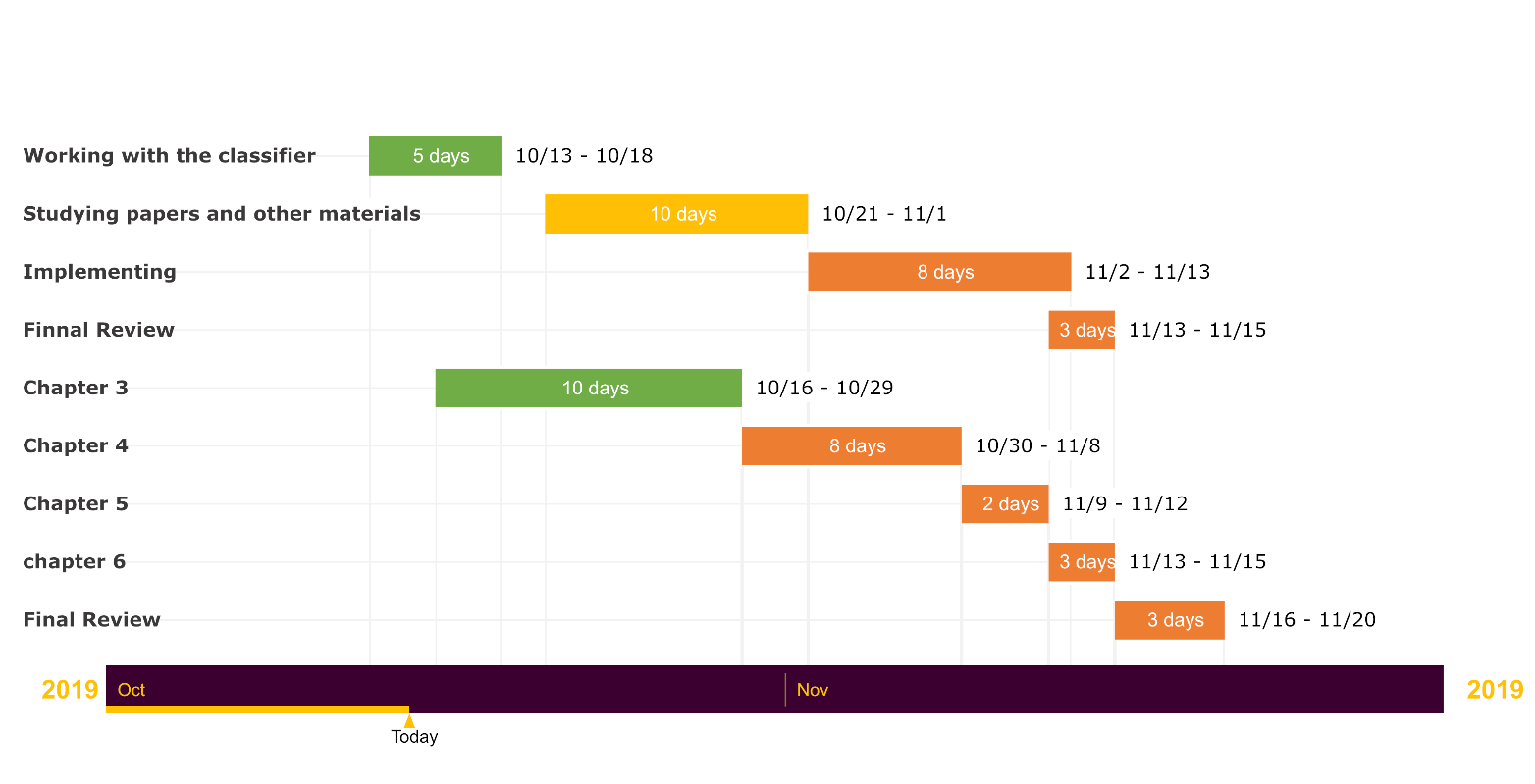
**NLTK.Corpus:** The NLTK corpus is a massive dump of all kinds of natural language data sets that are definitely worth taking a look at. Almost all of the files in the NLTK corpus follow the same rules for accessing them by using the NLTK module, but nothing is magical about them. These files are plain text files for the most part, some are XML and some are other formats, but they are all accessible by you manually, or via the module and Python.

**3.3 Implementation Issues and Challenges**

Sentiment analysis classifies text as positive, negative or else objective, so it can be thought as text classification task. Text classification has many classes as there are many topics but sentiment analysis has only three classes. However, there are many factors that make sentiment analysis difficult compared to traditional text classification.

**3.4 Timeline**

This project is estimated to be complete in a period of around 8 months. The following Grant Chart will show the timeline for each phases of the project. There are seven phases in this project and take different duration to complete. Working with the Classifier will take around 05 days to complete, Studying Papers and other Materials will take 10 days to complete, Implementing to complete in 08 days and Final Review to complete in 03 days, Chapter 04 to complete in 08 days, Chapter 05 to complete in 02 days , Chapter 06 to complete in 03 days. When the Chapter 06 was done, the Final Review will carry out, each of it will take around 03 days.



**Table-3.4:** Gantt chart of system Implementation

**3.5 Requirement Specification**

A software requirements specification (SRS) is a description of a software system to be developed. It is modeled after business requirementsspecification (CONOPS), also known as a stakeholder requirements specification (SRS). The software requirements specification lays out functional and non-functional requirements, and it may include a set of use cases that describe user interactions that the software must provide to the user for perfect interaction.

**3.6 Data Collection**

The raw IMDb dataset is structured in such a way that most of its attributes and information is organized and stored separately in compressed plain text files. For instance, all of the roughly 150,000 movie ratings from the database are stored in the compressed text file ratings. List (e.g. ratings.list.gz), which includes textual information about the data as well as a table of film rank, the number of votes and film titles. Thus, some sort of cleaning, integration and preprocessing is likely to be required in order to make good use of the data for the purpose of data mining through supervised machine learning techniques. The data was collected using IMDB (java movie database) which contains the IMDB movie dataset of more than 30,00000 movies in the dataset. The dataset was transferred to MySQL, in form of tables.

**3.7 Data Analysis**

The dataset is divided into training dataset and test dataset which contains the classes like Hit, Flop and Average and predicting variables like actor, actress, composer, genre, director producer and music director k-means clustering is used to analyze the training dataset to develop models which can be used for test dataset for analysis decision tree algorithm is used for predicting which factors.

**CHAPTER 4**

**SYSTEM IMPLEMENTATION**

**REFERENCES**

[1] E. Cambria, et al., "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, vol. 28, no. 2, 2013, pp.15-21.

[2] E. Cambria, "Affective computing and sentiment analysis," IEEE Intelligent Systems 31.2 (2016): pp. 102-107.

[3] P. Anantharam, K. Thirunarayan, and A. Sheth, "Topical anomaly detection from twitter stream," Proceedings of the 4th Annual ACM Web Science Conference, ACM, 2012.

[4] L. Chen, et al., "Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter," Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[5] L. Chen, W. Wang, and A. Sheth, "Are Twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican Presidential Primaries," International Conference.