INTRODUCTION TO NATURAL LANGUAGE PROCESSING

***Abstract - Natural language processing (NLP) as of late has received a ton of press for its computational portrayal and investigation of human language. It has a wide scope of utilization, including machine interpretation, email spam identification, information extraction, rundown, clinical, and question addressing, among others. The article is partitioned into four areas, starting with a conversation of various degrees of NLP and parts of Natural Language Generation (NLG), then, at that point, continuing to the set of experiences and development of NLP, the best in class, latest things and troubles and the future scope.***

***We will also be discussing the tools and techniques use in Natural Language processing and how these techniques come into action when we apply them. Comparison between techniques and how each technique works individually. Natural language processing hasn't reached perfection so far, but continuous improvements in this area can certainly touch the line of perfection. Today, various AIs use natural language processing algorithms to recognize and process voice commands from users.***

1. INTRODUCTION

Natural Language Processing refers to the area of computational linguistic which combines human language with statistical data and deep learning models. It allows

**Fig. 1. Broad Classification of NLP**

computers to process human language to the full of its extent, that is, understanding the human emotions and intent behind the words, the speaker or the writer has said or written.

NLP does not treat text like mere sequence of symbols. It understands the thought process of the speaker and why such words are used in an order. It digs a little deeper than just the surface.

Natural Language Processing (NLP) is a part of AI and semantics worried about causing PCs to get explanations or words written in human dialects. Natural language processing was made to make clients' lives more straightforward and to satisfy their craving to associate with PCs in regular language. NLP obliges those clients who need more of an ideal opportunity to learn new dialects or educate them, as not all clients are knowledgeable in machine specific language.

The reason why natural language processing is so important in future is it helps us to build models and processes which take chunks of information as input and in form of voice or text or both and manipulate them as per the algorithm inside the computer. Thus, the input can be speech, text or image where output of an NLP system can be processed Speech as well as Written Text. Different algorithms developed to increase the efficiency of processing the language in text form.

An assortment of rules or a bunch of images can be utilized to characterize a language. Images are blended and used to send or communicate data. The rules have a domineering grasp over images. Natural Language Processing is separated into two sections: Natural Language Understanding and Natural Language Generation, which advances crafted by grasping and creating text.

Phonetics is the investigation of language, and it contains Phonology, which manages sound, Morphology, which manages word creation, Syntax, which manages sentence structure, Semantics punctuation, and Pragmatics, which manages perception.

1. ****RELATED WORK****

Many researchers worked on NLP, building tools and systems which makes NLP what it is today. Tools like Sentiment Analyser, Parts of Speech (POS) Taggers, Chunking, Named Entity Recognitions (NER), Emotion detection, Semantic Role Labelling made NLP a good topic for research.

Jeonghee et al., 2003 [1] Sentiment analysis works by extracting emotions about a particular problem theme. Sentiment analysis consists of subject-specific feature terminology extraction, emotions extraction and association by relationship analysis.

(Namrata Tapswi, Suresh Jain., 2012) [2] Parts of speech taggers for the languages like European languages. The most procedures for part of speech can work efficiently on European languages, but it won’t on Asian languages or middle eastern languages.

McDonald et al., 2005 [3] Chunking – it is also known as Shadow Parsing, it works by labelling segments of sentences with syntactic correlated keywords like Noun Phrase and Verb Phrase. Every word has a unique tag often marked as Begin Chunk (B-NP) tag or Inside Chunk (I-NP) tag.

(Alan Ritter., 2011) [4] Usage of Named Entity Recognition in places such as Internet is a problem as people don’t use traditional or standard English. This degrades the performance of standard natural language processing tools substantially. By annotating the phrases or tweets and building tools trained on unlabelled, in domain and out domain data. It improves the performance as compared to standard natural language processing tools.

1. LEVELS OF NLP

The 'levels of language' are a basic method for communicating Natural Language Processing, which helps with the production of NLP message by finishing the Content Planning, Sentence Planning, and Surface Realization stages.



**Fig. 2. Phases of NLP Architecture**

Phonetics is a part of study that concentrates on the significance of language, its specific situation, and its many structures. Coming up next are a portion of the critical terms in Natural Language Processing: -

1. **Phonology**

Phonology is a part of etymology that arranges with the deliberate association of sound. Phonology comes from the Ancient Greek prefix phono-, which alludes to voice or sound, and the addition - logy, which alludes to word or discourse. (Clark et al.,2007) [5].

1. **Morphology**

Morphemes are the littlest units of importance addressed by the different components of the word. Morphemes are the beginning stages for morphology, which is the investigation of the idea of words. The word precancellation, for instance, might be separated morphologically into three morphemes: the prefix pre, the root cancellation, and the addition. People can part any obscure word into morphemes to appreciate the significance since the understanding of morpheme is something very similar across all words.

1. **Lexical**

People and NLP frameworks both comprehend the significance of individual words in Lexical. Word-level perception is helped by an assortment of handling procedures, the first is the expansion of a grammatical feature tag to each word. Semantic portrayals can be subbed by words with a solitary significance at the lexical level. The idea of the portrayal in an NLP framework change relying upon the semantic hypothesis utilized.

1. **Syntactic**

This level accentuates inspecting the expressions of a sentence to decide the expression's linguistic design. This level requires the utilization of both language structure and a parser. The portrayal of the expression that uncovers the underlying reliance joins between the words is the result of this degree of handling. There are assortment of sentence structures that might be discouraged, and which, thusly, confine the parser choice.

1. **Semantic**

The vast majority accept that importance is chosen in semantics, yet this isn't true. Semantic handling recognizes the different implications of an expression in view of the connections between the sentence's statement-level implications. This degree of handling can incorporate semantic disambiguation of words with various implications, like how syntactic disambiguation of expressions can be confounded as various grammatical features is refined at the syntactic level. (Elizabeth D. Liddy,2001) [6].

1. **Discourse**

While grammar and semantics work with sentence-length units, NLP's talk level works with message units that are longer than a sentence, for example it doesn't comprehend multi-sentence messages as a progression of single-sentence sentences

1. NLP TOOLS AND TECHNIQIUE

NLP can be used through SaaS (Software as a Service) tools or using open-source libraries.

SaaS tools are powerful, out-of-the-box, cloud-based solutions that can be implemented with little or no code. SaaS platforms often provide pre-trained NLP models and APIs. These are for users who need more flexible low-code options. A professional developer or programmer who wants to simplify their work.

Open-source libraries, on the other hand, are free and flexible, allowing you to fully customize your NLP tools. However, because they are aimed at developers, they are very complex to understand and require machine learning experience to build open-source NLP tools. Fortunately, however, most are community-driven frameworks, so you can count on a lot of support

The Natural Language Toolkit (NLTK) using Python is one of the leading tools for NLP modelling. NLTK focuses on research and education in the field of NLP and is supported by an active community and a variety of language processing tutorials, sample datasets, and resources, including comprehensive manuals on language processing and Python.

This library takes some time to master but is considered a great playground for hands-on experience with NLP. The modular structure of NLTK provides numerous components for NLP tasks such as tokenization, tagging, stemming, parsing, and classification.



**Fig. 3. NLP Toolkit**

There are different techniques in NLP that we can use to extract text from a given text snippet:

* **Sentence segmentation** - Defines sentence boundaries in the given text. That is, where one sentence ends and another begins. Sentences are often marked ended with the punctuation mark ‘.’.
* **Tokenization** - Identifies various words, numbers, and other punctuation mark and treat them individually.
* **Stemming** - It eliminates the endings from words, for example, 'eating,' which is diminished to 'eat.'
* **Part of speech (POS) tagging** - Assign a unique part-of-speech tag to each word in the sentence. Designating a word as a noun or adverb.
* **Parsing** - The specified text falls into various categories. To answer a question like this part of the sentence, modify another part of the sentence.
* **Named Entity Recognition** - Identifies people, places, times, and other entities in a document.
* **Co-Reference resolution** - This is to define the relationship between a particular word in a sentence and the previous and next sentences.

1. **APPLICATIONS OF NLP**

Machine interpretation, email spam discovery, data extraction, synopsis, and question addressing are only a couple of the spaces where Natural Language Processing might be utilized.

