LINGUISTIC PROCESSOR:

REALTIME EXTRACTION AND ANNOTATION OF USER GENERATED CONTENTS

**Tajammal Amin**

**Sohaib Ahmad**



DEPARTMENT OF COMPUTER SCIENCES

COMSATS UNIVERSITY ISLAMABAD, VEHARI CAMPUS

VEHARI – PAKISTAN

SESSION 2018-2021

LINGUISTIC PROCESSOR:

REALTIME EXTRACTION AND ANNOTATION OF USER GENERATED CONTENTS

*Undertaken By:*

**Tajammal Amin**

Reg. No. CIIT/FA18 -bCS-093/Vhr

**Sohaib Ahmad**

Reg. No. CIIT/FA18 -bCS-131/Vhr

*Supervised By:*

**Arslan Ali Raza**



A DISSERTATION SUBMITTED AS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELORS IN COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCES

COMSATS UNIVESITY ISLAMABAD, VEHARI CAMPUS

VEHARI – PAKISTAN

SESSION 2018-2022

**CERTIFICATE OF APPROVAL**

It is to certify that the final year project of BS (CS) “Linguistic processor: Realtime extraction and annotation of user generated contents” was developed by

**Tajammal Amin (CIIT/FA18-BCS/ 093)** and **Sohaib Ahmad**

**(CIIT/FA18-BCS-131)** under the supervision of “ARSLAN ALI RAZA” and that in his opinion; it is fully adequate, in scope and quality for the degree of Bachelors of Science in Computer Sciences.

---------------------------------------

**Supervisor**

---------------------------------------

**External Examiner**

---------------------------------------

**Head of Department**

**(Department of Computer Science)**

**DEDICATION**

We would like to dedicate this project to our parents and supervisor Sir Arslan Ali Raza for their endless love and efforts.

**ACKNOWLEDGEMENT**

Saying of **The Holy Prophet Muhammad (S.A.W)** a person who is not thankful to his benefactor is not thankful to **Allah**. All and every kind of praise is upon **Allah Almighty**, the strength of Universe, forever help in darkness and difficulties.

All and every kind of respect goes to **The Holy Prophet Muhammad** (S.A.W) for being a unique, comprehensive, and everlasting source of guidance and knowledge for humanity. We deem supreme pleasure to be able to express the heartiest gratitude and deep sense of devotion to our respected and worthy supervisor **Sir Arslan Ali Raza** for his kind guidance, useful suggestions, inspiring nature and unfailing patience as well as giving us extraordinary experiences throughout our project without which this work would have never been materialized and this dissertation would have not been possible. Above all and the most needed, he provided us with unflinching encouragement and support in various ways.

We deem it utmost pleasure in expressing my gratitude with the profound thanks to **Dr. Muhammad Ali Shahid,** HOD Computer Science, COMSATS University Islamabad, Vehari campus for providing us with strategic command at every step. we extend deep emotions of appreciation, gratitude, and indebtedness for his valuable guidance. We also acknowledge the members of our evaluation committee, **Dr. Qasim Umar, Dr. Muhammad Mudassar** and **Sir Yasir Munir**. They paid special attention regarding our project and enable us to reach at this destination.

Words are lacking to express our humble obligations to our parents help and prayers. Our deep thanks for our parents without their enduring support and advice we would not be able to complete our work.

**PROJECT BRIEF**

PROJECT NAME LINGUISTIC PROCESSOR

ORGANIZATION NAME CUI, VEHARI

OBJECTIVE REALTIME EXTRACTION AND ANNOTATION OF USER GENERATED CONTENTS

UNDERTAKEN BY TAJAMMAL AMIN

SOHAIB AHMAD

SUPERVISED BY ARSLAN ALI RAZA

LECTURER

DEPARTMENT COMPUTER SCIENCE

COMSATS UNIVERSITY ISLAMABAD VEHARI CAMPUS

Started On 20 SEPTEMBER 2021

Completed On -CONTINUE (EXPECTED SPRING 2022)

COMPUTER USED DELL LATITUDE E7240

HP ENVY

SOURCE LANGUAGE PYTHON

REAT.JS

OPERATING SYSTEM WINDOWS 10

TOOLS USED MICROSOFT VISUAL STUDIO CODE & NLTK

**ABSTRACT**

With the exponential growth of World Wide Web and rapid expansion of e-commerce, web opinion sources such as merchant sites, forums, discussion groups and blogs are used as a platform by individual users to share experiences or opinions. Online merchant sites provide space for customers to write feedback about their product and services, as a result number of customer reviews grow rapidly for each product. Such reviews are useful for customers in making purchase decision regarding a product based on the experiences of the existing users, whereas on the other hand, it helps product manufacturers in assessing strength and weaknesses of their products from the perspective of end users. Such information is very useful in developing marketing and product development plans.

“The aim of this project is to perform real time extraction, cleaning, normalization of contextualized and country specific user generated contents from different social media channels. Further aim is to pre-process and annotate extracted contents with subjective labels”.

