

COMSATS Institute of Information Technology

Park Road, Chak Shahzad, Islamabad

#### Department of Computer Science

**TextSol**

Submitted By:

Maham Aftab

FA15-BSE-019

Aimen Baseer

FA15-BSE-049

Supervised By:

Mr. Muhammad Rashid Mukhtar

**A-** Desktop Application/Information System **** **B-** Web Application/Web Application based Information System, ** C-** Problem Solving and AI ** D-**Simulation and Modeling ** E-** Smartphone Application ** F-** Smartphone Game

** G-** Other (specify category) \_\_\_\_\_\_\_\_ Submitted By:

Student 1 Name

Student 1 Registration Number

Student 2 Name

Student 2 Registration Number

Supervised By:

Supervisor name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 

c- Problem Solving and AI.

* 1. **Problem Statement:**

Today, we are living in an era of information age. The ever-increasing scale of textual information on web resources has presented a big range of information retrieval problems and area of research in natural language processing.

This information is scattered all around the internet, and this has a great potential value. However, it is also unstructured and written in a natural language, which poses great problems for data mining and data analysis. The scale, non-uniformity and complexity of this much information makes them a big data problem. We need to identify the context of the information, make it precise and extract only relevant information, which will require a lot of time and effort if done manually. We often need to identify whether the available documents/reports are similar. We have to go through multiple documents to know the quality of their content and it has become hard to find quality information. Semantic relatedness, public sentiment and classification of documents also falls under the same scope. Hence, bringing the attention to solve these issues.

**1.2 Problem Solution:**

Our agenda here is to accommodate maximum of all these problems in our final year project. The ultimate purpose is to save time and increase interactivity between humans and computer. We will summarize lengthy documents, and then we will check the similarity between different documents, extract the useful and important topics out of a lengthy document, which will help us understanding the core theme of document in a second, handle emotions by performing sentiment analysis. Further we will also penalize documents and texts to get the important and required information in front and organize documents into meaningful groups according to their content so that they can be browsed easily, then we will also try to identify the quality of information in a given document. In addition, all these issues will be resolved using the concepts of **Natural Language Processing**.

**1.3 Related System Analysis:**

|  |  |  |
| --- | --- | --- |
| **Application Name** | **Weakness** | **Proposed Project Solution** |
| Squash by software 995 | This application summarizes the text but the results given by it are not accurate. | Our proposed system will generate more accurate and more precise summaries. |
| Sentiment Analyzer | This online app performs sentimental analysis on text. It checks positive/ negative attitudes and tone of text, somehow, but results are not satisfactory. | We will add more quality to sentimental analysis. We will try to move a step forward from just checking the attitude and tone of text and will generate results that are more accurate. |

**Topic generation** and checking **text similarity** is a new idea that we will implement in our system.

* 1. **Advantages/Benefits Of Proposed System:**

Following are the advantages of the proposed system to the problem:

* Most significantly, it will save users time and effort, which is the basic requirement of this era.
* It will help refine raw data and differentiate the useful and useless information making it more understandable.
* It will improve productivity making it easier to find quality information.
* It will let user know what the piece of text or document is about without going through it thoroughly.
* When researching documents, the proposed system will make the selection process easier, as it will make it quite easy to figure out whether the document is relevant or not.
* It will help us classify documents with respect to their content, thus improving the browsing/querying process.
* Our system will have more accurate results than prior done research and development
  1. **Scope And Modules :**

**Scope:**

Our proposed system will perform multiple things, which includes summarization of text, identifying similarity between documents, topic modelling and sentiment analysis. In order to achieve these, we will touch different technologies with machine learning and deep learning models and produce state-of-the-art results. We will also pursue it after our final year.

**Modules**: [2]

So far, researchers have used multiple models to settle **NATURAL LANGUAGE PROCESS** issues. In our project, we will use the most up-to-dated models. Moreover, the modules involved in our project are:

1. **Preprocessing:**[6]

Preprocessing is the first step in any **NATURAL LANGUAGE PROCESS** related problem, and it includes following multiple steps are as follow:

1. **Data collection:**

The primary thing is that you need datasets to practice on when getting started with deep learning/machine learning for natural language processing tasks.

1. **Tokenization:**

Tokenization is a step, which splits longer strings of text into smaller pieces, or tokens. Larger chunks of text tokenized into sentences; sentences will tokenized into words.

1. **Normalization:**

Text normalization is the process of transforming [text](https://en.wikipedia.org/wiki/Writing) into a single [canonical form](https://en.wikipedia.org/wiki/Canonical_form) that it might not have had before.

1. **Lemmatization:**

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, known as the lemma.

1. **Stemming:**

Stemming is the process of reducing inflected (or sometimes derived) words to their [word stem](https://en.wikipedia.org/wiki/Word_stem), base or [root](https://en.wikipedia.org/wiki/Root_(linguistics)) form—generally a written word form.

1. **Entity Extraction:**

Identifying and extracting entities (people, places, companies, etc.) is a necessary step to simplify downstream processing.

1. **Word Embedding:[**1]

Word embedding is the modern way of representing words as vectors. The aim of word embedding is to redefine the high dimensional word features into low dimensional features. In addition, models used for this purpose are Word2Vec, GloVe and many others. We will select model according to our need.

1. **Model Training (Machine Learning Or Deep Learning):**

We will use machine learning and deep learning according to our needs. Depending upon the problem, we will decide that which models are appropriate to use. For example if we are having the problem of document classification we will use machine learning models which can be SVM or other model depending on our need , but if we are required to use neural networks then we will use deep learning models and so on.

1. **Training Of Model:[**4]

Models will be trained using training data that tunes the model. Training data contains the correct answer. The learning algorithm finds patterns in the training data that maps input data to the output that is required. It creates a model that captures these patterns.

1. **Testing And Evaluation:**[4]

Once we train our models, they will be evaluated using Test data that is not the part of training.

1. **Validation:**[4]

Validation includes eliminating biasness, and it is highly important for producing state-of-the-art results.

1. **Visualization And Data Representation:**[3]

We will be using Jupyter notebooks for visualization and data representation. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

* 1. **System Limitations / Constraints :**
* System’s memory and storage space must be enough to process Big data
* Data must be in huge amount; only then, our model will be able to learn different trends of data.
  1. **Tool And Technology :**

**Tools:**

Anaconda3, Jupyter Notebook.

**Technology:**

Scripting languages **–** Python & Different Machine/Deep learning libraries for computations and learning

* 1. **Roles And Project Stakeholders :**

|  |  |
| --- | --- |
| **Project Sponsor** | We are contributing to the research in NLP using machine learning and deep learning concepts to ease information overload and will produce state of the art results. |
| **Stakeholder** | **Sir Rashid Mukhtar**  Sir Rashid Mukhtar will be verifying the requirement and documents of the proposed system. |

**Project Stakeholders and Their Roles:**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student Registration No** | **Roles/Modules** |
| **Aimen Baseer** | FA15-BSE-049 | Module 1 – Module 3 |
| **Maham Aftab** | FA15-BSE-019 | Module 4 – Module 7 |

* 1. **Data Gathering Approach :**

Either we will scrap data from the website or we may use API for extracting the data. If we will be able to get data directly from the websites we will get it from there, otherwise we will generate data by ourselves. In addition, the formats of data that we will use will be excel, json or csv.

**1.8 References:**

1. [Word Embeddings](https://towardsdatascience.com/word-embeddings-exploration-explanation-and-exploitation-with-code-in-python-5dac99d5d795).
2. [Neural Network Approaches](http://www.fast.ai/).
3. [Data Visualization Libraries](https://plot.ly/).
4. [Testing, Training and Validation of Data Sets](https://l.facebook.com/l.php?u=https%3A%2F%2Ftowardsdatascience.com%2Ftrain-validation-and-test-sets-72cb40cba9e7&h=AT1x3ka5lZ6kL7sK44mJFlP2rwsrHxf7Gxk3HquudJvQi-EYx6kjnU0eSh5x4pcvCkCRzUGwLqs-O9zFVEveFAIpqxS9n5na4YiKYA6QkH1BldMcO4r91m9TG96vPMIAu-JRzg).
5. [Our Working Environment](https://www.anaconda.com/).
6. [Preprocessing in NLP](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.kdnuggets.com%2F2018%2F03%2Ftext-data-preprocessing-walkthrough-python.html&h=AT1x3ka5lZ6kL7sK44mJFlP2rwsrHxf7Gxk3HquudJvQi-EYx6kjnU0eSh5x4pcvCkCRzUGwLqs-O9zFVEveFAIpqxS9n5na4YiKYA6QkH1BldMcO4r91m9TG96vPMIAu-JRzg).

## 