1. **Machine Translation**

Because the internet connects the majority of the globe, the task of making knowledge open and available to everyone is a challenge. The language barrier is a significant hurdle in making information available. There are several dialects, each with its own sentence structure and punctuation. Machine Translation is the process of interpreting sentences from one language to the next with the use of a translation engine like Google Translate. The challenge with machine interpretation advancements isn’t only understanding words but maintaining the relevance of sentences, as well as syntax and tenses. The genuine AI gathers as a good deal record as they could that appears to be equal throughout dialects and crunches it to decide the chance that something in Language is comparable. Language B has anything to do with A. (Tillmann et al., 1997) 86].

Diagram

Description automatically generated

**Fig. 4. Machine Translation Diagram**

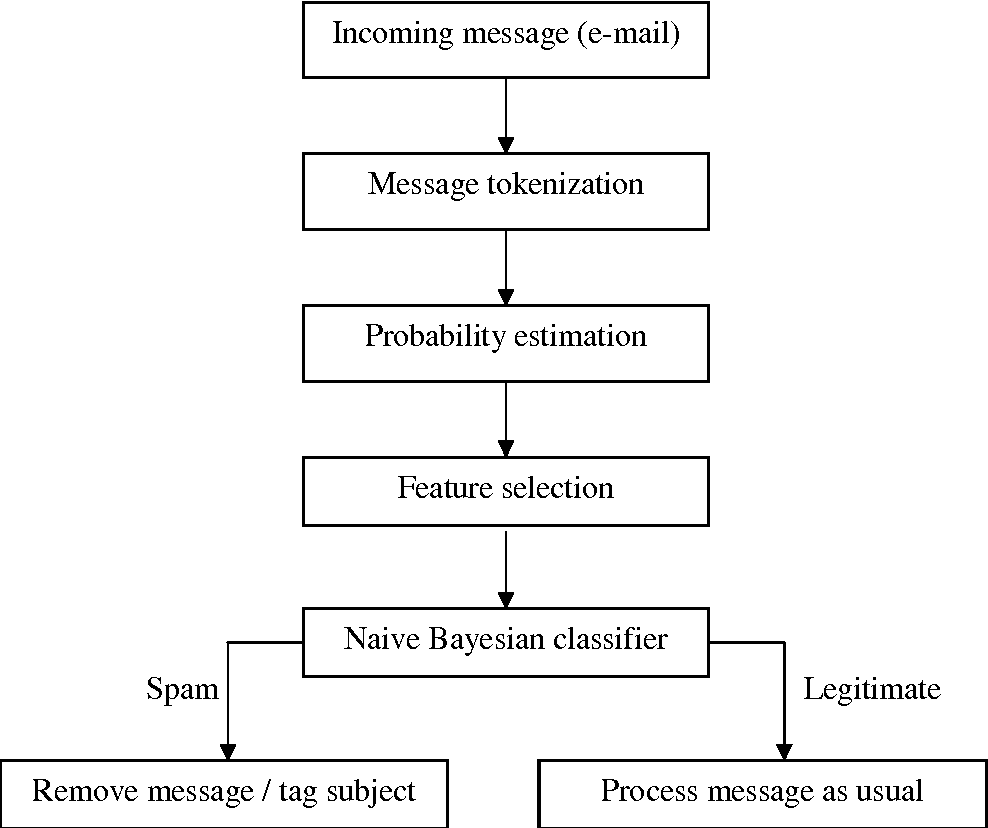
1. **Text Categorization**

Classification frameworks take a large amount of data, such as authority archives, military loss reports, market data, and newswires, and divide it into present groups or lists. For example, The Carnegie Group's Construe framework (Hayes PJ ,Westein ; 1991)[9] imports Reuters articles and saves time by doing the work that would otherwise be completed by staff or human indexers. Order frameworks have been used by a few businesses to categorise annoyance complaints or objection requests and direct them to the appropriate task groups. Email spam channels are another use of text order. Spam channels are becoming increasingly important as the primary line of defence against unwanted messages. A fictitious negative and fictitious positive Spam channels are

at the heart of NLP innovation, and it's been reduced to a test of eliminating significance from text strings.