# Contents

[Contents ix](#_Toc93178664)

[1. Introduction 1](#_Toc93178667)

[1.1 System Introduction 1](#_Toc93178668)

[1.2 Literature Review 2](#_Toc93178669)

[1.2.1 Tweepy 2](#_Toc93178670)

[1.2.2 TAG V6.1 3](#_Toc93178671)

[1.2.3 Selenium 3](#_Toc93178672)

[1.3 Objectives of the System 4](#_Toc93178673)

[1.4 Significance of the System 4](#_Toc93178674)

[2. Overall Description 5](#_Toc93178675)

[2.1 Product Perspective 5](#_Toc93178676)

[2.2 Product Scope 5](#_Toc93178677)

[2.3 Product Functionality 6](#_Toc93178678)

[2.4 Operating Environment 6](#_Toc93178679)

[3. Problem Statement 7](#_Toc93178680)

[3.1 Possible solution 7](#_Toc93178681)

[4. Methodology 9](#_Toc93178682)

[4.1 Text extraction: 9](#_Toc93178683)

[4.1.1 Twitter 9](#_Toc93178684)

[4.1.2 YouTube 9](#_Toc93178685)

[4.1.3 Blogging Websites 9](#_Toc93178686)

[4.2 Tokenization 10](#_Toc93178687)

[4.3 Stopwords removal 10](#_Toc93178688)

[4.4 Lemmatization 10](#_Toc93178689)

[4.5 Syntactical Analysis 10](#_Toc93178690)

[4.6 Subjectivity Classification 10](#_Toc93178691)

[4.7 Tools and technologies 10](#_Toc93178692)

[4.8 ReactJS 11](#_Toc93178693)

[4.8.1 Single page applications 11](#_Toc93178694)

[4.8.2 Data Binding 11](#_Toc93178695)

[4.8.3 Strong community support 11](#_Toc93178696)

[4.9 NLTK 11](#_Toc93178697)

[4.10 Django 11](#_Toc93178698)

[4.10.1 Pluggable 11](#_Toc93178700)

[4.10.2 ORM 12](#_Toc93178702)

[4.11 Algorithm Viewpoint 12](#_Toc93178703)

[5. Implementation 13](#_Toc93178704)

[5.1 Software interfaces 13](#_Toc93178705)

[6. Conclusion 18](#_Toc93178706)

[7. References 19](#_Toc93178707)

# LIST OF FIGURES

Figure ‎3.1Research design 8

Figure ‎5.1Sing in 13

[Figure ‎5.2 Sign up 14](#_Toc21248648)

[Figure 5.3 Home page/Twitter 14](#_Toc21248649)

[Figure 5.4 Home page/YouTube 15](#_Toc21248649)

[Figure 5.5 Home page/Blogging websites 15](#_Toc21248649)

[Figure 5.6 About us 16](#_Toc21248649)

[Figure 5.7 Contact us 17](#_Toc21248649)

# LIST OF TABLES

[Table ‎5.1 Tools and technologies](#_Toc21248675) 13

# Introduction

## System Introduction

The dawn of the internet opened the doors to the easy and widespread sharing of information on subject matters such as products, services, events and political opinions. The amount of publicly available text data is increasing every day and considerable number of resources are exhausted to find the data that matches the project requirements. In majority of cases the extracted data is in the raw form and requires dedicated pre-processing before performing any analysis or visualizations, which makes this whole process laborious and time taking.

As the most natural form of storing and exchanging information is written words, text mining (TM) has a very high commercial potential. In fact, a recent study indicated that 80% of a company's information is contained in text documents, such as emails, memos, customer correspondence, and reports [1]

Traditional document and text management tools are inadequate to meet the utilities. Document management systems work well with homogeneous collections of documents but not with the heterogeneous mix that knowledge workers face every day. Even the best Internet search tools suffer from poor precision and recall. (Precision is a measure of how many documents returned from a search actually meet the intended query criterion. Recall measures the percentage of documents returned versus how many should have been returned.) [2]

We present social media linguistic processor which is a modern, sophisticated and interactive web application that combines the steps of real time text extraction, cleaning, pre-processing, normalization and annotation of extracted contents in terms of subjective labels in a single step. An end user will be able to perform all required text processing by his/her own.

* Crawling of contents from different domain channels i.e., Twitter, YouTube and Blogging websites
* Text Cleaning
  + Noise reduction
  + Removal of URLs and undesired and symbols
  + Handling redundant contents
* Tokenization
  + Sentence level tokenization
  + Word level tokenization
* Stopwords removal
* Lemmatization
  + Conversion to root form
* Annotation
  + Syntactical annotation
  + Subjectivity classification

After text extraction the most crucial step is to perform cleaning as it improves overall efficiency significantly. Tokenization is used to tokenize the paragraphs into sentences and sentences into words. Lemmatization is a process to remove inflections and map a word to its root form. Stop words removal is performed to remove the words that don’t contribute to the meaning of sentence. Syntactical annotation is performed to check if the sentence has any feature that can declare it as subjective i.e., adjectives. Sentence level subjectivity classification is done to classify sentences to be subjective or objective.

## Literature Review

With the exponential growth of World Wide Web and rapid expansion of e-commerce, web opinion sources such as merchant sites, forums, discussion groups and blogs are used as a platform by individual users to share experiences or opinions. Online merchant sites provide space for customers to write feedback about their product and services, as a result number of customer reviews grow rapidly for each product and we have a huge amount of data publicly available online.

Social networks surfaced as communication and socialization tools. The vast amount of data these networks generate has led to a growing need of automatic knowledge extraction. The popular nature of these services is ideal for trends discovery. [3]

Twitter offers an open environment where people all around the world share information and opinions, emerging as a real-time repository of knowledge that can be exploited by researchers and applications. Tools like Twitter’s official API, Tag V6.1, selenium bots are available to perform twitter data extraction. All of these mentioned tools have certain limitations.

### Tweepy

Tweepy is an open-source Python package that gives you a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter’s models and API endpoints, and it transparently handles various implementation details, such as:

* Data encoding and decoding
* HTTP requests
* Results pagination
* OAuth authentication
* Rate limits
* Streams

If we weren’t using Tweepy, then you would have to deal with low-level details having to do with HTTP requests, data serialization, authentication, and rate limits. This could be time consuming and prone to error. Instead, thanks to Tweepy, we can focus on the functionality you want to build.

