e-ISSN: 2349-2678, p-ISSN:2961-6050

available online on www.ijpba.in

International Journal of Pharmaceutical and Biological Science Archive

NLM (National Library of Medicine ID: 101732687)

Index Copernicus Value 2022: 72.59

Volume 11 Issue 6; 2023; Page No. 55-62

Natural Language Processing in Machine-Learning Techniques

PV Ramana Murthy [1*] & Dr. Rakesh Kumar Giri [2]

[1] Research Scholar, Sunrise University, Alwar.

[2] Assistant Professor, Sunrise University, Alwar.

Article Info: Received: 14-07-2023 / Revised: 13-09-2023 / Accepted: 16-10-2023
Address for Correspondence: PV Ramana Murthy
Conflict of interest statement: No conflict of interest

Abstract

Natural Language Processing (NLP) is one of the most exciting areas of artificial intelligence (AI). **T**he main aim of the study is to Natural Language Processing In Machine-Learning Techniques**.** The primary objective of this study is to use machine learning techniques and algorithms for the assessment of technical documentation translations. A knowledge discovery method that included the steps of acquiring data, preprocessing data, selecting a suitable data mining technique to uncover patterns within the data, and interpreting them was used to accomplish this goal.

Keywords: Language, processing, Machine, Learning, Data, Technical, Documentation

1. INTRODUCTION

Natural Language Processing (NLP) is one of the most exciting areas of artificial intelligence (AI). This is because of tools like text generators that write well-organized essays, chatbots that make people think they're intelligent, and text-to-image programs that turn any description into a photorealistic picture. In the past few years, computers have become much better at understanding natural languages, programming languages, and even chemical and biological patterns that look like language, like DNA and protein structures. The newest AI models are opening these areas so they can figure out what text means and produce creative, meaningful output. Natural Language Processing (NLP) will be an important tool for businesses in most fields in the years to come. NLP is a way for computers to use AI to understand text or voice data and then write or speak some text or speech in response. This guide will explain what Natural Language Processing is and how it will change how businesses handle tasks that used to be done by hand and how they talk to their customers. NLP has been used for a long time in robots for customer service. It is also being used more and more in marketing, banking, human resources, healthcare, media, and more. A lot of people are interested in NLP now, especially since OpenAI released ChatGPT, a language model. Millions of people and businesses in all kinds of fields have started using ChatGPT to make text with AI. This is a big step toward making NLP useful for business. This guide will tell you everything you need to know about Natural Language Processing and show you how it is used in the real world!

NLP MODEL AND NLP MACHINE LEARNING

Machine learning is an important part of Natural Language Processing because it lets computers learn from data and get better at understanding voice or writing data over time. This is important because it lets NLP programs get better over time, which makes them work better and give users a better experience. Using deep learning to process natural language In the past few years, many deep learning models for

natural language processing (NLP) have been created to make text analytics and NLP tasks better, faster, and more automatic. When it comes to NLP jobs, machine learning, and especially deep learning methods, have become very useful. In deep learning, neural networks with many layers are used to learn how to describe data at higher and higher levels of abstraction. This lets the network find complicated trends in the data, which helps NLP models work better. Sentences in human language are made up of words and phrases that fit together in a certain way. Recurrent Neural Networks (RNNs) are a type of deep learning that works especially well with and can study linear data like text, time series, financial data, speech, audio, video, and more.

2. LITERATURE REVIEW

Awais, Muhammad & Ashrafi, Bilal & Abbas, Asad. (2023). This in-depth study looks at the wide range of effects deep learning has on artificial intelligence (AI) and chatbots. Deep learning, with its strong neural network designs, has greatly expanded AI's abilities and made it possible for chatbots to connect with people in smarter and more relevant ways. This study gives a review of the basics of deep learning, including neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs), as well as how they can be used in different AI projects. In particular, it looks into how deep learning methods like word embeddings, focus mechanisms, sequence-to-sequence models, and deep reinforcement learning can be used in robot systems. The study also talks about the problems and limits of using deep learning to make chatbots, emphasizing how important it is to be able to explain things and think about what is right and wrong. It also gives a full review of all the current deep learning-based chatbot systems, looking at how well they work, how scalable they are, and how satisfied their users are. At the end of the study, new trends and goals for future research are talked about. Deep learning is emphasized as a way to make chatbots smarter and improve user experiences in AI-powered talking agents.

