iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50147

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Analysis of Different Text Features Using NLP

Deepali Joshi[1], Harsh Zanwar[2], Keyur Soni[3], Sanika Yadav[4], Priya Wankhade[5]
Vishwakarma Institute of Technology, Pune, Maharashtra, India

Abstract: ***A single letter can create a word, then that single word can create a paragraph and many paragraphs together make a perfect bound of a text collection, make a perfect article. These texts, these words can be said a way through which one expresses themselves. Be it for personal reasons, or professional, Texts do play an important role to convey the feelings. So, with these texts, many operations can be performed on it. Be it making a very large paragraph shorter - as is done in para summarizer, be it changing the language of a given text - as in language translator, be it automatically guessing of the next word - as in automatic word generator, or be it the spelling a grammar corrector. Here, NLP takes its role. Natural Language Processing i.e., NLP includes studies of how computers and humans communicates in a particular language, more precisely natural language. It is a field that blends computer science, linguistics, and machine learning. NLP aspires to enable computers to understand and generate human language. There are many more tasks one can perform with the text given, to get a more fruitful output of it. In this paper, work done on some of these operations or features of the texts are discussed and summarized.***
Keywords: texts, words, paragraphs, NLP, spell checker, text generator, paraphraser, text summarizer, text language translator, grammar checker

I. INTRODUCTION

Natural Language Processing is the study of teaching computers to comprehend natural language. It is definitely not an easy task. Computers do understand organized data like the ones in spreadsheets and database tables, but they struggle with unstructured data like human languages, words, and voices, necessitating the usage of NLP. Natural language data is present in many forms around us, and computers interpreting and analyzing that data would make it a lot easier to grasp and handle. To match projected results, we can train the models using a variety of ways and techniques. For thousands of years humans are witting down, in the form of texts and manuscripts, and there is an enormous amount of literature available, which would be incredible if computers could comprehend. The task, on the other hand, will never be easy. Understanding the precise meaning of a sentence, accurate Named-Entity Recognition i.e., NER, accurate prediction of various portions of speech, and conference resolution are just a few of the difficulties that can be addressed.

Human language is incomprehensible to computers. If we insert a model with good amount of data and properly train it accordingly, it will absolutely work well. Based on previously provided data and experiences, it will be able to recognize and attempt to categorize different parts of speech such as noun, adjective, verbs and all. It tries to identify the nearest word when it encounters a new term, which can be embarrassingly incorrect at times.

A computer's ability to get the precise meaning of a particular sentence is extremely tough. Like, for instance, the boy exuded a fiery aura. Is it possible that the youngster had a highly inspiring personality or that he truly emanated fire? Parsing English with a computer will be difficult, as you can see here.

In order to train a model, there are several steps that must be completed. In Machine Learning, solving a difficult problem necessitates the creation of a pipeline. Simply said, it comprises breaking down a huge problem into smaller problems, developing models for each, and then integrating them. Somewhat, similar goes with NLP. The process of learning a language for a model can be broken down into several simple steps -

1) *Step - 1: Sentence Segmentation:* It means breaking down the piece of given text in various sentences. Segmenting the sentence, thus named Sentence Segmentation
2) *Step - 2: Word Tokenization:* Tokenization is the process of breaking down a statement into individual words. We can tokenize them anytime we come across a space and use that information to train a model. Punctuation marks are also regarded distinct tokens because they have meaning.
3) *Step - 3: Prediction parts of speech for each token created:* Identifying the type of word, i.e., whether it is a noun, adjective, verb, adverb, or anything else. This will make it easier to comprehend what the statement is about. This is accomplished by passing the tokens (together with the surrounding words) to a pre-trained (already trained) part-of-speech categorization model.
4) *Step - 4: Lemmatization:* Inserting root word into the model

5) *Step - 5: Identifying Stop Words:* There are many words in any language, suppose in English language we have 'a', 'and', 'the' and many like these which are used frequently. When undertaking statistical analysis, these words produce a lot of noise. These words can be deleted
6) *Step - 6.1: Dependency Parsing:* This requires figuring out the relationship between the words in the phrases and how they relate to one another. Here, a parse tree is created with root being placed as the main verb in the phrase
7) *Step - 6.2: Finding Noun Phrases:* It means grouping of the words that are representing the same idea
8) *Step - 7: Named Entity Recognition (NER):* It examines how a word is used in a phrase and applies various statistical models on it to determine which type of word it is.

There are different features that can be performed on the texts, some of which are described below.
A spell checker is a software feature that examines a text for misspellings. Word processors, email clients, electronic dictionaries, and search engines all have spell-checking features built in.
In computing jargon, a grammar checker is a programme (or component of a programme) that checks written text for grammatical errors. Grammar checkers are frequently included as part of a bigger programme, such as a word processor, but they can also be used as a standalone tool from within programme that work with editable text. Natural language processing, also known as NLP is used in the implementation of a grammar checker.
In and of itself, translation entails deciphering the meaning of a document and producing a new text that is functionally equivalent to the original but written in a different language. This is the work done by Text Language Translator.
A text summarizer is a tool that wraps up a given text to a specified short length. Simply, we can say it condenses a long article to main important points. It can be either abstractive or extractive.
The task of guessing what word will be spoken next, also known as Next Word Prediction or Language Modeling, is known as Next Word Prediction, Automatic text generator. It's one of NLP's most important duties, and it's useful in a variety of situations.

II. LITERATURE REVIEW

Summarizing Short Stories [1], This paper offers a method for creating extracting summaries of literary short stories automatically. The summaries are written with one goal in mind: to assist readers in deciding whether or not they want to read the entire novel. To that purpose, the summaries provide relevant information about the story's location without giving away the storyline. The approach depends on a variety of surface indicators regarding clauses in short stories, the most essential of which are those connected to a clause's aspectual type and the story's primary characters. The summaries were judged by fifteen assessors on a variety of extrinsic and intrinsic criteria. The results of this assessment indicate that the summaries are beneficial in accomplishing the original goal.
Text Summarization: An Overview [2], this paper provides an overview of the whole concept of text summarizing. Starting from the very introduction of the text summarizing technique, it goes on to the description of methodology of it to take place. It further discusses the types of it, extractive and abstractive in detail. It says, the goal of extractive document summarising is to choose a number of representative sentences, chapters, or paragraphs from the original content mechanically. Text summarising techniques based on neural networks, graph theory, fuzzy logic, and clustering have all been successful in producing an effective summary of a document to some extent. Methods that are both extractive and abstractive have been studied. The majority of summarising approaches rely on extractive techniques. The abstractive method is akin to human summaries. Abstractive summarization currently necessitates the use of complex language generating equipment and is difficult to reproduce in domain-specific domains.
Effectiveness of Grammarly [3], this paper is based on the review of grammarly application. Grammarly is a programme that detects and corrects problems in abstract English language. The goal of the research was to see how successful Grammarly was in writing abstract English. A pre-experiment method was used in this investigation. All seventh-semester students who were randomly selected as research samples made up the research population. The data was gathered using abstract text prepared before using Grammarly (pre-test) and abstract text edited with Grammarly (post-test) (post-test). Analysis The pre-test score was 64.86 with an 18.89 standard deviation, and the post-test value was 85.72 with a 6.82 standard deviation. The hypothesis test utilising a Paired t-test yielded a p-value of 0.00 (0.05), indicating that the Grammarly application was effective in creating English abstracts text by HangTuah Tanjungpinang students. Grammarly showed problems in the use of punctuation 10.5 percent of the time, spelling 35.5 percent of the time, word choice 13.4 percent of the time, and sentence structure 33.0 percent of the time, according to an analysis of the correction results in abstract writing. Finally, this study found that Grammarly was a useful aid for improving students' writing of abstract English texts involving the right use of English grammar.

Spelling Checker Algorithm Methods for Many Languages [4], This paper examines a number of studies that were undertaken in languages except English. Indonesia, India, Africa, China, Arabia and Thailand are just a few examples. The spelling check approach attempts to validate and fix misspelt words by using a series of suggested words that are closer to the incorrect term. To papers published between 2008 and 2018, they utilised a comprehensive literature review approach. There are 23 papers in total, one for each language and approach. The findings show that each language uses a different approach to spell checking. The Damerau-Levenshtein algorithm is most widely used in spelling checkers.

Survey of Automatic Spelling Correction [5], From 1991 to 2019, this study provides a survey of spelling correction papers indexed in Scopus and Web of Science. For this, they divided the work in three groups. The very first group follows a set of pre-determined rules. Next, second group uses a context model that is distinct from the first. In the survey, the third category of automatic spelling correction systems may adapt their model to the problem at hand. Each system's application area, string metrics, language and context model are listed in the summary tables. The overview discusses a variety of concepts within a common conceptual perspective centered upon Shannon's noisy channel. Methods of evaluation and benchmarking are described in a separate section.