#### Advantages

The Twitter API give developers access to most of Twitter’s functionality, some of the advantages of using twitter for data extraction are as follows,

* As it is provided by Twitter itself so it has a very extensive and through documentation and very huge online community all these things ensure a very smooth development life cycle.
* Tweepy directly interacts with twitter database without any third-party dependency so it completes the task in short time span as compared to other data extraction tools.
* Data extracted using Tweepy always has more credibility as compared to Twint, tag v6 etc.

#### Limitations

* Twitter’s API limits you to the last 3200 tweets in a timeline.
* Doesn’t provide any method to determine the country from where the tweet was made
* The standard API only allows you to retrieve tweets that were made 7 days ago and is limited to scraping 18,000 tweets per a 15-minute window.
* Returns data in the json format which requires additional pre-processing i.e., cleaning, normalization etc.

### TAG V6.1

TAGS is a free Google Sheet template which lets you setup and run automated collection of search results from Twitter.

#### Advantages

* Doesn’t require any API key or access token
* Provides information about city/country of user
* Comes with an interactive and easy to learn GUI i.e., doesn’t require to write code

#### Limitations

* Unlike Twint or Tweepy it’s not an API rather it’s a google sheet hosted online
* As it’s a google sheet so can’t be automated or used in some other project
* Has a rate limit of 18000 tweets
* There must be a gap of around 30 mins between two consecutive queries
* Can only scrape data from last 8-9 days
* No official documentation

### Selenium

Selenium is an umbrella project for a range of tools and libraries that enable and support the automation of web browsers. [4]

It provides extensions to emulate user interaction with browsers, a distribution server for scaling browser allocation, and the infrastructure for implementations of the W3C WebDriver specification that lets you write interchangeable code for all major web browsers.

#### Advantages:

* Fully customizable

#### Limitation:

* Doesn’t provide any metadata

In order to tackle with limitations mentioned above our system uses a third-party tool named Twint. Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API [5]. Twint utilizes Twitter's search operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers.

## Objectives of the System

* Perform real time extraction of contextualized and country specific user generated contents from variant social media channels.
* Realtime preprocessing i.e., noise removal, tokenization, Stopwords removal, lemmatization
* Syntactical analysis to check if the sentence contains any feature to declare it as subjective.
* Subjectivity classification to classify extracted data as opinion bearing or factual
* Data visualizations

## Significance of the System

Our system is capable of performing real time extraction and normalization of contextualized and country specific user generated contents from multiple social media channels i.e., Twitter, YouTube and micro blogging websites. In NLP pipeline normalization is of prime of importance because it is done to reduce the randomness, bringing the text closer to a predefined “standard”. This helps us to reduce the amount of different information that the computer has to deal with, and therefore improves efficiency [6]. The goal of normalization techniques like stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. This system performs normalization in following steps,

* Text Cleaning
  + Noise reduction
  + Removal of URLs and undesired and symbols
  + Handling redundant contents
* Tokenization
  + Sentence level tokenization
  + Word level tokenization
* Stopwords removal
* Lemmatization
  + Conversion to root form

One of the biggest concerns in NLP projects is to have data of right nature because factual data can lower the accuracy significantly. In order to solve this problem our system performs subjectivity classification and annotates the extracted contents as subjective or objective.

# Overall Description

## Product Perspective

The dawn of the internet opened the doors to the easy and widespread sharing of information on subject matters such as products, services, events, and political opinions. While the volume of studies conducted on sentiment analysis, decision making and confidence indexing is rapidly expanding, these studies mostly address English language concerns. People use short and informal language during communication. They usually adopt acronyms, slangs, and short terms for conveying their opinion and suggestions. Online merchant sites provide space for customers to write feedback about their product and services, as a result number of customer reviews grow rapidly for each product. Such reviews are useful for customers in making purchase decision regarding a product based on the experiences of the existing users.

The aim of the study is to develop an automated system that extracts the social media contents from variant blogging websites and to provide an optimized mechanism of normalization. Further aim is to classify extracted contents as subjective or objective. Following tasks are performed.

* Crawling of contents from different domain channels i.e., Twitter, YouTube and Blogging websites
* Text Cleaning
  + Noise reduction
  + Removal of URLs and undesired and symbols
  + Handling redundant contents
* Tokenization
  + Sentence level tokenization
  + Word level tokenization
* Stopwords removal
* Lemmatization
  + Conversion to root form
* Annotation
  + Syntactical annotation
  + Subjectivity classification

## Product Scope

With the exponential growth of World Wide Web and rapid expansion of e-commerce, web opinion sources such as merchant sites, forums, discussion groups and blogs are used as a platform by individual users to share experiences or opinions. Online merchant sites provide space for customers to write feedback about their product and services, as a result number of customer reviews grow rapidly for each product. Such reviews are useful for customers in making purchase decision regarding a product based on the experiences of the existing users, whereas on the other hand, it helps product manufacturers in assessing strength and weaknesses of their products from the perspective of end users. Such information is very useful in developing marketing and product development plans.

Extracting right data from such huge archive can be a problem and require substantial amount of man power. To make things go faster and rely on minimal amount of man power we have designed a unified tool which extracts relevant information from a large volume of data coming from multiple source channels in variant forms.

Our system extracts data from Twitter, YouTube and Blogging websites. Some of these resources do provide their official APIs. These APIs return data in an unstructured and raw format. Our tool receives data from APIs and performs all necessary processing. For the resources that don’t provide their API, we have developed custom bots to scrape necessary data autonomously.

## Product Functionality

1. Realtime extraction of user generated contents from various social media channels,
   1. Twitter
   2. YouTube
   3. Blogging websites
2. Data preprocessing
3. Noise removal
4. Tokenization
5. Stopwords removal
6. Lemmatization
7. Data visualization
8. Bar graph
9. Pie graph
10. Line graph
11. Subjectivity classification
12. Saving results in the form a .csv file

## Operating Environment

This is system is designed to perform real time dataset preparation for big organizations and individual users. By having this application organizations and individuals can generate a dataset on any trendy topic by themselves.

As it’s a web base application so it can be accessed by any operating system i.e., Windows, Macintosh, Android, iOS etc.

Following requirements needs to met in order to run this application,

* Good internet connection
* Latest internet browser

# Problem Statement

The internet has brought all things to our fingertips, from buying groceries to researching our next new automobile purchase. The internet is now evolving to be a forum where consumers evaluate products and services based on impressions and feedback from other, like-minded consumers.

Statistics show that the majority of all online shoppers use reviews to determine what products to buy and which services to purchase. Reviews not only have the power to influence consumer decisions but can strengthen a company’s credibility. Reviews have the power to gain customer trust, and they encourage people to interact with the company. Studies show that a customer is more likely to leave a review after a negative experience than a positive one. Negative reviews can have a dramatic, and sometimes detrimental effect on a business.

The above-mentioned points make it clear that in this modern age it is crucial to analyze user feedback to run a successful business. Companies are nowadays potentially capable of collecting data about their product lifecycle activity and performance. In order to analyze the performance of any product data needs to be collected from different channels but the problem is every channel has its own unique nature and the most desired results are obtained by the best tool selection and employment of most suitable techniques, algorithm or methodologies. The complexity of application and need of expertise for successfully implementing the right tool to right data domain has deprived the user or an organization from the benefits of these tools.

We present a system that is concerned with removing the above anomaly which otherwise results to discovery of incorrect or ambiguous data pattern and irrelevant information retrieval. The aim of social media linguistic processor is to combine the steps of real time text extraction from variant social media channels, cleaning, pre-processing, normalization and annotation of extracted contents in terms of subjective labels.

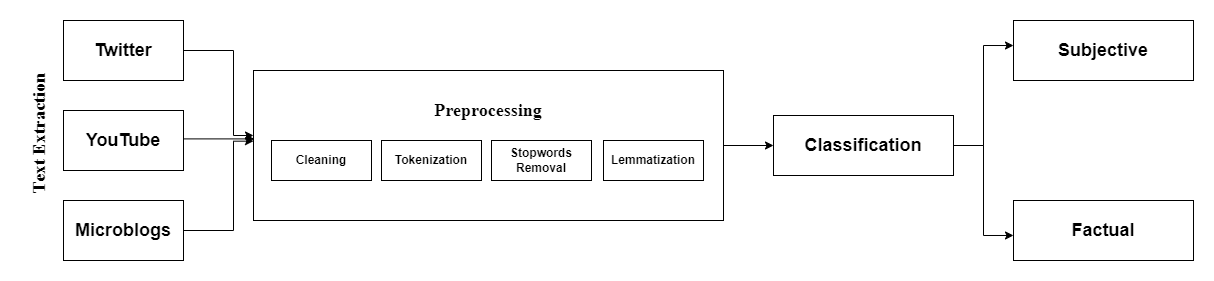
## Possible solution

Language is considered as one of the most significant achievements of humans that has accelerated the progress of humanity. So, it is not a surprise that there is plenty of work being done to integrate language into the field of artificial intelligence in the form of Natural Language Processing (NLP).

Different tasks will be performed to solve the problem mentioned above.

* Crawling of contents from different domain channels i.e., Twitter, YouTube and Blogging websites
* Text Cleaning
  + Noise reduction
  + Removal of URLs and undesired and symbols
  + Handling redundant contents
* Tokenization
  + Sentence level tokenization
  + Word level tokenization
* Stopwords removal
* Lemmatization
  + Conversion to root form
* Annotation
  + Syntactical annotation
  + Subjectivity classification

After text extraction the most crucial step is to perform cleaning as it improves overall efficiency significantly. Tokenization is used to tokenize the paragraphs into sentences and sentences into words. Lemmatization is a process to remove inflections and map a word to its root form. Stop words removal is be performed to remove the words that don’t contribute to the meaning of sentence. Syntactical annotation is performed to check if the sentence has any feature that can declare it as subjective i.e., adjectives. Sentence level subjectivity classification is done to classify sentences to be subjective or objective.



##### Fig 3.1: Working Mechanism

# Methodology

Software development methodology refers to structured processes involved when working on a project. Following essential steps are performed to achieve the target objectives,

1. Text Extraction
2. Tokenization
3. Stopwords Removal
4. Lemmatization
5. Syntactical Analysis
6. Subjectivity classification

## Text extraction:

Text data contains a lot of information but not all of it will be important to you. The task of extracting meaningful information from unstructured text data and presenting in a structured format is known as text extraction. Our system performs real time text extraction from different resource channels like Twitter, YouTube and blogging websites.

### Twitter

Twitter holds prime importance for getting insights about current affairs. The easiest way to extract data from twitter is to use twitter’s official API. But twitter’s API does come with number of limitations so in order to overcome the limitations of Tweepy (Twitter’s official API) we will make use of a third-party python-based tool for extracting data from twitter. Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. Twint utilizes Twitter's search operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers.

### YouTube

YouTube is certainly one of the most important social media platforms. About 122 million users consume a billion hours of video every day and 500 hours of content are uploaded every minute. These factors made YouTube a tremendous valuable source of data. They provide official API to easily extract information about virtually everything stored in their servers, so these petabytes of information can be leveraged to get useful insights of a wide variety of interesting problems. As YouTube API V3 doesn’t have too many limitations that’s why we have used it in our project.

### Blogging Websites

Majority of the blogging websites don’t have any official or third-party module for extracting relevant data. To solve this problem, we have developed our own custom bots and they can extract all relevant data autonomously.

## Tokenization

Tokenization is the task to divide up character streams into tokens which can be used for further processing and understanding. Tokens can be words, numbers, identifiers or punctuation (depending on the use case)

## Stopwords removal

Stop words are available in abundance in any human language. By removing these words, we remove the low-level information from our text in order to give more focus to the important information. In other words, we can say that the removal of such words does not show any negative consequences on the model we train for our task.

Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training.

## Lemmatization

Lemmatization is the process of finding the normalized form of a word [7]. It is one of the most common text pre-processing techniques used in Natural Language Processing (NLP). In lemmatization, we try to reduce a given word to its root word. The root word is called a stem in the stemming process, and it is called a lemma in the lemmatization process.

Stemming and Lemmatization both are for preprocessing in NLP but the lemmatization proves to more accurate. For NLP applications like subjectivity classification where understanding the meaning of dialogue is crucial lemmatization proves to be more accurate.

## Syntactical Analysis

Syntactical analysis is done to check if the sentence has any feature which can classify it as subjective i.e., presence of adjectives. Another reason to perform syntactical analysis is to draw exact meaning, or you can say dictionary meaning from the text. It is defined as analysis that tells us the logical meaning of certain given sentences or parts of those sentences.

## Subjectivity Classification

Classifying the sentence to be subjective or objective is called sentence level subjectivity classification.

## Tools and technologies

Following tools will used in our proposed system

* ReactJS
* NLTK
  + **Word\_tokenize**
  + **Sent\_tokenize**
  + Stopwords removal
* Django

## ReactJS

**React** is a JavaScript library for building user interfaces [8]. For frontend we are using ReactJS to achieve modern interactive and user interface. ReactJS is an open-source JavaScript library used for frontend development, which was developed by Facebook.

The reasons to select react over Angular, Vanilla JavaScript or simple HTML, CSS are given below,

### Single page applications

A single-page application is an app that works inside a browser and does not require page reloading during use

### Data Binding

React uses one-way data binding and an application architecture called Flux controls the flow of data to components through one control point, the dispatcher. It’s easier to debug self-contained components of large ReactJS apps.

### Strong community support

The large ecosystem of React.js is one of the reasons to choose React.js for this. Since it was launched by Facebook, React.js is still supported by the company.

## NLTK

NLTK, the Natural Language Toolkit, is a suite of open-source program modules, tutorials and problem sets, providing ready-to-use computational linguistics courseware [9]. 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.

## Django

## Django is a high-level Python web framework that enables rapid development of secure and maintainable websites [10].

The reasons for using Django as a backend framework for this application are as follows,

### Pluggable

### Django is pluggable by nature and can be extended with plugins. Plugins are software components that allow developers to add a specific feature to an app, leaving a lot of scope for customization.

### ORM

Django is valued for its object-relational mapper that helps developers interact with databases. An object-relational mapper (ORM) is a library that automatically transfers data stored in databases such as PostgreSQL and MySQL into objects commonly used in application code.

## Algorithm Viewpoint

1. **Input**: Text
2. **Output**: Text
3. **Function Result (.CSV file)**
4. Enter text for performing real time extraction
5. Select channel to perform extraction
   1. Twitter
   2. YouTube
   3. Blogging websites
6. Text normalization applies on extracted text
   1. Noise removal
   2. Tokenization
   3. Stopwords removal
   4. Lemmatization
7. Perform subjectivity classification to classify extracted contents as
   1. Subjective (opinion bearing)
   2. Objective (factual)
8. Display: Pre-processed and classified text
9. Save output to a .CSV file
10. **End function.**

# Implementation

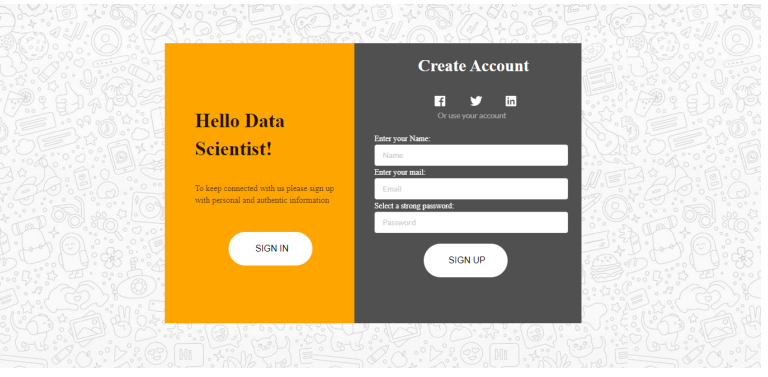
Following tools and technologies will be used in this project

|  |  |  |  |
| --- | --- | --- | --- |
| **Tools**  **And**  **Technologies** | **Tools** | **Version** | **Rationale** |
| Spyder | 4.5 | IDE |
| MS SQL lite | 2018 | DBMS |
| Adobe Photoshop | CC 2020 | Design Work |
| Adobe XD | CC 2019 | Mockups Creation |
| MS Power Point | 2019 | Presentation |
| MS Word | 2019 | Documentation |
| **Technology** | **Version** | **Rationale** |
| Python | 3.9.7 | Programming language |
| NLTK | 3.6 | Pre processing |
| SQL | 2018 | Query Language |
| React.js | 17.0 | Frontend development |
|  | Django | 3.2.5 | Backend development |