Johri, et.al. (2021). A speech helper, a chatbot, or a recommender system would not be smart without natural language processing (NLP). It all starts with an initial unit that reads the data (voice or writing) that is given and then starts to make sense of it. Once the data has been properly processed, the machine moves on to the next steps to answer or finish the job. This is because NLP is not a single field of study. Instead, it is a combination of computer science, information engineering, artificial intelligence (AI), and languages. NLP is the study of how computers and human speech work together. Speech recognition, machine translation, automatic text summary, part-of-speech tagging, and other areas are all part of NLP. In general, NLP is used in a lot of real-time situations, like in smart homes and smart offices with Alexa, Cortana, Siri, and Google Assistant. NLP has a past that goes back to the 1950s. It has come a long way and gotten better since then. This essay talks about the background of NLP, how it has changed over time, its tools and methods, and how it can be used in various fields. The study also talks about how machine learning and artificial neural networks (ANNs) can help NLP get better.

Imamguluyev, Rahib. (2023). A speech helper, a chatbot, or a recommender system would not be smart without natural language processing (NLP). It all starts with an initial unit that reads the data (voice or writing) that is given and then starts to make sense of it. Once the data has been properly processed, the machine moves on to the next steps to answer or finish the job. This is because NLP is not a single field of study. Instead, it is a combination of computer science, information engineering, artificial intelligence (AI), and languages. NLP is the study of how computers and human speech work together. Speech recognition, machine translation, automatic text summary, part-of-speech tagging, and other areas are all part of NLP. In general, NLP is used in a lot of real-time situations, like in smart homes and smart offices with Alexa, Cortana, Siri, and Google Assistant. NLP has a past that goes back to the 1950s. It has come a long way and gotten better since then. This essay talks about the background of NLP, how it has changed over time, its tools and methods, and how it can be used in various fields. The study also talks about how machine learning and artificial

neural networks (ANNs) can help NLP get better.

Khan, Saad & Khan, Konal. (2023). Improvements in Machine Learning have made huge steps forward in many areas, from computer vision and natural language processing to healthcare and business. This paper gives an outline of recent progress in Machine Learning, closing the gap between new ideas in theory and real-world uses. We start with a quick look at some basic ideas and algorithms and then move on to cutting-edge methods that have completely changed the field, like deep learning and reinforcement learning. We show how these advances are leading to new ideas and changing businesses by focusing on real-life examples. We also talk about the problems and moral issues that come up as Machine Learning becomes more common, which shows how important it is to use methods that are responsible and include everyone.

Sharifani, Koosha & Amini, Mahyar. (2023). Machine learning and deep learning have become very useful tools very quickly in many areas, like health, natural language processing, speech and picture recognition, and more. We look at the methods and uses of machine learning and deep learning in this piece. We also talk about their pros and cons and where they might be going in the future. We also talk about the problems that come with these tools, such as data privacy, moral issues, and the need for openness in the way decisions are made. Many people think that machine learning and deep learning are two of the most important new developments in artificial intelligence. In the past few years, they've become more famous because they can make predictions, look at big data sets, and give us information that we couldn't get before. This piece will talk about what machine learning and deep learning are, how they vary, how they can be used, and how they affect different fields. Deep learning and machine learning are changing how we use technology and opening up new ways to make things better. There are already big changes happening in many fields because of these technologies, and they could continue to change the world. This book talks about the basics of machine learning and deep learning, as well as their differences, how they can be used, and how they affect society. This piece wants to help you understand the possibilities of machine learning and deep learning better by focusing on recent study and writing. It also wants to show you what these technologies mean for the future.

3. METHODOLOGY

3.1 RESEARCH METHODOLOGY

The primary objective of this study is to use machine learning techniques and algorithms for the assessment of technical documentation translations. This thesis aims to address two distinct issues. Initially, technical document translations will undergo classification and evaluation using a machine learning system that has access to the original content. In the second iteration, an algorithm will be enhanced for the identical goal without any prior knowledge of the original.

3.2 SPECIFICATION OF A DATA MINING APPROACH

In this part, we will discuss the KDD-steps 5 to 7, which include the selection of a data mining strategy, the selection of an algorithm, and the actual data mining step.

4. RESULTS

4.1 EMPIRICAL DATA

In the next part, the numerical representations of the methodology discussed in chapter 3 are presented. Table 4.1 provides an overview of the total number of phrases, characteristics, and documents that were used for the tasks that were provided.

Table 4.1: Statistics of used data sets.

Number of Automated Translation Systems	3
Number of Technical Documents	14
Original Number of Sentences in all Documents	30, 000
Number of Sentences used per Translation System for the Data Sets	22, 327
Approximate Number of Sentences in the Data Sets	44, 654
Total Number of Data Sets	9
Number of Attributes needing a Reference Translation	14
Number of Attributes needing no Reference Translation	18
Total Number of Attributes	32
Maximum Amount of Available Attribute Values	1, 428, 928
Number of Created Fictitious Documents	19, 190

During the process of extracting sentences from technical documentation, there were thirty thousand lines that had text fragments. 8000 of these lines had not been extracted properly, which led to erroneous text fragments that did not constitute legitimate sentences. This was since sentence extraction is not a straightforward process. Considering this, each final data set that was used for the purpose of training and testing the machine learning algorithms had a total of 22327 phrases for each translation language. The purpose of combining each data set was to guarantee that there was a fair distribution of the label's professional translation and automated translation. This was accomplished by combining two sets of phrases, one of which was translated professionally and the other of which was translated automatically. As a consequence of this, every single data set ended up holding a total of 44654 sentences. It was determined that these words were taken from fourteen different technical publications that dealt with the management of a virtual machine software. In order to conduct a more in-depth investigation into the validity of the algorithm on various sizes of technical papers, the produced documents were distributed in sizes ranging from five to three thousand words per documentation. The construction of documents of bigger sizes was more difficult than the creation of smaller documents owing to the restricted number of words and the quantity of training data that was required to create a useful classification model. It is vital to highlight that this was the case. In addition, the importance of the information obtained from papers of a smaller size cannot be overstated. This is because it was anticipated that the number of sentences required to accurately categorize a document would fall somewhere between sixty and three hundred sentences. The nine data sets that were used are shown in Table 4.2. These data sets were generated by using the three machine translation systems as candidates and references, respectively, and by utilizing the Freetranslation software to generate a round-trip translation for six of the nine data sets.

Table 4.2: Used combinations of references and candidates.

Candidate	Reference 1	Reference 2	Round-Trip Reference
Google Translate	Bing	—	Google RTT via Free translation
Google Translate	Free translation	—	Google RTT via Free translation
Google Translate	Bing	Free translation	Google RTT via Free translation
Bing Translator	Google	—	Bing RTT via Free translation
Bing Translator	Free translation	—	Bing RTT via Free translation
Bing Translator	Google	Free translation	Bing RTT via Free translation
Free translation	Google	—	—
Free translation	Bing	—	—
Free translation	Google	Bing	—

The computation of the attributes, as well as the construction of the database, and the setup from, were all carried out in Java, with Eclipse serving as the development environment. Open-source tools, such as the Stanford Natural Language Processing Package and the Language Tool Core Package, were used to include the metrics that were stated. Most of the metrics were modified to conform to the specifications of the machine learning job that was presented, and then they were computed to create the final database that was used to train and verify the machine learning algorithm. A further preparation of the data set was carried out to transform it into an appropriate environment for the algorithm to train on before the optimization of each method was carried out. These processes, which include the optimization process, were carried out with the assistance of the RapidMiner program, and figure 4.1 provides a detailed illustration of the procedure.

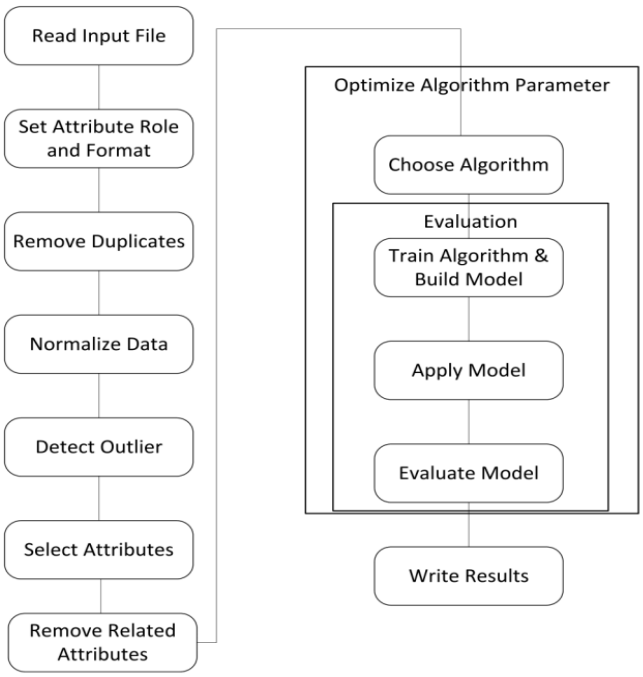

Figure 4.1: Preprocessing of the data set and optimization of the machine learning algorithm using RapidMiner.

The first phase of the preprocessing procedure, which is called Remove Duplicates, is carried out once the input data set has been read and the attribute types have been established. It is essential to restrict the effect of numerous identical data entries, and this phase eliminates all of the items in the data set that are similar to one another, with the exception of one. The range transformation is then used to standardize the data to the next step. The following stage, which is the identification of outliers, is made possible because of this mapping of all characteristics to the same value range. This makes it possible to compare any two attributes more easily. Using class outlier factors, which are characterized by comparing data entries to one another using the Euclidean distance metric, outliers may be identified and identified from the data. To the machine learning job, the five percent of data entries that deviate the greatest are referred to as outlying values and are not taken into consideration. In addition, the qualities that are permitted for the work at hand are chosen in accordance with the research question. This process involves the elimination of fourteen attributes for the second research question. Identification of strongly correlated attributes and elimination of those attributes if the correlation is more than 90% is the last stage in the preprocessing process. This phase is performed in order to further reduce the amount of time required for calculation and to eliminate redundancies among the attributes.

Table 4.3: created models and optimizations.

	Research Question 1	Research Question 2	**Total**
Built Decision Trees	480, 000	300, 000	780, 000
Built Neural Networks	6, 000	2, 000	8, 000
Built k-Nearest Neighbors	3, 500	1, 500	5, 000
Built Support Vector Machines	2, 500	1, 000	3, 500
Total Number of Optimizations	492, 000	304, 500	796, 500

4.2 RESULTS

These are the outcomes of the experiments that will be presented in this section. The part will discuss the outcomes that occurred when the algorithm was able to access the original documents but will demonstrate the accomplishments that were made for the identical work when the algorithm was unable to access the contents of the papers. The findings of the tested algorithms and the document-based analysis that was produced will be presented in each subsection, with the best and most significant outcomes being highlighted.

4.2.1 Research Question 1

The assessment of translations with the presence of knowledge about the source material is the subject of the first research topic. Following this, the best results for each of the several configurations that were evaluated will be shown, arranged according to the algorithm that was used. The precise ranges that were used in the process of improving the models are detailed. To offer a concise summary, table 4.4 displays an accumulation of the findings that were provided in a more extensive manner later on. Additionally, it represents the overall performance of the algorithms that were used.

Table 4.4: Averages and standard deviations of the tested algorithms.

Algorithm	Accuracy	Standard Deviation
Decision Tree	68.69%	0.014
Artificial Neural Network	**70.33%**	0.014
k-Nearest Neighbor	69.74%	**0.009**
Naive Bayes	65.84%	0.019
Support Vector Machines	61.31%	0.017

The results that provide the best accuracies for sentence predictions when utilizing Decision Trees are shown in Table 4.5. The F1-scores for each of these outcomes are also included. To assess each candidate-reference combination that is shown in the table, fifty thousand Decision Trees were constructed and analyzed. Every accessible characteristic was included in the data that was utilized, and it was normalized. The data was reduced by up to five percent because of an outlier identification and further reductions owing to the elimination of duplicates.

Table 4.5: best Decision Tree results for the respective candidate reference combinations.

Decision Tree					
Candidate	Reference 1	Reference 2	Accuracy	F1-Automated	F1-Professional
Google	Bing	—	69.09%	0.724	0.649
Google	Free translation	—	67.91%	0.706	0.646
Google	Bing	Free translation	**70.48%**	0.744	0.651
Bing	Google	—	70.18%	0.706	0.698
Bing	Free translation	—	68.75%	0.692	0.683
Bing	Google	Free translation	70.23%	0.715	0.689
Free translation	Bing	—	66.22%	0.689	0.629
Free translation	Google	—	67.85%	0.687	0.669
Free translation	Bing	Google	67.52%	0.676	0.675

The outcomes that are the most impressive for artificial neural networks are shown in the table that follows. 150 iterations of optimization were performed on each candidate-reference combination. The data preparation is comparable to the optimization of the Decision Tree that came before it. Every lineup includes all of the qualities that are offered. In addition to the normalization of the characteristics, the data set is reduced by up to five percent of the outliers. It is not possible to consider duplicates.

Table 4.6: best Artificial Neural Network results for the respective candidate-reference combinations.

Artificial Neural Network					
Candidate	Reference 1	Reference 2	Accuracy	F1-Automated	F1-Professional
Google	Bing	—	69.93%	0.731	0.659
Google	Free translation	—	68.08%	0.709	0.647
Google	Bing	Free translation	70.93%	0.744	0.663
Bing	Google	—	**72.24%**	0.729	0.715
Bing	Free translation	—	69.14%	0.696	0.687
Bing	Google	Free translation	71.67%	0.723	0.711

Table 4.7 displays the outcomes of the Application of the k-Nearest Neighbor method that was used. Fifty models were constructed and assessed for each candidate-reference combination that was considered. In this case, the results are representative of the various models that attained the greatest levels of accuracy. The data preparation is quite like the settings that were stated before. The data set is reduced by up to 5% outlier and owing to the elimination of duplicates. Additionally, the characteristics are normalized to execute the outlier identification.

5. CONCLUSION

A knowledge discovery method that included the steps of acquiring data, preprocessing data, selecting a suitable data mining technique to uncover patterns within the data, and interpreting them was used to accomplish this goal. Finally, the findings were used for more study. The document database was partitioned into a sentence level, which resulted in the production of nine data sets, each of which had 22, 327 data items for each of the two kinds of translations (professional translation and automatic translation). A total of 32 metrics and qualities were selected and put into action; however, only 18 of them required a reference translation for the calculation procedure, while 14 of them did or did not.

REFERENCES

1. Awais, Muhammad & Ashrafi, Bilal & Abbas, Asad. (2023). A Comprehensive Study on Deep Learning in Artificial Intelligence and Chatbots.
2. Johri, Prashant & Khatri, Sunil Kumar & Al-Taani, Ahmad & Sabharwal, Munish & Suvanov, Shakhzod & Chauhan, Avneesh. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. 10.1007/978-981-15-9712-1_31.
3. Imamguluyev, Rahib. (2023). The Rise of GPT-3: Implications for Natural Language Processing and Beyond. International Journal of Research Publication and Reviews. 4. 4893-4903. 10.55248/gengpi.2023.4.33987.
4. Khan, Saad & Khan, Konal. (2023). Advancements in Machine Learning: From Theory to Practice.
5. Sharifani, Koosha & Amini, Mahyar. (2023). Machine Learning and Deep Learning: A Review of Methods and Applications. 10. 3897-3904.
6. Liew, Cher Don. (2021). Survey of Machine Learning Algorithms Used in Natural Language Processing and Understanding Tasks. 10.13140/RG.2.2.25017.65127.
7. Chotirat, Saranlita & Meesad, Phayung. (2021). Natural Language Processing with "More Than Words – BERT". 10.1007/978-3-030-79757-7_11.

8. Shaik, Thanveer & Tao, Xiaohui & Li, Yan & Dann, Christopher & Mcdonald, Jacquie & Redmond, