

A Review Study of Natural Language Processing Techniques for Text Mining

Aastha Tyagi
Amity University
Noida, India

Abstract:- In today's technological Eire, text analysis and natural language processing are very much important aspects in artificial intelligence and many other types of technology. Using natural language processing (NLP), text mining (also known as text analytics) transforms unstructured text within documents and databases into normalized, structured data that may be used for analysis or to train machine learning algorithms. The technological advancement is having responsibility for helping people towards getting advantageous benefits with usage of this. In this paper, text analysis and natural language processing in artificial intelligence is being convicted. Natural language is that particular language which helps the machine for reading the text with the help of stimulating the ability of human for understanding a natural language like English or any other type. Content analysis method was taken into consideration during the time of research i.e., the literature and study which is already in existence was taken into consideration to be more profound about the topic. The conclusion of the study has been convicting that importance of text analysis and natural language processing in artificial intelligence is essential to a greater extent.

Keywords: Artificial intelligence, Natural Language Processing, Text analysis, etc.

I. INTRODUCTION

Natural language processing is basically being defined as the branch of artificial intelligence which provides help to the computers for understanding, manipulating, and interpreting the human language. Natural language processing draws so many types of disciplines, with an inclusion of computer science and computational linguistics which is being put for fulfilling the gap in existence in between human communication and understanding of computer. Artificial intelligence text analysis is simply defined as the process for extracting the information from inside a larger amount of textual data which is present [5]. There are basically two types

of NLP techniques which are Syntactic Analysis and Semantic analysis. Syntactic analysis - or parsing - examines text using fundamental grammatical principles to detect a sentence structure, arrange words and how they connect. Semantic analysis relies on text capture. First, it examines the significance of each term (lexical semantics). If we talk specifically about businesses, natural language processing is the prospect which can particularly help in automating the complete process pertaining the understanding of comments exerted by customers on a larger scale. It will be helping them for making decisions on the basis of data towards improving the business [6].

The importance of natural language processing can be defined in such a way that it helps the computers for communicating with the humans in their own languages and scale other types of tasks in relation to the languages [8]. If we take example, natural language processing is defined to make it possible for all the computers to words reading the text, hearing the speech, interpreting it, measurement of the sentiment and also determining the parts which are important to be taken care of. Human languages are very much complex and diverse in nature. We have been expressing ourselves in so many types of ways and both verbally and in writing also [10]. Natural language processing is very much important in artificial intelligence as it helps for resolving the ambiguity in the language and adding a much more useful structure numerically for the data to words many types of downstream applications like as speech recognition or text analytics.

II. NLP AND TEXT ANALYTICS

Natural language processing is defined to be going hand-in-hand with text analysis which specifically will be counting, grouping and categorizing words for extracting the structure including meaningful content from larger amount of content. The text analysis is being used for exploring the textual content and deriving new type of variables from raw data which might be visualized, Filtered and further being used in the form of your inputs for predicting the models and other types of statistical methods [2]. Natural language processing and text analysis are being used altogether in so many types of applications such as:

- Investigation discovery: identification of the patterns including the clues residing in the emails or reports which has been written for helping in the detection and solving process of crimes happened.
- Subject matter expertise: classification of the content to words the topic which is meaningful so that people will be having ability for acting and discovering the trends.
- Social media analytics: tracking awareness including the sentiment pertaining specific types of topics and identification of key influencers [6].

Analyzing the text is very much important skill for the readers who are successful. Analysis of a text is being involved of breaking down the ideas and structure for understanding it in a better way and thinking critically about it and drawing the conclusion [1]. Text analysis in artificial intelligence is clearly working in this particular direction of breaking down the ideas and structure of the text so that it can be clearly understood by the human operating that [8].

Natural Language Processing

Natural Language Processing (NLP) allows computers to comprehend the language of humans. NLP examines the characteristics and the significance of the phrases behind the scenes and then utilizes algorithms to extract meanings and provide results. In other words, the human language makes sense to do various jobs automatically [9].

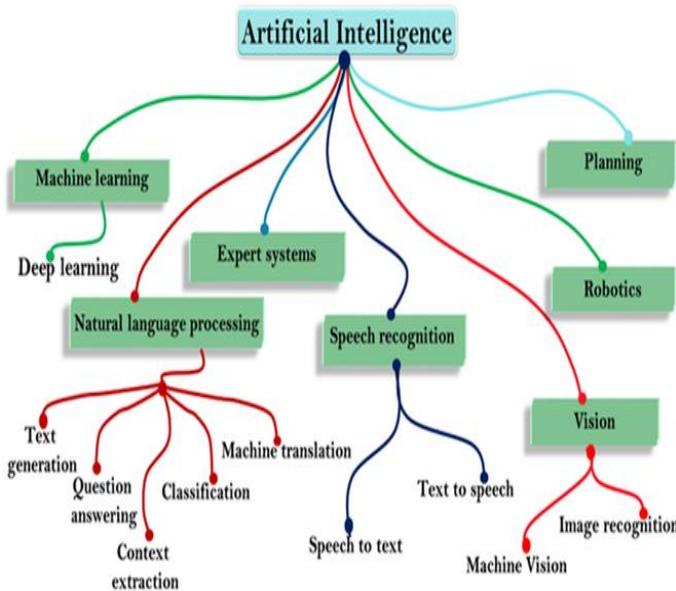


Fig1: NLP in AI

Natural Language Processing (NLP) examines the understanding of human language by computers and translations. With NLP computers can understand the written or spoken language and execute tasks like translation, extraction of keywords, categorization of the subjects, and more [11].

- The correctness of the response rises with the quantity of relevant information in the questions.
- Users may ask inquiries about any topic and get a straight answer in seconds.
- It is simple to deploy.
- It is less expensive to use a software than to hire a human. A human may take two or three times longer to do the same activities than a computer.
- The NLP system gives responses to natural language queries.
- Allow you to do more language-based data without weariness in an impartial and consistent manner compared to a person.
- The NLP method helps a computer to speak in its language with a person, and measures other language activities.

NLP Techniques

In order to assist computers, comprehend text Natural Language Processing (NLP) uses two techniques: syntactic analytics and semantic analysis.

Syntactic Analysis

Syntactic analysis - or parsing - examines text using fundamental grammatical principles to detect a sentence structure, arrange words and how they connect.

Some of its principal subtasks are:

- Tokenization involves dividing a text up into smaller pieces called tokens (which may be phrases or words) to simplify handling of material.
- Part of the speech tag labels tokens like verb, adverb, adjective, substance, etc. This helps determine the meaning of words (for example, the term "book" refers to several objects whether employed as a verb or a substantive).
- Lemmatization & stemming consists in reducing inflected phrases to their basic form to facilitate analysis.
- Stop-word removal often eliminates words that don't contribute semantic value, such as I, they, have, etc. [12].

Semantic analysis

Semantic analysis relies on text capture. First, it examines the significance of each term (lexical semantics). Then, the arrangement of words and also what they signify is examined in context. The primary tasks of semantical analysis are as follows:

Word meaning disambiguation attempts to determine the meaning in which a word is used in a particular context [4].

Uses of NLP

Some of the major uses of NLP are as follows:

- *Translation of Language*

In recent years, machine translation technology has advanced greatly, with the translations of Facebook reaching superhuman performance in 2019.

Translation technologies allow companies to interact in various languages, enhance their worldwide communication or open up new markets [11].

- *Extraction of text*

Text extraction allows you to extract predefined text content. This application helps you identify and extract keywords and important characteristics (such as product codes, colors and specifications) and named entities when dealing with huge quantities of data [3].

- *Chatbots*

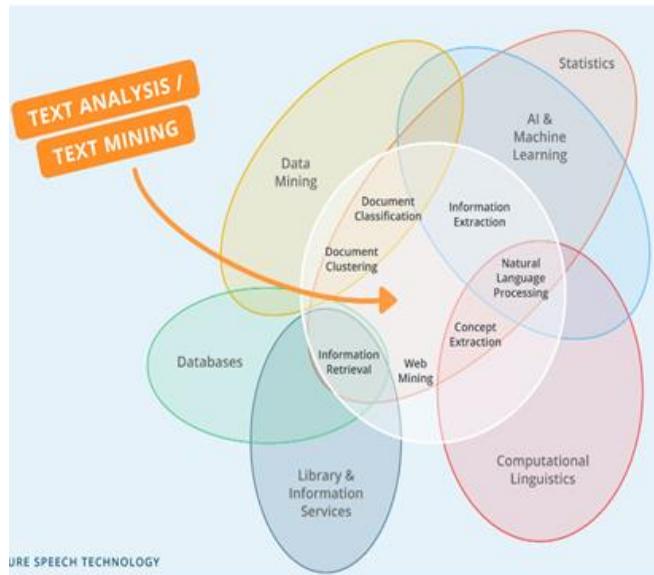
Chatbots are AI systems intended for text or voice interactions with people.

Chatbots are being increasingly used for customer services since they are able to provide help 24/7 (speed up response times) concurrently manage numerous inquiries and relieve human workers from replying to repeated questions. Chatbots learn active from every contact and improve user intent so that you can depend on them to do repeated and easy jobs [13].

Text analytics

Text analysis is the automated text analysis. Although used interchangeably occasionally, there is a distinction between text analysis and text mining. In the meanwhile, text analysis offers you with more in-depth quantitative information to

make educated choices. If applied to unstructured input (also called as open-ended input), text analysis offers insight into trends, patterns and consumer feelings to discover and prioritize methods to improve customer experience [6].



The benefit of unstructured feedback is that it is consistent with the words of your consumers and gives the greatest insight into the thinking of your customers. In this response there are many methods you may digest much more value as possible.

Text analytics techniques

Here are just a few of the major technologies involved:

- *Artificial Intelligence (AI)*

The computer system's capacity to execute activities that usually require human intelligence. These tasks include voice recognition and decision-making but are not restricted to these. This is why Text Analytics processes huge amounts of text and categorizes it automatically to ease the examination of your unstructured comments [10].

- *Machine Learning (ML)*

Machine learning is an AI component, but it is distinct. ML focuses on the capacity of a computer algorithms to effectively learn from experience and automatically adapt to enhance performance without human programming. Text Analytics utilizes machine learning to identify how fresh text pieces should be classified based on previously processed text, and to evaluate whether the categories used to categorize these text pieces should be improved according to the patterns identified.

- *Deep learning (DL)*

A controlled, specialized subset of artificial intelligence that encompasses the capacity of a computer system to analyze data and make judgments on other data. In Text Analytics, deep learning may be used to better grasp the context in unstructured feedback and to enhance the accuracy of automatic text analysis.

- *Sentiment Analysis*

By use of NLP and text analysis to process and automatically evaluate texts for a positive, negative or neutral attitude of the

text. Sentiment Analysis is an excellent beginning point for you to start evaluating your unorganized feedback and rapidly discover significant and developing problems, improvement opportunities from your interviewees [3].

- *Classification of text based on rules*

By use of a word dictionary or a lexicon to teach a computer system how to classify fresh materials. In Text Analytics, rule-based text categorization is used to assign feelings or subjects to any text it analyzes automatically [3].

III. LITERATURE REVIEW

(Kongthon, 2009) indicated the implementation of online tax system with the usage of natural language processing in artificial intelligence. This particular implementation was shown up for convicting the idea of usage of natural language processing and text analysis in artificial intelligence so that future can be secured. Majority of the high-level natural language processing applications be involved of aspects which has been emulating the intelligent behavior including the apparent comprehension of the natural languages.

Following the review of methods in this subject (Jean, 2014) offer a strategy based on the sampling of significance, which enables us to utilize a large-scale vocabulary without raising the training complexity of the NMT model, to solve machine translation. Then they propose an approximation training method based on (biased) sampling that will enable you to train an NMT model with a considerably wider target wording, using a single neural network tuned to optimize translation performance.

The process of refining or creating, or approaching, a practical method to natural language processing (NLP) may become fairly difficult. (Collobert, 2011) feel that their contribution to this area of research is a significant milestone in addressing linear training algorithms that best benefit from the enormous advancements in computer hardware. In this instance, these authors are based on huge, unlabeled data sets utilizing the NLP tagger, and the training algorithm discovers internal depictions that are helpful for all jobs.

(Vinyals, 2015) fittingly conduct extremely concentrated research into foreign language grammar by utilizing a Recurrent Neural Network with an attention 4 strategy for producing phrase parse trees. Syntactic constituency parsing is a basic issue with a broad variety of applications in linguistics and natural language processing. For decades, this issue has been the focus of intensive study and thus extremely precise domain-specific parsers exist.

(Fisher, 2010) Text mining is the examination of a vast number or collection of textual resources to produce new knowledge. For example, by converting text into data that can be utilized for further research, the objective of text mining is to uncover useful information in text. NLP is one of the methods used in text mining to achieve this goal.

Thesaurus, lexicon, ontology and up-to-date entities are all required for any natural language processing program to function.

(Mnasri, 2019) One of the most frequent uses of text mining and NLP is social media monitoring, where a pool of user-generated material is analyzed to identify mood, emotions, and awareness connected to a topic.

(Moreno,2016) Natural Language Processing is a collection of techniques for processing such data to understand underlying meaning (NLP). Machine Learning (ML) approaches are used to develop applications that classify, extract structure, summarize, and translate data, which can be speech, text, or even an image.

IV. METHODOLOGY

Objectives of the study

Some of the major goals considered throughout the research were:

- To study about Text Analytics and its techniques used in Artificial Intelligence.
- To study about NLP in Artificial Intelligence and its various techniques.

Research Methodology

Content analyses method regarded for performing the analysis of the literature which is already in existence and having a detailed knowledge Text Analytics and NLP in Artificial Intelligence. Various research paper and previous analysis have been considered and literature that before now in existence have been extracted with the help of keywords. Extraction of the information from those literatures have been done and post to the extraction, verification of the information has been done through the appropriate sources for its accuracy and reliability. After the testing of accuracy, details were contained in paper and irrelevant information have been removed. After the inclusion of the relevant information in the paper, subsequent analysis has been achieved about this particular topic.

V. CONCLUSION

In accordance with the research conducted it can be clearly sustained that NLP is much better methods used in comparison to other methods as NLP is having ability for recognizing the text and speech also and other method such as text mining only deals with the evaluation of text quality. For NLP system you are having less knowledge requirement of skills like NLTK, proficiency in neural networks while in text mining you are having requirement of knowledge about different aspects such as cosine similarity or feature hashing, text processing, like Perl or python. Knowledge about statistical method is also very important aspect in text mining. In accordance with the research of last 10 years it can be clearly illustrated that NLP is being used more in comparison to other statistical methods as it is more feasible, easy to use and less knowledge aspect requirement.

REFERENCES

- [1] Lehnert, Wendy, and Beth Sundheim. "A performance evaluation of text-analysis technologies." *AI magazine* 12, no. 3 (1991): 81-81.
- [2] Ahonen, Helena, Oskari Heinonen, Mika Klemettinen, and A. Inkeli Verkamo. Applying data mining techniques in text analysis. Report C-1997-23, Dept. of Computer Science, University of Helsinki, 1997.
- [3] Lalwani, T., Bhalotia, S., Pal, A., Rathod, V., & Bisen, S. (2018). Implementation of a Chatbot System using AI and NLP. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)* Volume-6, Issue-3.
- [4] Cavazza, Marc, Srikanth Bandi, and Ian Palmer. "Situated AI in video games: integrating NLP, path planning and 3D animation." In *AAAI 1999 Spring Symposium on Artificial Intelligence and Computer Games*, pp. 6-12. 1999.
- [5] Mnasri, Maali. "Recent advances in conversational NLP: Towards the standardization of Chatbot building." *arXiv preprint arXiv:1903.09025* (2019).
- [6] Lee, L. (2002, July). A non-programming introduction to computer science via NLP, IR, and AI. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* (pp. 33-38).
- [7] Mamou, J., Pereg, O., Wasserblat, M., Eirew, A., Green, Y., Guskin, S., ... & Korat, D. (2018). Term set expansion based nlp architect by intel ai lab. *arXiv preprint arXiv:1808.08953*.
- [8] Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, 3(6), 57-64.
- [9] Yang, Yiwei, Eser Kandogan, Yunyao Li, Prithviraj Sen, and Walter S. Lasecki. "A study on interaction in human-in-the-loop machine learning for text analytics." In *IUI Workshops*. 2019.
- [10] Ise, Orobora Anderson. "Integration and analysis of unstructured data for decision making: Text analytics approach." *International Journal of Open Information Technologies* 4, no. 10 (2016).
- [11] Fisher, Ingrid E., Margaret R. Garnsey, Sunita Goel, and Kinsun Tam. "The role of text analytics and information retrieval in the accounting domain." *Journal of Emerging Technologies in Accounting* 7, no. 1 (2010): 1-24.
- [12] Day, M. Y. (2020). Artificial Intelligence for Text Analytics.