2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN) | 979-8-3503-3802-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CICTN57981.2023.10140271

Text Generator using Natural Language Processing Methods

Rupinder Kaur
Computer Science and Engineering
Chandigarh University,
Punjab, India
rupinder.e12697@cumail.in

Amandeep Kaur
Computer Science and Engineering
Chandigarh University,
Punjab
amandeep.e11813@cumail.in

***Abstract*—The field of NLP (Natural Language Processing) deals with the interaction between computers and human languages, focusing on the automated analysis of human language to extract useful information.For organisations that place a premium on efficiency, automation is the ultimate aim, and natural language processing is key to getting there. A machine can only completely comprehend humans when it talks with them in human language; it's like speaking to someone in your mother tongue as opposed to someone you hardly understand or can barely speak. It is closed using natural language processing. . This paper will examine how a machine can truly comprehend human language while also delving into the specific algorithms that are employed. This essay will examine text conversions in-depth and go over how natural language processing proceeds. For those who are curious about the fundamentals of comprehending and mastering natural language processing, this review essay may be helpful. Natural language processing has a wide range of applications, including text processing, summarization, translation, voice and speech recognition, AI-powered user interfaces, expert systems, and more. However, Despite advances, NLP still faces challenges in accurately detecting emotions and expressions conveyed through non-verbal cues such as voice tone and body language.There are far too many difficulties in processing the complex linguistic system of humans. It will soon be necessary for everyone to communicate with a machine for both domestic and space-related tasks, making it crucial to comprehend how artificial intelligence works and how crucial NLP is to it.**

Index Terms*—Natural language processing (NLP) encompasses techniques such as machine learning, stemming, lemmatization, TF-IDF, artificial intelligence (AI), bag of words, and vectorization to process and understand human language.

I. INTRODUCTION

In the field of artificial intelligence known as natural language processing (NLP), machines are taught to converse with humans. The natural language is transformed into language that computers can understand using a variety of techniques[1]. Humans speak or utilise natural language to communicate with one another. Humans can communicate by writing or speaking (audio) (text). Humans pick up their language from their environment, and different geographical areas have unique languages. Learning a language involves memorising word meanings and then putting them together in sentences, but communication does not solely rely on alphabets[2]. Humans can express themselves through body language as well; for example, a person's smile may give away their happiness. Regardless of the method of communication, we use natural languages to express our thoughts and feelings as well as our responses to other people and our environment. As of right now, technology is not developed enough for computers to fully comprehend the variety of human languages. We are introduced to the collection of strategies used to attempt to accomplish that goal using natural language processing. The first thing to grasp is that computers can only interpret numbers. Text files or audio files can both contain raw natural language data.

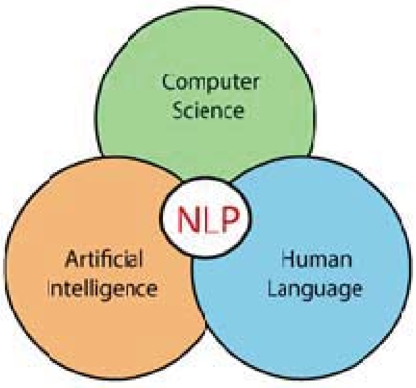
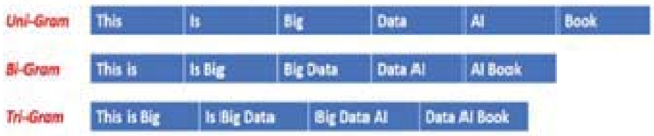


Fig. 1. Overview of Natural Language Processing

Both classified and unstructured data fall under this category. To make categorical data understandable to machines, it must be transformed into numerical data. Unstructured data cannot be transformed into rows and columns to make sense in a regular table, such as one found in a book. It consists of a continuous string of words, each of which is given a meaning based on the sentence it is used in. The majority of human language data is in unstructured form, which also needs to be converted into structured data for processing[3]. Natural language processing (NLP) uses techniques such as grammatical structure extraction to fulfill specific tasks. As a result, natural language generation is capable of producing output tailored to task requirements and language rules.