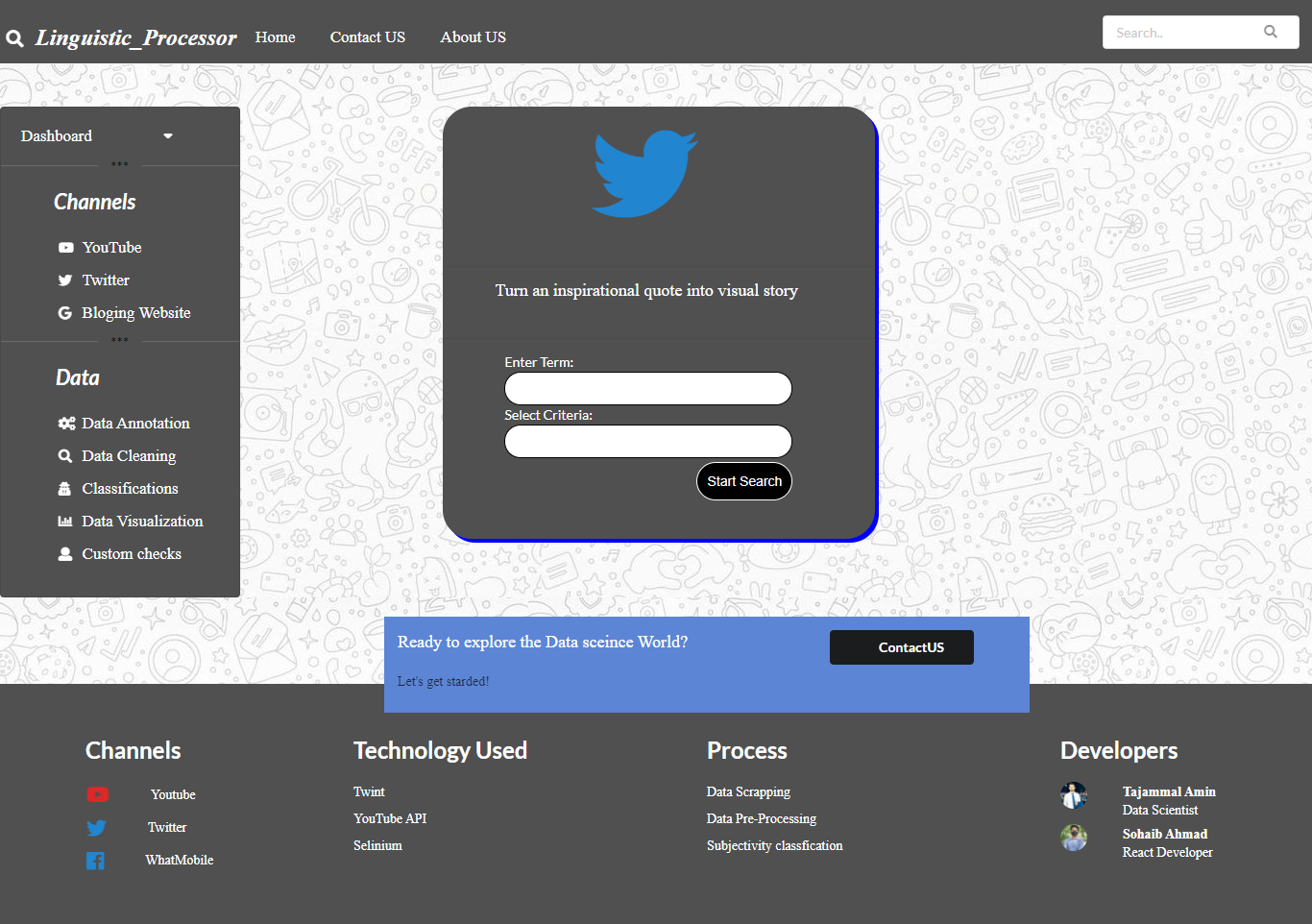
##### Table 5.1: Tools and technologies

## Software interfaces

##### Fig 5.1: Sing in page

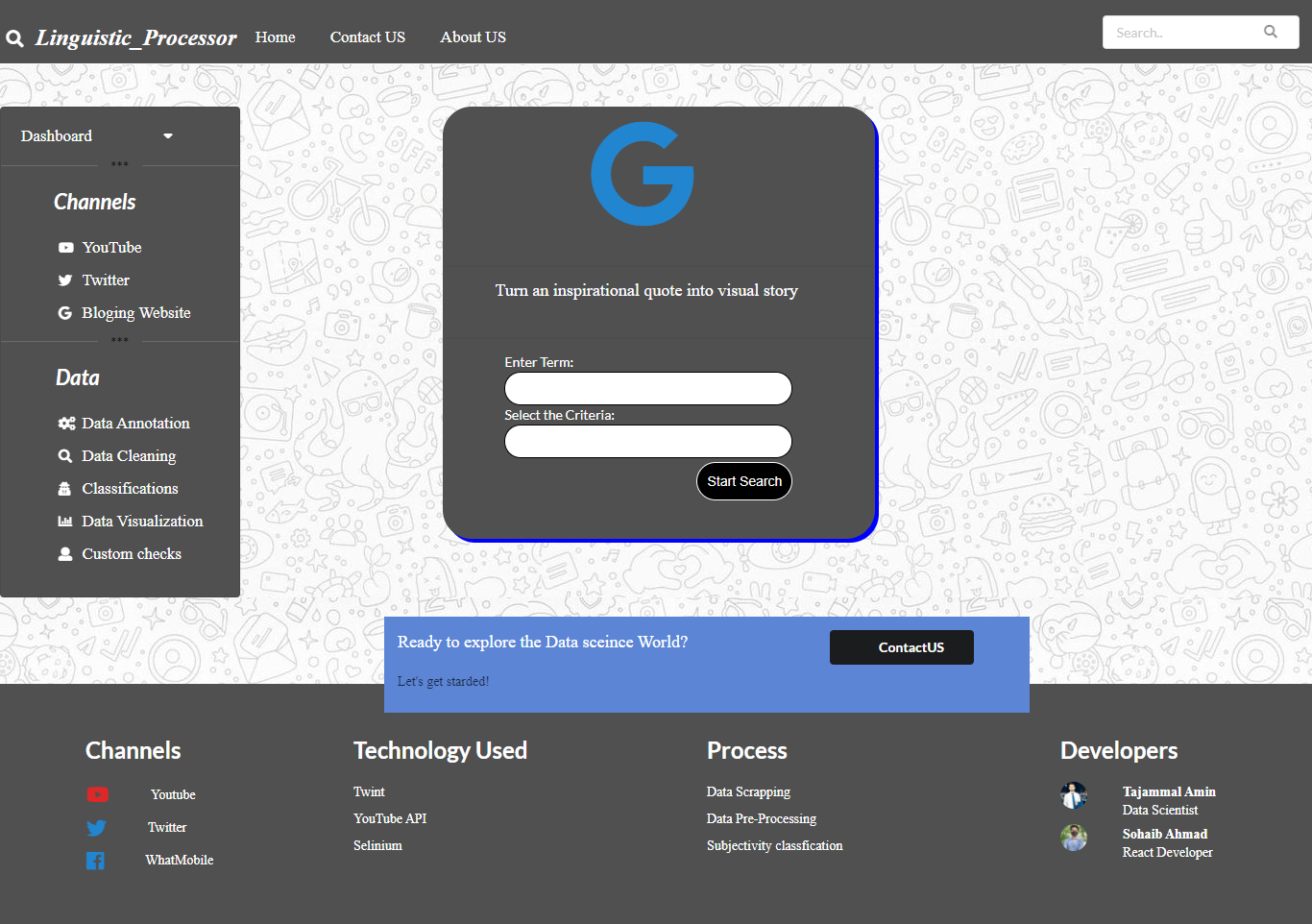


##### Fig 5.2: Sing up page

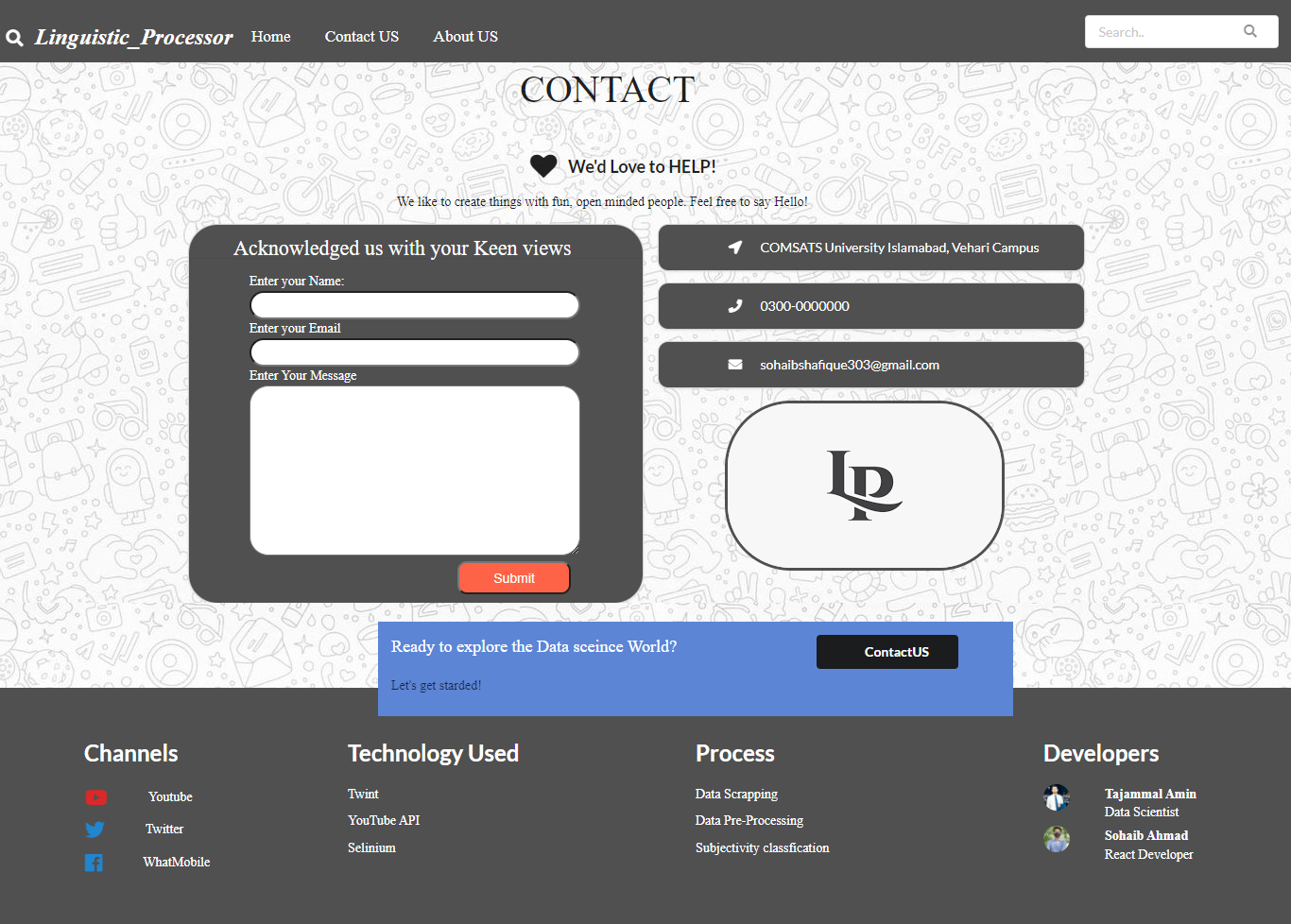
****

##### Fig 5.3: Home page/Twitter

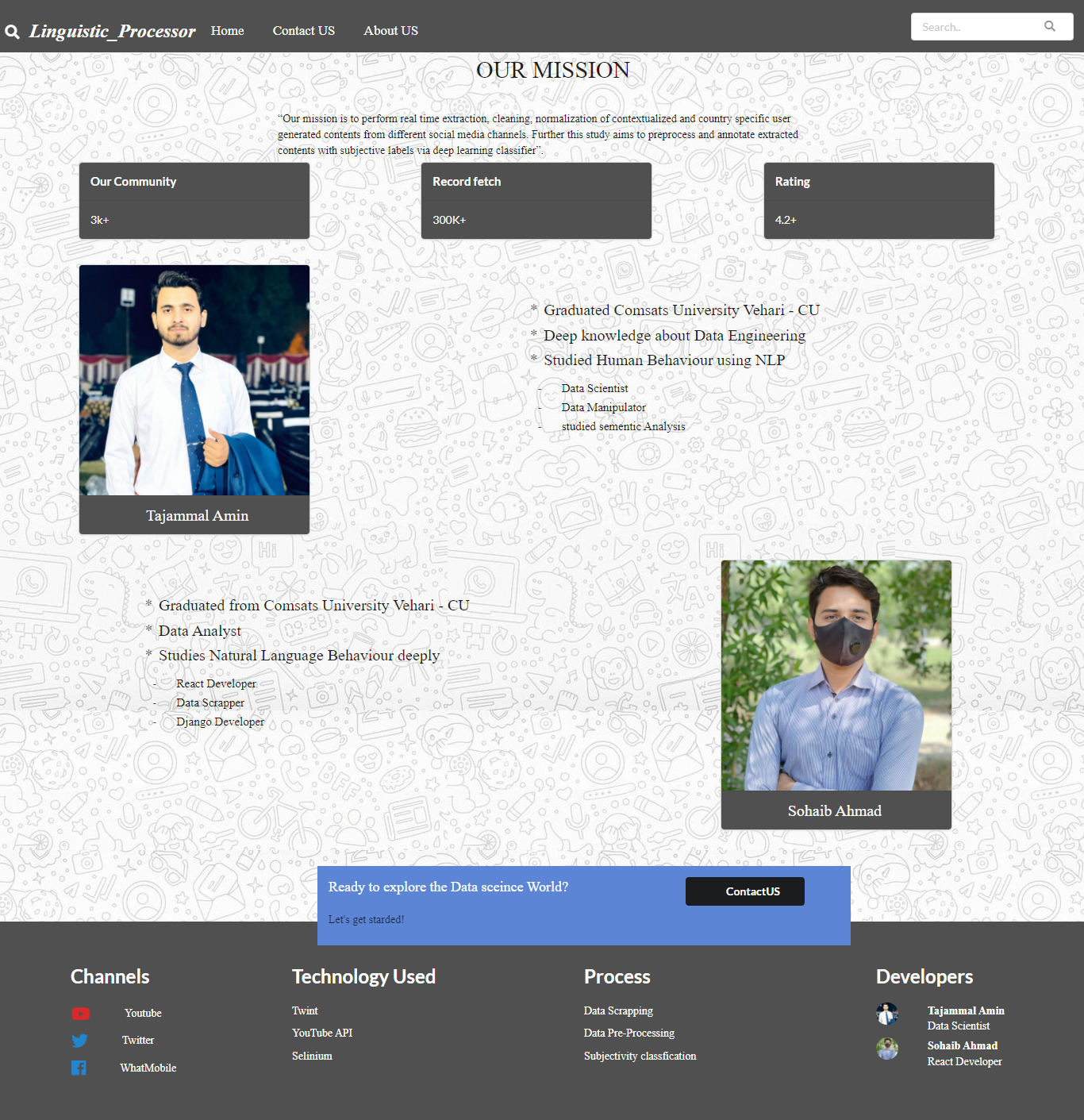
##### Fig 5.4: Home page/YouTube

****

##### Fig 5.5: Home page/Blogging websites

****

**Fig 5.6: Contact us**



**Fig 5.7: About us**

# Conclusion

The exponential growth of World Wide Web and rapid expansion of e-commerce, web opinion sources such as merchant sites, forums, discussion groups and blogs are used as a platform by individual users to share experiences or opinions. A huge amount, of text exists on these sites. Text mining has become one of the trendy fields that have been incorporated in several research fields’ Computational linguistics, Information Retrieval (IR) and data mining. Natural Language Processing (NLP) were used to extract knowledge from the textual text to convert unstructured data into structured data. Text mining reads an unstructured form of data to provide meaningful information patterns in a shortest time. Social networking sites are a great source of communication as most of the people in today’s world use these sites in their daily lives to keep connected to each other. This application performs real time text extraction and its normalization which includes noise removal, lemmatization and tokenization and prepares a dataset of the entered keyword. So, we developed an automated system that extracts the social media contents from variant blogging websites and provides an optimized mechanism of normalization. Further this system will also perform subjectivity classification to differentiate opinion bearing and factual data.

# References

|  |  |
| --- | --- |
| [1] | A.-H. Tan, "Text Mining: The state of the art and the challenges". |
| [2] | M. M. E. Wakil, "Introducing Text Mining". |
| [3] | D. F.Barrero and M. D. R-Moreno, "A Framework for Massive Twitter Data Extraction and Analysis". |
| [4] | "Slenium," [Online]. Available: https://www.selenium.dev/documentation/. [Accessed 29 December 2021]. |
| [5] | "GitHub," [Online]. Available: https://github.com/twintproject/twint. [Accessed 29 December 2021]. |
| [6] | "Towards Data Science," [Online]. Available: https://towardsdatascience.com/text-normalization-7ecc8e084e31. [Accessed 29 December 2021]. |
| [7] | J. Plisson, N. Lavrac and D. Mladenic, "A Rule based Approach to". |
| [8] | "React," [Online]. Available: https://reactjs.org/. [Accessed 29 December 2021]. |
| [9] | E. Loper and S. Bird, "NLTK: The Natural Language Toolkit". |
| [10] | "Django," [Online]. Available: https://www.djangoproject.com/. [Accessed 29 December 2021]. |