Survey of Automatic Spelling Correction [6], this paper is about the translation of a text into different language. The research is intended to analyse significant approaches to literary translation quality assessment. The qualitative method was used to explore the processes and linguistic aspects connected with literary text translation. The study looked into and evaluated the subject of literary translation, emphasising the need of maintaining text originality throughout translation. A theoretical perspective was used to evaluate the examination of methodologies and linguistic approach. Various previous studies have been reviewed to discover the connected factors based on literary text qualities. According to this study, translated literary texts are known as hybrids to some extent since they can be seen as a transplantation of the source text into another language based on the cultural milieu.

Automatic Text Generation by Learning from Literary Structures [7], This paper describes the technique for automatic text generation. According to their findings, they fulfilled the overall goal by avoiding over-structured texts and developing coherent stories with original sentences based on what had previously been published, as well as by creating a narrative flow in texts that resulted in coherent stories. The suggested architecture is a methodology for converting a single word into a complete tale by combining syntactic and semantic properties of words and phrases, so bridging the gap between language syntax and semantics. They developed a literary style architecture for generating texts that are more closely related to the language used in fiction texts. Similarly, we were able to emulate creativity in the sense that we don't have influence over the content of the stories and can't predict the algorithm's outcome until it generates a new narrative arc, all based on a single word as input.

Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection [8], This paper demonstrated a unique data augmentation technique for grammatical error detection that uses neural machine translation to learn the distribution of language-learner defects and introduce them into grammatically correct text. To increase the quality of our synthetic errors, different sampling methods were tried. They improved prior state-of-the-art results on the canonical test with even a basic BiLSTM after creating artificial training examples with an off-the-shelf NMT, and established a new state of art with a stronger model. They also showed that we could use corruptions in an out-of-domain dataset to create new benchmarks for two different, likewise out-of-domain tests without explicitly optimising for either.

Error Detection in Highly Inflectional Languages [9], In this study, this paper explains why error detection is challenging for inflectional languages, especially when contrasted to languages such as english. The topics of scale, coverage, hamming distance, and classifier accuracies were all considered. The difficulties are exacerbated by the complexity of the Indian script. It investigated the issue and devised a number of mistake models and methods for detecting them. For Malayalam, F-score was 0.66, while for Telugu, it was F-score of 0.78. They applied simple dictionary-frequency based mistake correction in Malayalam and were able to reduce word error rates by 10%.

Qualitative research and translation dilemmas [10], The goal of this work is to look at qualitative research translation challenges. It focuses on three issues: whether or not the act of translation should be identified from a methodological position; the epistemological implications of who makes the translation; and the consequences for the final output of the researcher's decision to include a translator in the investigation. Some of the methods researchers have explored to handle linguistic issues are discussed.

NATURAL LANGUAGE GENERATION [11], This paper introduces newcomers to the field of computational approaches to the former—natural language generation (NLG) - by highlighting some of the theoretical and practical challenges that linguists, computer scientists, and psychologists have faced in attempting to explain how language works in machines or in our minds.

A Review on Grammar error correction in different domains [12], In this paper, numerous feature extraction and selection strategies in diverse domains are investigated. These methods have a significant impact on system performance. The main objective of the English grammar project is the correction of articles, prepositions, clause connectives, and spelling errors.

N-gram features, dependence features, POS features, and other features are all examined. Many metaheuristic algorithms, such as the genetic algorithm, PSO, firefly algorithm, binary grey wolf optimization algorithm, and others, are studied for feature selection. For error correction and performance evaluation, algorithms such as kNN, Nave Bayes, SVM, and others are utilised. Accuracy, f-measure, precision, and other performance metrics are taken into account. The key issues, such as optimal error correction, excessive memory usage, and time consumption, are still present.

Grammatical Error Correction (GEC): Research Approaches till now [13], This paper explains how the researchers applied GEC approaches and whether setbacks or technological advancements opened the way for newer and more advanced GEC approaches. It also explains how the GEC work might be enhanced and what choices will be made for achieving higher performance in the future. Grammatical Error Correction (GEC) in the "Natural Language Processing" Domain is the process of detecting and repairing grammar problems in a text. "Rule-based," "Classification-based," and "Machine Translation" were the three methodologies utilised to solve the GEC job, with Machine Translation being further divided into "Statistical Machine Translation" and "Neural Machine Translation." All these approaches are considered in this paper.

Understanding the Processes of Translation and Transliteration in Qualitative Research [14], This paper aims to explain and investigate some important procedures and concepts which are involved in qualitative research transliteration and translation. The use of qualitative methodologies in health research is gaining popularity among many professionals like that of health and social care. Qualitative cross-cultural research analysis is a difficult endeavour that necessitates an understanding of many methodologies, procedures, and mastery of the proper languages.