The information may be contained in a text file or an audio file. There are many uses for processing both types of data, like Google autocomplete, chatbots, voice interface bots, etc. Before being supplied to the machine, each word in the text file needs to be converted into a vector. There are several ways to do it; some of the more popular ways will be covered later. The audio file is examined by a machine that compares its wavelength and frequency to the sounds of the words already in use.



Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Fig. 2. Big data AI book

II. Bag Of Words Approach

Every word in a sentence serves a purpose in the overall meaning of the sentence. Every sentence or document is viewed as a collection of words in a bag or container in natural language processing, hence the name ”bag of words”[4]. All additional information regarding the arrangement or structure of the words in a document is ignored, and each word is considered as an independent variable. Symbols of any kind, including punctuation, are also considered independent variables.

III. Processes Of Natural Language Processing

Before the data can be supplied to the machine learning algorithm, it must undergo a number of procedures. There is a systematic discussion of these operations. These activities, as well as data automation processes, are frequently carried out using Python and R. Natural language processing might flow differently depending on the issue and the data available, but some commonplace activities are regarded as crucial.

A. Tokenization:

In order to break the text into small pieces we can use tokenization technique. A document may be broken up into various paragraphs, phrases, or even single words[5]. In order to solve our dilemma, a programmer can construct more precise tokens, such as words beginning with ”ta” or words ending in ”ed,” or he can eliminate all numeric data and solely use categorical data. In rare cases, a programmer may need to use a collection of words as a token rather than just one word and this process is known as N-gram. The N-gram is a tokenzation technique which is a group of words that are presented in a document or a sentence and are made up of characters or words. N denotes a number, like 1, 2, 3, 4, etc., that reflects the quantity of elements that make up a gram. The term ”Uni-gram” refers to a single word, ”Bi-grams” to two words combined, ”Tri-grams” to three words combined, and so forth[6]. Consider the sentence ”This is Big Data AI Book.” The image below illustrates how this language is broken up into various n-gram tokens. After tokenization, only distinct words remain, resulting in a list of the words that make up the sentence. Because it is simple to infer the meaning of a text by looking at its terms, tokenization is crucial.

B. Data Pre-Processing:

Data must first undergo preprocessing or cleaning after tokenization before being sent to the computer. Many of the tokens obtained in the previous phase might only include a few pieces of information or have no value at all. Numbers and special characters like ”@” and ”” can be used in sentences. These tokens don’t offer any useful details about the sentence’s context. For instance, if the tokens ”684” and ”” are the only ones eliminated from a list of tokens that reads [”Raman’s”, ”house”, ”no.”, ”is”, ”Raman”, ”lives,” ”alone”], the list becomes [”Raman’s”, ”house”, ”no”, ”is”, ”Raman”, ”lives,” ”alone”]. It is clear that the sentence’s meaning may still be inferred from the collection of words.The algorithm’s space and time complexity can be decreased by removing numerals and other special characters[7]. Data may also contain words in various case types (upper case and lower case). All words must be transformed to the same case, typically lower case, because a machine might not be able to discriminate between them. Additionally, the algorithm’s effectiveness will rise as a result.

C. Removing Stop Words:

Stop words are frequently used words in a language such as ”a,” ”the,” ”and,” and ”are” in English, that typically carry low informational content and make up a significant portion of a sentence.. Most people consider stop words to be a ”single collection of words.” Depending on the context, it can mean a wide range of things. For instance, in some situations, removing all commas and adjectives that give the sentence quality may be an appropriate stop word list.However, in other circumstances, doing so could result in the loss of important information. For instance, in sentimental analysis, deleting the adjectives (’good,’ ’bad,’ etc.) can cause the loss of important information regarding the sentence[8]. In this scenario, a list consisting solely of determiners, determiners combined with prepositions, or coordinating conjunctions can be utilized.

D. Finding the Root Word:

By combining fundamental words with numerous prefixes and suffixes, numerous words can be created. The root word acts as the foundation for a word, providing its core meaning. Prefixes and suffixes are added to the root word to form new words with altered meanings. The root word is what determines a word’s meaning, as demonstrated by the examples in the Table 1:[9].

The process of finding the root word for various tokens and representing them through it can result in additional increases in the time and space complexity of the model.. For instance, the token set [”Program,” ”Programmed,” and ”Programming”] can be represented by the single root word ”programme,” which can be used to represent all three terms. In natural language processing, lemmatization and stemming are the basic methods for locating a root word.

TABLE I. Table showing Prefixes and suffixes

Root word	Prefix	Suffix
Friendly	Unfriendly	Friendliness
Faith	Unfaith	Faithfully
Respect	Disrespect	Respectful
Order	Disorder	Order
Manage	Unmanageable	Manageable
Mature	Immature	Maturity
Comfort	Discomfort	Comfortable
Success	Unsuccess	Successful
Appear	Disappear	Appearance

E. Stemming :

The technique of stemming entails finding the root word inside a given term. When a word is stemmed, its context in the phrase is ignored and it is viewed as an independent unit. By deleting the affixes from the word[10], it accomplishes this. An affix, which is a group of letters, can be added to the beginning or end of a word to alter its meaning. When an affix is added to the beginning of a word, it’s called a prefix, and when added to the end, it’s called a suffix. Stemming reduces the term’s size so that it can finally be reduced to its underlying word by removing affixes. The majority of stemming algorithms merely remove suffixes. Porter's Stemmer is one of the most popular algorithms for suffix removal. Porter Stemmer considers nearly all of the suffixes and keeps track of how words are put together; rules are then developed as a result of these observations. For each type of suffix of a word, this algorithm employs a separate set of rules. For instance, the rule that stipulates the ending should be altered to ”EE” if the word has at least one vowel and consonant in addition to the ending ”EED.” Thus, ”disagreed” changes to ”disagree” Because the algorithm produced irrelevant results when you searched for a word on the internet, it is also employed in search engine optimization. For instance, when you searched for ”employment,” you also got results for ”Unemployment” and ”Employee.” The time it takes to search longer due to this issue[11]. While applying the aforementioned stemming strategy, mistakes may occur. Errors of the Over Stemming and Under Stemming varieties are also possible.

1) Over stemming:

Over stemming is a concept of deleting a word's stem far more often than is necessary. By doing this, it can combine two or more words into one, when there should be two or more, with each root word standing for a separate phrase or token. These words include ”City” and ”Citation,” for instance. Both names may be reduced by some stemming algorithms to the word ”City,” which would erroneously imply that they have the same meaning. An excessive amount of a word's stem is removed when it is over-stemmed.As a result, when two or more words should have been reduced to two or more stem words, they are instead incorrectly reduced to the same root word or stem. Two examples are the university and the universe. Some stemming algorithms may reduce both words to the stem univers, inferring that they have the same meaning—which is obviously false—despite the fact that they don't.

2) Under Stemming:

When words are understemmed, they may be incorrectly reduced to multiple root words when they should have been reduced to only one. Some algorithms, for instance, will shorten the terms ”trouble” and ”troublesome” to ”troubl” and ”troublesom.” The same root term needs to be used to translate both words. So, attention should be taken when selecting the algorithm to stem the word. Under stemming and over stemming are proportionate, meaning that if one is reduced, there is a greater likelihood that the other will grow. When two or more words are wrongly reduced to many root words when they should only be reduced to one, this can happen.For instance, consider the phrases ”data” and ”datum.” These phrases could be translated as dat and datu by some algorithms, which is obviously erroneous. It is necessary to unite these two into a single stem dat. Optimizing such models, meanwhile, might also lead to excessive stemming. Consequently, we must be extremely cautious while dealing with stemming. be reduced to only one.

F. Lemmatization:

A lemma is a base form of a word found in a dictionary and can refer to multiple related words, such as ”come,” ”coming,” and ”came,” which all share the lemma ”come.” Lemma is the basic root word for the terms that are presented. The primary distinction between stemming and lemmatization—which are sometimes confused—is that while both include the trimming or stemming of words, stemming involves a set of rules, whereas lemmatization involves a dictionary in which lemmas are connected to various terms. While lemmatization uses root words that are already defined in the dictionary, stemming occasionally produces roots that have no meaning. Programming languages come with packages that contain the corpus or database types where lemmas of various words can be located. Lemmatization also allows for the definition of parts of speech and the discovery of lemmas in various tenses. Stemming is less precise than lemmatization.[12] The Table II: that illustrates how stemming and lemmatization vary is provided below[19].

G. Word Vectorization:

To construct a list of distinct tokens up to this point, all numerals, special characters, stop words, and uppercase letters have been removed. Each token has undergone lemmatization or stemming to be reduced to its base word. As was mentioned earlier, a machine can only interpret numbers and perform mathematical operations on numbers. It cannot comprehend categorical data. Tokens are transformed into numbers or vectors as a result. Words or phrases from the dictionary are represented by a vector of real numbers in a procedure called word vectorization or word embeddings. To determine word predictions and word similarity/semantics, mathematicians run computations using these numbers. Word similarities are further calculated using either Euclidean Distance or Cosine Similarities [13].

TABLE II. TABLE DISPLAYS STEMMING AND LEMMATIZATION

Word	Stemming	Lemmatization
Wrote	Wrote	Write
Thinking	Think	Think
Remembered	Remember	Remember
Relies	Reli	Rely
Ate	Ate	Eat
Gone	Gone	Go
Won	Won	Win
Ran	Ran	Run
Mistered	Mistered	Mistered

TABLE III. TABLE DSIPLAYS ONE HOT ENCODING

Token set	”name”	”tom”	”live”	”Australia”	”mother”	”him”
Set 1	1	1	0	0	0	0
Set 2	0	1	1	1	0	0
Set 3	0	0	1	0	1	1

Vectorization is the process of converting text data into numerical vectors for processing. There are several methods for vectorization, ranging from basic binary word occurrence features to advanced context-aware feature representations, and the appropriate method depends on the data and problem at hand. Two common vectorization methods are One-Hot Encoding and TF-IDF Vectorization.

H. One hot Encoding:

Each unique token is represented by a column in a table, and each document index is a row in the matrix when the tokens and document indexes are organised in a tabular format. Each cell only has a value of 1 if the word is present in the specific text sample and 0 otherwise. Data that don't relate to one another can benefit from one hot encoding.

Create a vector with as many dimensions as there are words or other tokens in your corpus. Each token has a distinct dimension and is represented by a number between 0 and 1. Consider that three sets of tokens were produced after using the methods that were previously mentioned.[14] 1. [”name”, ”tom”], 2. [”tom”,”live”, ”Australia”], 3.[”mother”, ”live”,”him”]. Below is the Table III that shows how one hot encoding will transform into vector matrix[19].

One-Hot-Encoding has the advantage of producing binary outputs in an orthogonal vector space, but the disadvantage of not scaling well when the number of output labels is large. For instance, in language modeling, where the number of output labels equals the vocabulary size, each input word feature would require a large vector representation[15].

I. TF-IDF Vectorizer:

The full name of IDF is Inverse Document Frequency, which measures the rarity of a word across a set of documents. TF, or Term Frequency, on the other hand, measures the frequency of a term within a single document.The premise of the TF-IDF vectorizer is that if a word appears more frequently in the corpus, it has greater significance and will be given more weight. These more frequently occurring ring terms will have a greater impact on the computations when conducting mathematical operations on these allocated vectors[16].

There are some mathematical formulas for determining the values of TF and IDF for every word in the document T F =Number of times a word is present in doc Total number of words in the doc(1) IDF = log Total number of documents Number of documents that contain the word(2)

Simply multiplying the TF multiplied by IDF values from above yields the value of TF-IDF. Each word or token will have a TF-IDF value, and these scores indicate how important the word is. The more significant a word is to the document, the higher its TF-IDF score. The Table IV below displays the TF-IDF values for the previously used data [19]. The most popular method for identifying the most crucial word in a corpus is TF-IDF. Programmers have the freedom to select words with higher weights thanks to the numerical weights provided to each word. As a result, it simplifies by eliminating words with little meaning, in contrast to hot encoding once which cannot recognize the importance of words, it just shows if a word is present or not[17].

TABLE IV. TF-IDF VALUES

Token set	“name”	“tom”	“live”	“Australia”	“mother”	”him”
Set 1	0.23	0.08	0	0	0	0
Set 2	0	0.23	0.058	0.15	0	0
Set 3	0	0	0.058	0	0.15	0.15

J. Named Entity Recognition(NER):

One of the most common jobs in semantic analysis is named entity recognition, which entails identifying entities within a document. Entities include things like names, locations, businesses, email addresses, and more. Another NLP sub-task called relationship extraction goes a step farther and looks for connections between two nouns. For instance, the semantic category ”lives in” links a person (Susan) to a location (Los Angeles) in the sentence ”Susan lives in Los Angeles.”

IV. Applications Of Nlp

Since a few years ago, there has been tremendous advancement in the field of natural language processing that no one could have anticipated. Understanding the practical applications is essential to comprehending how natural language processing affects our lives.

A. Auto-correct and Auto-complete

When 3-4 letters are entered into any search engine to find something, it returns a list of related search terms. Alternatively, if a word with incorrect spelling is searched, the web browser returns the correct spelling for the word and returns relevant results. This type of problem is faced by everyone in their everyday lives but to which they pay little attention.Both auto-complete and auto-correct assist us in quickly locating accurate search results. Various other firms, such as Google, Facebook and Instagram have begun to use this functionality on their websites[18].

B. Language Translator

I know every one among us tried to find out about a word, or a particular phrase in some different language using Google translator and had noticed that the ease with which it translates our words and phrases into another language is pretty amazing. The technique behind it is Machine Translation. The technique of automatically translating a text given in one language to another language while keeping the meaning of the material is known as translation. Previously, machine translation systems were dictionary- and rule-based, and they were only somewhat successful[18].

C. Social Media Monitoring

People are increasingly embracing social media sites such as Twitter, Instagram, and Facebook to voice their thoughts about a product, policy, or issue. These platforms contain information about a person’s social behavior and can point out what are the likes and dislikes of that person. As a result, observations from such type of data can give information which can be valuable to the organisations. Various organisations now uses a different NLP techniques to evaluate posts on var ious social media platforms and learn how their products are doing in the market. It is used by the government as well as businesses to identify possible dangers to national security. For Example, if there is a terrorists threat to a country, government will do surveillance on suspicious keywords on internet and as well on calling.

D. Voice Assistants

I am sure that at present everyone knows about voice assistant applications like Google Assistant, Amazon Alexa, Siri, Cortana. These all applications uses speech recognition, natural language processing and understanding to understands the voice commands used by us and perform task as per the user instruction. For example, if you says google assistant to play music then it will play music for you. Many people are fully dependent on voice assistance for their day to day life. Voice assistance has turned out to be a very dependable and powerful companion throughout the years.

E. Targeted Advertising

Have you ever noticed that if you look for something on Amazon or any other eCommerce website like Flipkart, Myntra, Bewakoof, etc., Google starts giving you advertisements on other webpages connected to the similar things you searched for on Amazon or any other eCommerce site? This is called targeted advertising.The core of targeted advertising is keyword matching. The advertisements are associated with a key word or phrase, and they are only displayed to those who search for an item using a term similar to the advertisement's keyword.

F. Grammer Checker

When producing professional reports for your superiors or tasks for your professors, grammar and spelling are important aspects. Huge mistakes could, after all, result in your failure or termination.Grammar and spelling checkers are therefore an useful factor for any professional writer. They can increase the readability of your text overall and offer better synonym suggestions in addition to checking spelling and grammar. You'll be surprised to learn that they use natural language processing to produce the greatest possible writing! Millions of sentences are used to train the NLP algorithm so that it can recognise the proper structure. Because of this, it may offer a better synonym, the proper verb tense, or a clearer sentence structure than what you have typed.

V. Conclusion

Although natural language processing is a relatively new field compared to other automation techniques, it has already made significant advancements and is poised to be a major area of information systems research and development in the future. With ongoing improvements in computational text analysis, NLP is constantly evolving. Researchers are exploring how to better understand human language and how people use language to communicate, resulting in the development of tools and techniques that allow computers to understand and manipulate natural language for various tasks, such as string matching, grammar correction, speech recognition and synthesis. Technologies like string matching, grammar checkers, automatic speech recognition, and speech synthesis have all benefited from the application of NLP approaches. There are still many challenges that Natural language processing has to deal with. Human beings communicate in a complex way, a single word can be said in a different tone or expression and at different times it can mean different things for example, a word "Alright" if said in a loud and quick manner then one can assume that the other person is angry but if the same word is said politely, then everything is alright. Other than this problem humans also have different accents which can lead in pronouncing a single word at different frequencies. It can confuse the machine while reading its frequency. Body language can also be used as a medium of communication but still there is less work done in how a machine can deduce the body language of a person and relate it what the person is saying.

References

[1] Natural Language Processing (2001), Elizabeth D. Liddy Syracuse University, liddy@syr.edu

[2] 1995. Allen, J. Natural Language Understanding, 2nd Ed. Reading, MA: Addison-Wesley.

[3] Unstructured Data Analysis-A Survey (2015),K.V.Kanimozhi , Dr.M.Venkatesan, International Journal of Advanced Research in Computer and Communication Engineering

[4] Understanding bag-of-words model: A statistical framework (2010), Yin Zhang, Zhi-Hua Zhou, International Journal of Machine Learning and Cybernetics 1(1):43-52 DOI:10.1007/s13042-010-0001-0

[5] Tokenization as the initial phase in NLP (1992),Jonathan Webster, Chunyu Kit, Proceedings of the 14th conference on Computational linguistics, DOI:10.3115/992424.992434

[6] N-Gram Models (2006), Djoerd Hiemstra, DOI:10.1007/978-0-387-39940-9-935

[7] NLP diagnostics (2021),Richard Gray ,DOI:10.4324/9781003198864-4 In book: Neurolinguistic Programming in Clinical Settings (pp.67-83) Authors:

[8] Stop words(2021), DOI:10.1201,Supervised Machine Learning for Text Analysis in R (pp.37-52), Emil Hvitfeldt,Julia Silge

[9] Rockin' Root Words,2021,Manisha Shelley Kaura,S.R. Kaura, Zak Hamby

[10] Explainability for NLP (2002), DOI:10.1007,Practical Explainable AI Using Python (pp.193-227),Pradeepta Mishra

[11] Special Issue on NLP Semantics 2021, Kunstliche Intelligenz, DOI:10.1007/s13218-021-00728-4, Daniel Hershcovich,Lucia Donatelli

[12] Adaptive Learning for Lemmatization in Morphology Analysis 2015 Communications in Computer and Information Science, DOI:10.1007/978-3-319-17530-0-24,International Conference on Intelligent Software Methodologies, Tools, and Techniques, Mary Ting,Rabiah Kadir,Fatimah Ahmad

[13] Word Embeddings(2019),DOI:10.1201/9780429469275-13,Text Mining with Machine Learning (pp.287-300),Jan Zi ˇ zka,Franti ˇ sek Da ˇ ˇrena,Arnostˇ Svoboda

[14] The research of BP Neural Network based on One-Hot Encoding and Principle Component Analysis in determining the therapeutic effect of diabetes mellitus(2019) IOP Conference Series Earth and Environmental Science, Yuchen Qiao, Xu Yang, Enhong Wu,

[15] Binarsity: a penalization for one-hot encoded features August 2019, Journal of Machine Learning Research 20:1-34, Mokhtar Z. Alaya, Simon Bussy

[16] TF* IDF, January 2009, DOI:10.1007/978-0-387-39940-9-956, Encyclopedia of Database Systems (pp.3085-3086)Publisher: Springer US Ibrahim Abu El-Khair

[17] Introduction to Information Retrieval,January 2008 Publisher: Cam bridge University Press Cambridge, UK, C D Manning,P Raghavan, H Schtextbackslashutze

[18] Application of NLP for Information Extraction from Unstructured Doc uments, 2021,DOI:10.1007/978-981-16-2126-0-54,Expert Clouds and Applications, Shushanta Pudasaini, Subarna Shakya

[19] Talib ul Haq, Darpan Anand, Nitika Kapoor. "The manuscript evaluation through Artificial Intelligence using Natural Language Processing and Machine Learning", 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT),2022