1. **Spam Filtering**

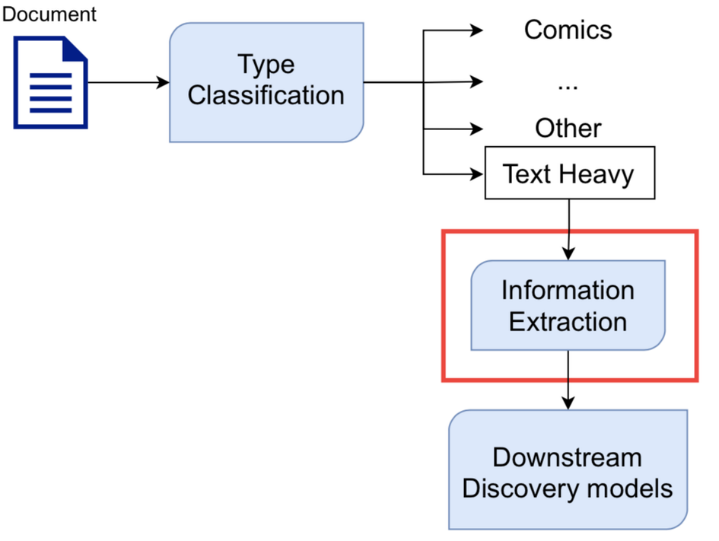
It uses text order, and several AI algorithms, such as Rule Learning (Cohen, 1996)[10], have recently been used to message categorization or Anti-Spam Filtering, Bayes, Nave (Sahami et al., 1998)[9], Memory based Learning (Androutsopoulos et al.,2000b)[11]. Consolidating various pupils every now and again is a good idea. It is preferable to use these techniques since the classifier is obtained by preparing information rather than by supplying it. Despite its effortlessness, the credulous bayes is preferred due to its display. Two types of models have been used in text categorization (McCallum and Nigam, 1998) [12]. Both modules assume that a correct vocabulary is present. However, in the first approach, a record is created by first selecting a subset of jargon and then repeatedly using the chosen words, to some extent once regardless of request. The model is known as the Multi-variate Bernoulli model. It records which words are used in an archive, independent of the number of words or the request. A record is constructed in the second model by selecting a group of word events and organising them in any request.



**Fig. 5. Spam Filtering Diagram**

1. **Information Extraction**

The recognition of indications of interest in text-based information is a concern for data extraction. Extraction of components such as names, locations, events, dates, times, and expenses are a powerful approach for certain applications to summarise the data relevant to a client's requirements. The programmed identifiable evidence of crucial facts may increase precision and productivity of a coordinated quest thanks to a space explicit internet searcher. To extract the important fields of examination papers, stowed away Markov models (HMMs) are used. These deleted text sections are utilised to allow for the examination of specific areas, the display of indexed lists, and the matching of references to papers. For example, you may have seen the spring up advertisements on any sites showcasing the new products you may have looked at on an online based store with restrictions.

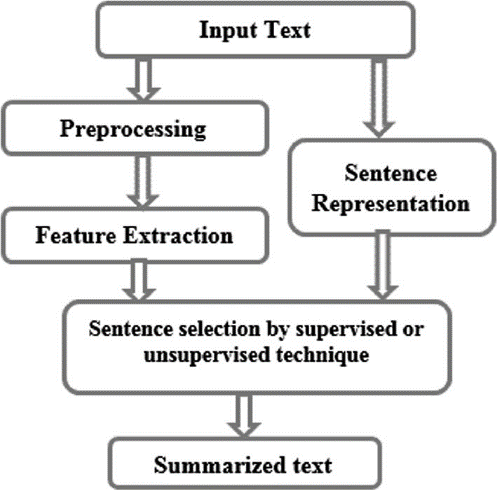


**Fig. 6. Information Extraction Diagram**

Over the next few years, information disclosure will be a hot topic of discussion. Information disclosure research employs a variety of techniques to extract important data from source documents, including

1. **Text Summarization**

This NLP application is used to summarize text by extracting the most important information. The main goal here is to reduce the process of going through vast amounts of data in news content, legal documentation and scientific papers. There are 2 ways of using natural language processing for text summarization: extraction based, which extracts key phrases and creates a summary without adding any extra information AND abstraction-based summarization, which paraphrases the original content to create new phrases.



**Fig. 7. Text Summarization Diagram**

1. **Predictive text**

Things like autocorrect, autocomplete, and predictive text are so commonplace on our smartphones that we take them for granted. Autocomplete and predictive text are similar to search engines in that they predict things to say based on what you type, finishing the word or suggesting a relevant one. And autocorrect will sometimes even change words so that the overall message makes more sense. They also learn from you. Predictive text will customize itself to your personal language quirks the longer you use it. This makes for fun experiments where individuals will share entire sentences made up entirely of predictive text on their phones. The results are surprisingly personal and enlightening; they’ve even been highlighted by several media outlets

1. CONCLUSION

As explained in the context above, NLP has a long history of research dating back to the 1950s, but many of its applications have only recently emerged. With the introduction of Google as the leading search engine, the increasingly digital world, and the rise in employment, NLP has sneaked into our lives almost unnoticed. But this is the reason for multiple conveniences in our daily lives. Developers can use NLP to perform tasks such as speech recognition, sentiment analysis, translation, automatic grammar correction when typing, and automatic answer generation. NLP is a challenging field because it deals with human languages ​​that are very diverse and can be spoken in different ways. NLP's various algorithms and techniques give developers a wider range of improvements and accurate results. NLP adoption is expected to gain momentum over the next few years with the introduction of more personal assistants, smartphone enhancements, and the development of big data to automate everyday human tasks.

**FUTURE SCOPE**

With the available information constantly growing and increasingly sophisticated, accurate algorithms, NLP is surely going to grow in popularity. It’s altering the way of interaction between humans and machines. The previously mentioned uses of NLP are proof of the fact that it’s a technology that improves our quality of life by a significant margin.As much as 80% of the information that surrounds us is unstructured. For this reason, NLP is one of the largest fields of data science. Organising this data is a considerable challenge that’s being tackled daily by countless researchers. Continuous advancements are being made in the area of NLP, and we can expect it to affect more and more aspects of our lives.

Considering the market scenario in the case of NLP. The buzz of NLP in the market is growing in an exponential manner which is expected to touch the mark of $ 16 billion by 2021 with the compound growth rate of 16 % annually. The reason behind this growth is rising of the chatbots, urge of discovering the customer insights, transfer of technology of messaging from manual to automated and many other tasks which are required to be automated and involve language/Speech at some point. Though, as stated above the functionality of NLP revolves around language/speech which refers to words in its basic raw form. No matter what the medium of the communication is, whether it is verbal or written, words are the fundamental unit of the functionality of NLP. But in the current NLP, there seems to be a difference in the performance of NLP, when it is handling texts and when it is handling voice. This challenge is going to be addressed soon surely. Let’s consider the different scenarios concerning NLP and future

**REFERENCES**

1. Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 427-434). IEEE.
2. Tapaswi, N., & Jain, S. (2012, September). Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences. In Software Engineering (CONSEG), 2012 CSI Sixth International Conference on (pp. 1-4). IEEE.
3. McDonald, R., Crammer, K., & Pereira, F. (2005, October). Flexible text segmentation with structured multilabel classification. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 987-994). Association for Computational Linguistics.
4. Ritter, A., Clark, S., & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1524-1534). Association for Computational Linguistics.
5. Nation, K., Snowling, M. J., & Clarke, P. (2007). Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. Advances in Speech Language Pathology, 9(2), 131-139.
6. Liddy, E. D. (2001). Natural language processing.
7. Jain, Aditya & Kulkarni, Gandhar & Shah, Vraj. (2018). Natural Language Processing. International Journal of Computer Sciences and Engineering. 6. 161-167. 10.26438/ijcse/v6i1.161167.
8. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997, September). Accelerated DP based search for statistical translation. In Eurospeech.
9. Hayes, P. J. (1992). Intelligent high-volume text processing using shallow, domain-specific techniques. Text-based intelligent systems: Current research and practice in information extraction and retrieval, 227-242.
10. Cohen, W. W. (1996, March). Learning rules that classify e-mail. In AAAI spring symposium on machine learning in information access (Vol. 18, p. 25).
11. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).
12. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.
13. McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).