Natural Language Generation with Computational Intelligence [15], This research looks at how a software agent learns to play Mario Bros. via explanations. The authors had two goals in mind when it came to improving learning from explanations: 1) to filter explanations into warnings and advises, and 2) to figure out the policies from sentences that don't include any state information Sentiment analysis was utilised by the creators to filter explanations into suggestions for what to do and what not to do, a list of things to stay away from. To depict the agent's activities when it comes to items, the developers created object-focused advice.

III. CONCLUSION

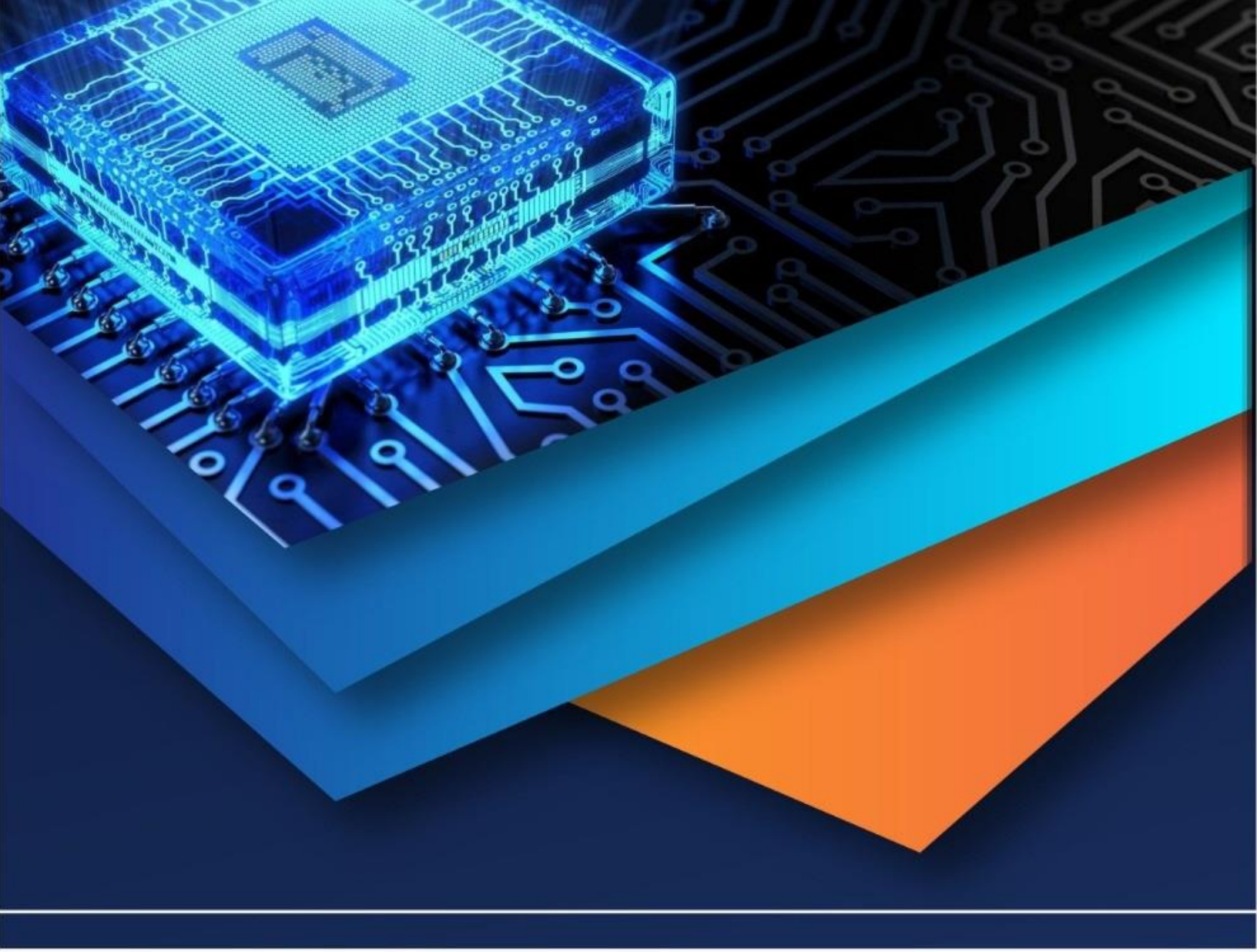
Firstly, in this paper, at the very beginning the need of NLP (Natural Language Processing) is discussed. This paper contains some of the tasks that can be performed with a given text, in context with NLP technique. It altogether groups down and summarizes few works done in this matter. Some are the surveys done on this topic, some are the Research papers published on the same. Just a short summary of their papers is jotted in this paper. It can be difficult to explain your views in an organised manner when writing an email or document. The most crucial concepts could be lost in the shuffle, and the paragraph arrangement could be confusing. As a result, a reader may miss what you're attempting to express, especially if they're skimming the content. SO, here comes word predictor, grammar and spell checker, word translators into picture. NLP works into this. You put the text through a huge language model to create a latent representation for each sentence, and then use labelled training data to educate the model which sentences are important and which aren't. The model then determines which sentences have the best chance of becoming the major points. There are various modules and libraries made for this work, like NLTK, Standard core NLP, TextBlob, GingerIt. Spacy.

The Natural Language Toolkit (NLTK) is the most well-known Python module for text analysis and natural language processing (NLP). This library focuses on research and instruction, therefore there are plenty of resources to get you started, such as data sets, pre-trained models, and a textbook. Stanford Core NLP is a well-known text analysis library that was developed by Stanford's NLP group. It's written in Java, but you may also use a Python wrapper to access it. This library offers various capabilities for grammatical analysis and can help you construct sophisticated solutions. It is robust, fast, and scalable. Many firms, in fact, use it to create sentiment analysis models or chatbots. Textblob is a python package for processing textual data that is open-source. It conducts a variety of operations on textual data, including noun phrase extraction, sentiment analysis, categorization, and translation, among other things. Ginger is a Python library and is open-sourced. Ginger is an artificial intelligence-powered writing assistant that can detect and repair spelling and grammatical errors in your content based on the context of the entire sentence. NLTK has been upgraded with SpaCy. This ultrafast Python and Python toolkit are the go-to library for complex text analysis and is focused on building real products.

REFERENCES

[1] “Summarizing Short Stories”, Anna Kazantsev, Stan Szpakowicz, University of Ottawa Polish, 2010
[2] “Text Summarization: An Overview”, Samrat Babar, 2013
[3] “Effectiveness of Grammarly Application for Writing, English Abstract”, Umu Fadhilah1, Lizawati2, Hotmaria Julia Dolok Saribu, 2018

[4] “Spelling Checker Algorithm Methods for Many Languages”, Novan Zukarnain, Bahtiar Saleh Abbas, Suparta Wayan, Agung Trisetyarso, Chul Ho Kang, 2019
[5] “Survey of Automatic Spelling Correction”, Daniel Hládek, Ján Staš, Matúš Pleva, 2020
[6] “Translating a Literary Text: Enigma or Enterprise”, Syed Sarwar Hussain, 2018
[7] “Automatic Text Generation by Learning from Literary Structures”, Angel Daza, Hiram Calvo, Jes´us Figueroa-Nazuno, 2015
[8] "Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection", Sudhanshu Kasewa, Pontus Stenetorp, Sebastian Riedel, 2018
[9] "Error Detection in Highly Inflectional Languages", Naveen Sankaran, C. V. Jawahar, 2013
[10] "Qualitative research and translation dilemmas", BOGUSIA TEMPLE, ALYS YOUNG, 2004
[11] "NATURAL LANGUAGE GENERATION", JOHN BATEMAN, MICHAEL ZOCK, 2012
[12] "A Review on Grammar error correction in different domains", Kiran R. Borade, Prof. N. M. Shahane, 2019
[13] "Grammatical Error Correction (GEC): Research Approaches till now", Sagar Ailani, Ashwini Dalvi, Irfan Siddavatam, 2019
[14] "Understanding the Processes of Translation and Transliteration in Qualitative Research", Krishna Regmi, Jennie Naidoo, Paul Pilkington, 2010
[15] "Natural Language Generation with Computational Intelligence", José M. Alonso, Alberto Bugarí, 2017
[16] https://www.geeksforgeeks.org/introduction-to-natural-language-processing/
[17] https://www.geeksforgeeks.org/natural-language-processing-overview/#:~:text=Natural%20Language%20Processing%20(NLP)%20is%20a%20field%20that%20combines%20computer,interpret%20and%20generate%20human%20language.
[18] https://monkeylearn.com/blog/best-text-analysis-apis/
[19] https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can.
[20] https://www.grammarly.com/blog/engineering/grammarly-nlp-building-future-communication/#:~:text=At%20Grammarly%2C%20we%20are%20passionate,30%20million%20daily%20active%20users.
[21] https://www.javatpoint.com/nlp
[22] https://www.techtimes.com/articles/224774/20180420/14-reasons-why-you-should-use-grammarly.htm
[23] https://www.apoven.com/grammarly-benefits/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY