IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10555

Opinion Mining and Text Summarization using Natural Language Processing

Swarajsingh Rajput[1] , Abhishek Deshmukh[2] , Ketan Patil[3] , Chaitanya Sawarkar[4]
Prof. P.M.Kamde[5]

[1-5] B.E. Student, Dept of Computer Engineering, Sinhgad College of Engineering, Vadgaon,Pune-411041, Maharashtra, India

Abstract: The conventional approaches to analyze the sentiment of text do not consider the relationship between words that show emotion and modifiers, they simply accumulate the words of the sentence to obtain the sentiment of short text. Along with sentiment analysis, the increasing amount of online information, and recourse texts, text summarization has also become an important subject with aims to preserve and show the main purpose of textual information. It is exceedingly difficult for human beings to manually summarize large documents of text. With this project, users would save a lot of time as well as conserve the resources which can be employed for better enterprises. The project would also sustain the bulk of the spam as well and give a brief of the complete information. It also helps to analyze the overall sentiment of the article and classifies it based on the data present.

Keywords: Sentiment recognition, text summarization, News classifier, Spam Check.

I. INTRODUCTION

With the incredible development of the internet, usage of social media has increased substantially where people nowadays express their views, opinions, etc. This has led to a huge increase in data that is unsaturated and unstructured. This is where Natural Language Processing comes into play. It helps the computer to read text, interpret it, determine the sentiment, and extract the important parts of the text. Today, the staggering amount of data that is generated from online sources is unstructured which makes it critical to analyze it efficiently. With the help of NLP, it is possible to analyze a huge amount of data without fatigue and inconsistent manner. Human Language is very complex and diverse. It differs from region to region. People express themselves in various languages through text or speech. Each language contains different grammar, syntax rules, terms, and slang. While many of the supervised and unsupervised learning methods, specifically deep learning is used for modeling the Human Language. But there is a need for semantic and syntactic understanding that is not possible using the standard machine learning approach. The project was designed to help users overcome the nuance difficulties of the English language. For users who are not as proficient in the language, the system can help them understand the sentiment behind the text as well as give them a classification of the text. It can also help organizations summarize text and extract valuable information without expending valuable resources and manpower. With the added functionality of spam checking, the system can be considered as an umbrella application for all NLP applications. In the future as well, as organizations invest more into chatbots, and other AI customer services. Its processing power can be increased manifold to aid them in these utilities.

II. LITERATURE SURVEY

Paper [1] demonstrates the use of dependency parsing to determine sentiment relationship migration and modified distance of microblogs. It also establishes a relationship between modifiers and emotion words and how they contribute to the sentiment calculation of text.

Authors of [2] have proposed a news text classification model that uses Latent Dirichlet Allocation(LDA). As the dimensions of the text of the news are too high, the model uses a topic model to reduce the dimensions of the text and get the important features. The text is converted from VSM to a topic vector. It also makes use of the SoftMax regression algorithm which is an extension of logistic regression for solving multi-class of text problems in life and uses it as the model's classifier. A real news Dataset was used to evaluate the proposed model and it performs better than other classification models thus making it an improved model. This model can specialize in reducing the dimensions of the text, feature extraction of the news text that help in getting better classification results.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10555

Paper [3], This paper mostly focuses on text mining techniques that help in getting different traits from textual data. Well-organized and accurate text mining is done with the help of several methods and techniques. It discusses the process of mining the text and its applications such as information retrieval, summarization, and many other applications. Nowadays there is enormous growth in digital data of which 85% is unsaturated and unstructured. The biggest issue in text mining is to determine the patterns and trends to examine the textual data. The struggle and time taken to mine important information from the text is reduced greatly by selecting the proper technique. Text data mining is used to obtain different and fascinating designs from the unstructured data that is obtained from various sources. The parts of the text such as words, sentences, phrases are converted into mathematical values and linked with the saturated information in the database which is then analyzed using traditional data mining techniques. Information can be obtained from different platforms, but unstructured data is a source of maximum knowledge.

Authors [4] constructed a domain sentiment dictionary using text data. There are various classification algorithms such as Support Vector Machine(SVM), Gradient Boosting Decision Tree(GBDT), etc. Both above algorithms mentioned are commonly used for text classification. However, they both are single classification algorithms and have weaknesses, for example, Support Vector Machine does not perform well when the data is not sufficient, or sparse and Gradient Boosting Decision Tree tends to overfit on a specific dataset. This paper mostly focuses on overcoming the weaknesses of the above algorithms by combining them i.e., making an ensemble model using SVM and GBDT. This model helped them in reaching higher accuracy to confirm they performed various experiments to show that the hybrid ensemble model performs better than SVM and GBDT.

Authors [5] implemented extractive text summarization by ranking the sentences. The input text that is given for summarization is tokenized i.e the sentences are split and assigned different tokens. After tokenization removal of stop words and punctuation takes place and the remaining words are considered as keywords that later help in summarizing the text appropriately. Each keyword has attached a part of the tag after being taken as input. This was all part of preprocessing of the input text. After this step, the frequency of each word is calculated i.e how many times the word has appeared in the text and then the maximum frequency of a word is used for calculating the weighted frequency. Weighted frequency is obtained by dividing the frequency of words by maximum frequency. This helps in getting individual term ranks. Finally, the higher weighted frequency sentences are selected and used in the summary of the given input text.

The authors [6] have conducted comprehensive research in summarization which includes traditional algorithms used as well as more recent methods for determining keywords of a specific domain or genre summarization as well as the evaluation of summarization. They have also discussed the challenges particularly in the need for language generation and semantic understanding of the language which is necessary for advancements in the field.

Paper [7] reviews various Natural Language Processing (NLP) and useful toolkits dealing with tokenization. It also reviews opinion mining and sentiment analysis. It was analyzed that there are a huge number of NLP techniques used for opinion mining and sentiment analysis during the review. The most important aspect of opinion mining and sentiments is to extract the sentiments from the given text writings. The process can be divided into three dimensions: sentence, document, and fine-grained level. In level 1 it chooses the type of presentation sentence imparts, also subjectivity arrangement that separates the important and unimportant parts of a given text in the document. Level 2 takes the task of evaluating the opinion/sentiment whether it is positive, negative, or neutral.

Authors [8] discuss the relationship between text mining and text summarization. Different summarization approaches and their parameters are also discussed in the study. Extracting pre-dominant sentences, identifying the main stages of the summarizing process, and significant extracting criteria are some of the important methodologies discussed in detail in this study. The study gives two evaluation approaches: judge by human and the second approach is compared to references summary. In the first approach, a summary generated by a machine is compared to a summary generated by a human. The only problem with this approach is the differences in personal taste and opinion. And in the second approach, there is more than one document with its summary in a dataset. Hence the document and its summary can be a good evaluation approach.

III. METHODOLOGY

Launching the application takes the user to the application home screen. At the home screen, one can see the four major NLP functionalities implemented. (A clear view of which can be seen in fig [1.1].)

The home screen is designed in such a way that navigating through functionalities could be done easily. Clicking on any of these functionalities will navigate the user to the respective page of that function. The operational detail of these can be further understood by the following diagram.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10555

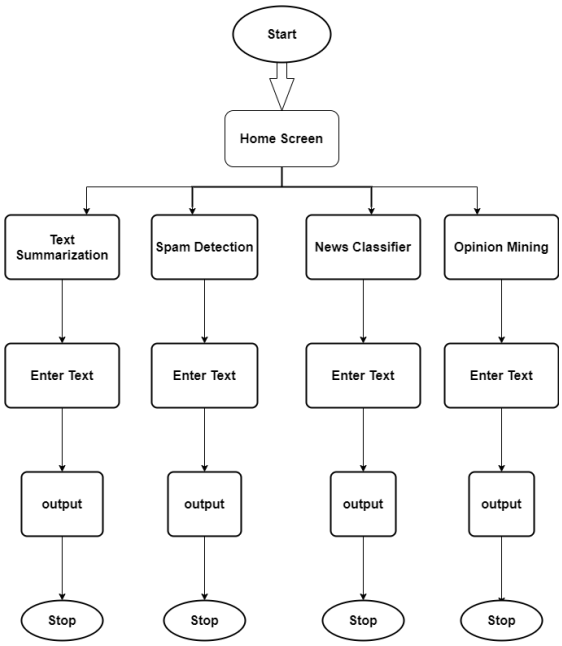

Fig. 1 System Flow of the project

The diagram depicts the stages involved in achieving the project's goals.

A. Text Summarization
The approach in summarization can be understood by the following steps:
i. Tokenization of the inputted file is done to get tokens from the terms. ii. From a predefined list of stop words the tokens are checked for stop words and are removed if found any. iii. Stop word removal is followed by lemmatization which gives us the root words. iv. From the obtained words a frequency distribution table is created to get the number of times that word appeared. v. Finally, the weighted frequency is calculated for each term. After this, the highest weighted sentences are selected and added to the summary.

B. Spam Detection
Spam detection follows the following approach for detecting :
i. After the raw text is inputted preprocessing is done. ii. Pre-processing includes a similar process as before. The first step being Tokenization where the text is tokenized into tokens. iii. Lemmatization is carried out to get the root words from different forms of the same word. iv. Here also the stop words are removed to retrieve keywords.
v. The remaining keywords are then sent to the classifier (built using logistic regression algorithm) which classifies the text into spam or not spam.

C. News classifier
The approach for news classifier can be understood by the following steps:
i. To get the inputted article classified, similar preprocessing has to be done. ii. Tokens of the inputted text are formed using a tokenizer. iii. Further retrieval of the keywords is done using lemmatization. iv. These root words are fed to the classifier which implements the Support vector classifier which finally classifies the article.

D. Opinion Mining
Opinion mining undergoes the following steps to get the overall opinion :
i. Opinion mining is carried out on the text to get the overall sentiment of the input. ii. For that, first tokenization is done to get tokens of the text. iii. Stop words are removed from the input. iv. Lemmatization is done to get the keywords that are used as final input to the classifier. Here also SVC algorithm is implemented by the classifier.
v. Finally the classifier extracts the overall opinion from the text.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10555

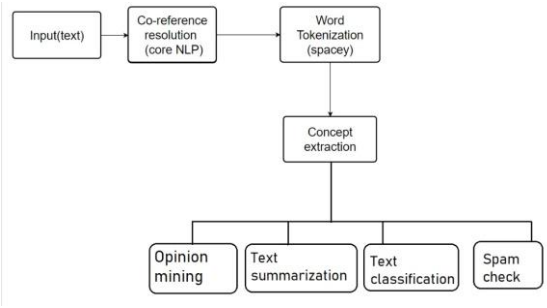

Fig. 2 System Architecture

E. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). In Spam detection, Logistic regression is best suited because the data was present in binary form that is 0 or 1 in the dataset for spam detection.

F. Support Vector Classifier

The purpose of a Linear SVC (Support Vector Classifier) is to fit the data provided, returning a "best fit" hyperplane that divides or categorizes the data. After getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. For classifying Linear SVC is preferred because of the precise hyperplanes it creates which sets perfect decision boundaries. As the distribution in our project is quite overlapped Linear SVC was best suited.

IV. PROJECT HIGHLIGHTS

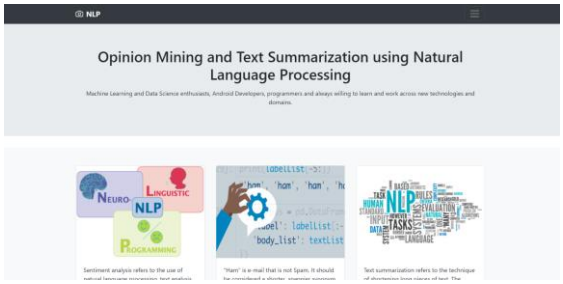

Fig. 3 Home page

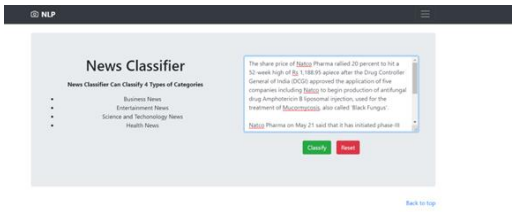

Fig. 4 News Classification page

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10555

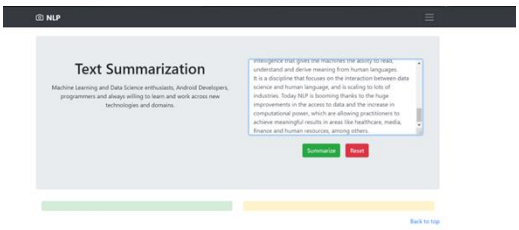

Fig. 5 Text summarization page

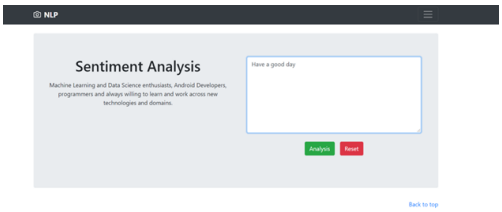

Fig. 6 Sentiment analysis page

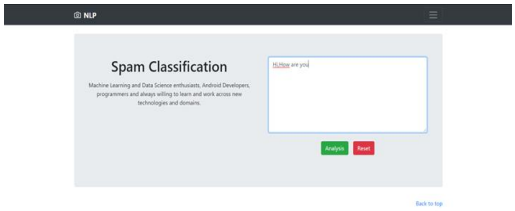

Fig. 7 Spam classification page

V. CONCLUSION

A comprehensive and integrated system has been successfully created which unifies the NLP applications under a single umbrella. This app helps users to perform text-based analysis with ease. The four functionalities namely Sentiment Analysis, Text Summarization, News Classification, and Spam checks form a one-step platform for analysis purposes.

ACKNOWLEDGMENT

The researchers would like to thank Prof. M.P. Wankhede(Head of Department), Prof. P.M. Kamde, and Prof. S.N. Bhosale for the endless support throughout the study.

REFERENCES

[1] J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017, pp. 776-779, DOI: 10.1109/CSE-EUC.2017.153.

[2] Z. Li, W. Shang, and M. Yan, "News text classification model based on topic model," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1-5, DOI: 10.1109/ICIS.2016.7550929.

[3] S. S. Tandel, A. Jamadar, and S. Dudugu, "A Survey on Text Mining Techniques," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 1022-1026, DOI: 10.1109/ICACCS.2019.8728547.

[4] Kai Yang, Yi Cai, Dongping Huang, Jingnan Li, Zikai Zhou and Xue Lei, "An effective hybrid model for opinion mining and sentiment analysis," 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017, pp. 465-466, DOI: 10.1109/BIGCOMP.2017.7881759.

[5] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, DOI: 10.1109/IconDSC.2019.8817040.

[6] Ani Nenkova; Kathleen McKeown, Automatic Summarization, now, 2011.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10555

[7] Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah, and A. Shah, "Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis," 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2018, pp. 1-4, DOI: 10.1109/ICETAS.2018.8629198.

[8] S. R. Rahimi, A. T. Mozhdehi and M. Abdolahi, "An overview on extractive text summarization," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI),2017, pp. 0054-0062, DOI: 10.1109/KBEI.2017.8324874.