International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Natural Language Processing in the Era of BIG DATA

P.ROHITH VARMA

ABSTRACT :

Natural Language Processing (NLP) has become a revolution in the age of big data. The large and diverse vocabulary generated by big data makes it possible to develop advanced NLP models using machine learning algorithms and distributed computing techniques. The combination of NLP and big data has led to the emergence of powerful language models such as BERT and GPT, allowing NLP to better understand content and provide insights in many applications such as thinking, machine translation, response and custom NLP text. The application of NLP in the big data environment provides solutions to many problems in industries. Business intelligence can benefit from collecting data and providing real-time insights, while collaboration can be enhanced through networking. Sentiment analysis helps improve product and market research by allowing organisations to understand customers' thoughts and preferences.This study demonstrates the effectiveness of NLP in analysing large datasets, especially for sentiment analysis using the Map Reduce framework.

Overall, this paper highlights the potential and challenges of integrating NLP with Big Data and gives insight into how the combination works. This power can be used in many ways, leading to a better understanding of language and data analysis.

1.INTRODUCTION

In the age of big data, Natural Language Processing (NLP) has become a revolutionary method. The large and diverse vocabulary derived from big data supports the development of NLP models using advanced machine learning algorithms and distributed computing techniques. This evolution has allowed NLP to better understand context and context by creating powerful language models such as BERT and GPT. Sentiment analysis, machine translation, question answers, and custom NLP

NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is a text classification whose purpose is to extract information from text. Text classification can be done using rule-scripts continue to evolve, providing insights in many ways. However, with the benefits of big data come concerns about privacy and bias. Overall, the integration of NLP and big data has revolutionised the understanding of language, allowing us to harness the power of human language in unprecedented ways.

2.PREVIOUS WORK

With the emergence of segmented connectivity models and their short-term communication model (LSTM) in 2014, the NLP landscape continues to evolve. This development opens applications such as machine translation to question answering. In 2018, Google's Bidirectional Encoder Agents for Transformers (BERT) introduced the first Transformer model, which changes the context of content by understanding bidirectional perception. At the same time, the revolutionary Transformer architecture emerged, which took advantage of the self-development mechanism to improve the structure of long-term dependencies. Later successes such as XLNet, ERNIE and the educational transformation concept have added to these advances, demonstrating the power of integrating external knowledge.

In particular, the Text-to-Text Converter (T5) provides a unified framework for various NLP projects in 2020, while the KEVID-19 NLP project (2020-2021) demonstrates the potential of NLP to examine the pace of many medical records. helping scientists understand the nuances and consequences of disease.

exponential growth of textual information in the digital era, data analysis has become an essential component of NLP, allowing us to make sense of unstructured text and derive valuable knowledge from it. This interdisciplinary approach combines the principles of natural language processing, machine learning, and statistics to handle the complexity and diversity of language data.

3. NLP Applications

Table T1 shows different percentages which are used in different sectors.

based or skill-based methods. Therefore, the Al component is not required for NLP. Rule-based methods are hard-coded rules or expressions for finding in text. This is also called translation. The challenge for a system using speech recognition is to respond appropriately even when there are errors in words.

The word or phrase must be added to the dictionary by the human/scientist using the legal method. When it comes to AI methods, letting the software build its own dictionary is essential. The machine detects words that appear together in a sentence to form a sentence, and then detects which words in the same sentence make up a sentence. It provides a deeper understanding of the text.

Data analysis

Data analysis is just one of my favourite things to do. One of the most exciting things when it comes to NLP is discovering new information. Distributed text is a new type of document that I have not worked with before, so I look forward to delving into all these resources.

Data analysis over NLP refers to the process of applying analytical techniques to extract insights, patterns, and meaningful information from large volumes of natural language data. And With the

4 .How NLP can help in the following areas ;

NLP can help with the following tasks:

1. Conversation

Siri on iOS is a good example of NLP in conversation. Online marketing and sales of self-help tools and translation apps can be done using NLP. Using NLP, professionals in relationship management have advanced to the level where these skills can be used to solve traditional customer service problems.

2. Marketing Intelligence

Tracking a social media hashtag requires an analyst to access all possible "hashtags" and keywords related to the topic. NLP can probe questions embedded in natural language, including all possible situations, and reduce data errors in determining how many people are talking about word of mouth.

Table 1. % NLP APPLICATIONS IN INDUSTRY

INDUSTRY	**C.S**	**Health**	**finance**	**E-Retail**	**S.Media**
Emotion Analysis	80%	60%	70%	75%	905
Chatbot	85%	40%	65%	70%	60%
Text Tagging	70%	50%	80%	60%	65%
NER	75%	55%	60%	50%	70%
Intent Detection	80%	45%	55%	60%	65%
Content Abstraction	70%	30%	40%	55%	50%

Keywords:

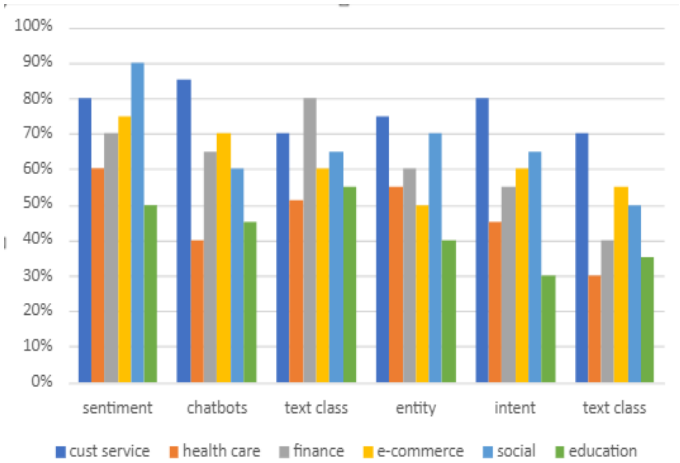
1. C.S=customer service
2. NER=Named entity recognition
3. S.Media=Social media

This table mainly describes about NLP application are widely used in many industry and also it shows different percentages in each and every industry

products or services are popular in a particular target. All big data analytics provided by all major solutions are expected to be done using NLP, because the world's big data will be more than 44 trillion GB, yes, so NLP's scope for big data will only expand.

A natural language representation can be a stream of Unicode characters (usually UTF-8) and requires a simple mechanism to convert the stream of characters into words, phrases, and numbers. Major techniques include language recognition, sentence discovery, lemmatization, parsing, pattern

Figure 1. Graphical representation of NLP Applications



3. Sentiment Analysis

Brands can now collect data beyond customer feedback via direct chat. NLP can identify which

5. Steps to improve

Step 1 : Decide level of understanding and Assess Possibility

Next, you have to decide which understanding points you need - macro or micro. While micro-understanding (comprehension from a single sentence or sentence) often helps with comprehensive understanding the two can be very different.

Also, when determining the level of understanding, you should evaluate the feasibility of the project because not all understanding of NLP can be done at a reasonable cost and time.

Step 2 : Extricate substance for large scale or smaller scale understanding

When you decide to start an NLP project, you will need to understand the data more thoroughly, so this " large scale understanding " comes into play. It is important that you do the following:

• Classify

• Cluster records

• Extract topics

extraction, symbolization, location and sentence extraction.

Table 2 . NLP in Big Data Growth %

YEAR	NLP Growth in Big Data %
2011	5%
2012	8%
2013	10%
2014	12%
2015	18%
2016	27%
2017	33%
2018	40%
2019	55%
2020	70%
2021	85%
2022	96%

• Semantic search

If you want to understand your words and sentences, go to the micro-understand, a place from the text, real or make a relationship. This is useful for:

• Extracting abbreviations and their meanings

• People, companies, products, places, dates, etc

Step 4: Enhancing Visibility

Obtaining content from multiple sources and then extracting information from that content will involve multiple steps and multi-level calculations. Therefore, it is important to ensure traceability for all yields. You'll be able at that point backtrack through the framework and bolster analytics and target examination to confirm precisely how the information was produced.

• Content Tagging

• Duplicate and near-duplicate detection

Step 3 : Provide human input

Understanding content cannot happen without human intervention. Find new designs, create, clean, or choose the names of known sites, etc. You need people for

Many of these processes can cause unconsciousness. In large systems, you must take the human element into account and incorporate it into your NLP system architecture.

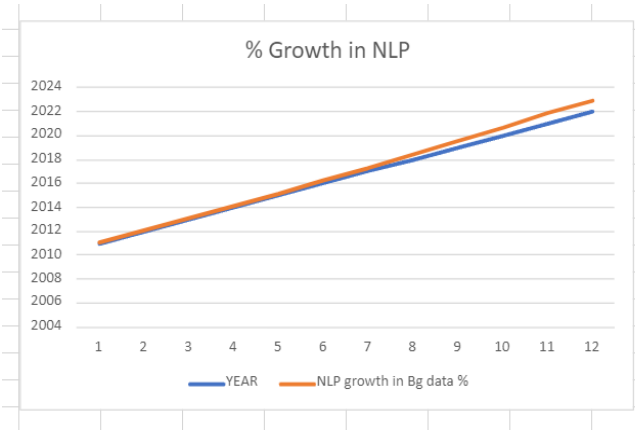
Remember that it is important to continually analyse data at every step of the process for the most excellent understanding of natural dialect substance. The entire process could appear overwhelming, but utilizing these steps and methods as a direct can assist you make a more effective and effective way to secure, collect, and change non-essential information.

6. What problems can natural language processing for big data solve?

Whatever the industry, every business today relies on collecting a lot of data. For example, law firms conduct a large amount of research, past and current legal documents, correspondence, e-mail correspondence and various government documents and other evidence. Companies can obtain clinical trial data and information, physician information, patient information and data, patent and regulatory information, and research, new research from competitors.

Because this data is mostly words, processing big data's words presents a growth opportunity,

Figure 2. Representation of % growth of NLP



especially for using content from the big store and expressing the structure, connection and differentiation of different materials.

Collaboration:

Natural language technology is used in many interactive applications today, such as mobile assistants such as online banking and personal shopping tools, and some automatic translations. Users ask questions every day and get correct answers instantly. It is a win-win for both the customer and the company, because customers can easily communicate with the companies they do business with anytime, anywhere, while companies know how to save money by reducing the number of calls from live chat.

Business Intelligence:

Big data dictionaries play an important role in managing large and complex data and gaining useful insights. This detailed information is a powerful tool for compressing and illuminating the vast amounts of information contained in large databases. Unlike traditional methods that require specific content to store information, users do not need to describe the content because they can interact with the content using their personal guidelines. Thus, this effective interaction increases the efficiency and effectiveness of data search and analysis, enabling users to uncover hidden patterns, trends and interactions.

Table 3. Estimated Usage of NLP Application in future

NLP Application	Current Usage	Future Usage	Used year Estimated
Text Analysis	40%	60%	2025
Emotion Analysis	30%	50%	2024
NER	20%	40%	2026
D.C	25%	45%	2023
Topic Modelling	15%	35%	2025
Language translation	10%	30%	2024
Speech - Text	20%	40%	2026

Keyword;

1. NER=Named entity recognition
2. DC=document classification

Sentiment Analysis:

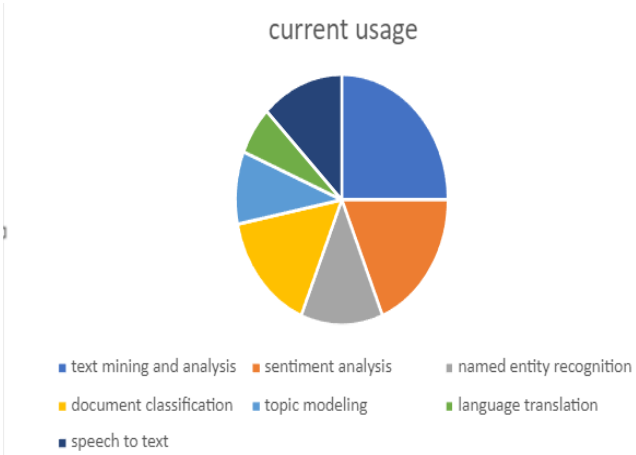
As online consumers continue to grow, social networks are important sources of information, rich but noisy. Using natural language for sentiment analysis, organisations can understand what people think and talk about their brands and products - what users think about the service, product or concept/strategy. This is a good way to find information about the market and current and potential customers that might not otherwise be found (including opinions as well as customers, interests and needs/opinions, demographics). This information can be used for product development, business intelligence and business research.

Text Modelling

Modelling in NLP is an essential tool for tapping into the potential of big data. With the increasing prevalence of text from sources such as social media, reviews and articles, text models can understand and analyse large volumes of content.

This model supports activities such as sentiment analysis, classification, aggregation, and interpretation, enabling businesses to uncover valuable insights, make informed decisions, and deliver user experiences. Paper models leverage advanced technologies and distributed computing to transform the challenges of managing and making sense of large volumes of data into opportunities for data-driven innovation and insight.

Figure 3. Pie chart represents different % of NLP usages



Using advanced techniques, such as removing stop words, URLs, and other user comments, removes data noise and improves accuracy. Also, tweaking the numbers and hashtags allowed the model to focus on the emotional content of tweets rather than irrelevant information. The integration of the MapReduce framework was successful in processing large volumes of tweet data. This approach simplifies business parallelization, are assessed through a speculation test utilising the

6 . RESULTS AND METHODOLOGY ;

In this study, we evaluated the effectiveness of big data analytics using descriptive and NLP techniques. Analysis is performed from tweets' data to perform sentiment analysis and classify tweets as positive or negative. The MapReduce framework has been adopted to process big data and improve the versatility of the investigation strategy. After past information utilisation tokenization, piece of discourse (PoS) labelling, and lemmatization, halt words, URLs, etc. Numerous progressed strategies were utilised, counting evacuating clients, numbers and hashtags. Assessment Utilising Gullible Bayes Classification Calculation. It turned out that by expanding the exactness of the

.

Discussion

Sensitivity analysis results demonstrate the effectiveness of NLP in processing large datasets and extracting useful information from text. Classification of positive and negative emotions in tweets with 73% accuracy shows that NLP techniques can be used to shine a light on the public, which is important for businesses and organisations to understand customer thoughts and make informed decisions.

Preliminary steps such as word segmentation, PoS tagging, and lemmatization play an important role in improving the accuracy of inference models. This step helps keep the text consistent by making it easier for the Naive Bayes classifier to identify patterns and relationships in the data.

CONCLUSION

The increment within the number of clients of the networks and the increment within the information transferred to these systems have made huge volumes and diverse sorts of information. Information volumes are tremendous and require explanatory strategies that can scale successfully as information grows in measure. In expansion, conventional investigation strategies ought to be created to prepare the diverse sorts of information created. In this consider, the viability of enormous information investigation was assessed utilising expressive and NLP strategies. These intelligences MapReduce framework. Sentiment examination employs a Gullible Bayesian classification calculation to classify tweets as positive and negative feelings. To extend the exactness of credulous Bayesian classification, information is preprocessed utilising tokenization, PoS labelling and lemmatization. In expansion, numerous progressed methods are utilised: expel halt words, URLs, notices of other clients, numbers and hashtags. Utilise MapReduce as a system to make strides in the scalability of the examination strategy. By progressing the exactness of the Gullible Bayes classification by 5%, we accomplished 73% exactness within the information utilised. At long last, the precision of the strategy can be assisted by making strides by including other NLP and preprocessing strategies.reduces computation time and shortens analysis time of big data.

However, despite the positive results, there are still issues to be addressed in the combination of NLP and big data. Privacy concerns are still a major concern, as the analysis of large volumes of personal data requires careful consideration of data protection laws. Also, bias in NLP models is a concern. As NLP models learn from training data, biases in training data can be exposed and cause bias. Continuous evaluation and reduction of bias in NLP models are important to ensure fair and unbiased results.

Taken together, the results of this study demonstrate the potential of NLP to extract useful insights from big data. Sentiment analysis of tweets using NLP techniques provides a good example of how businesses can use data to understand customer sentiment and improve decision making. However, addressing privacy and impartiality concerns remains important for the role and ethics of NLP in big data. Further research and development is needed to improve the ability and effectiveness of language modelling techniques to process diverse and expanding information in the context of big data.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, pp. 1-135, 2008.

[2] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 241-249.

[3] Wikipedia. (2015). *List of Applications in NLP* .

[4] L. Zhang, Y. Jia, B. Zhou, and Y. Han, "Microblogging sentiment analysis using emotional vector," in *Cloud and Green Computing (CGC), 2012 Second International Conference on*, 2012, pp. 430-433

[5] P. Nakov, Z. Kozareva, A. Ritter, "Semeval-2013 task 2: Sentiment analysis ," 2013.

[6] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning utilising twitter hashtags and smileys," in Proceedings of the 23rd Worldwide Conference on Computational Etymology: Blurbs, 2010, pp. 241-249.

[7]. NLP Application with in the article of india today https://timesofindia.indiatimes.com/education/upskill/natural-language-processing-11-real-life-examples-of-nlp-in-action/articleshow/101521214.cms

[8] How NLP used in data industry in Analyst insight https://www.analyticsinsight.net/a-loot-at-the-10-best-nlp-companies-in-india/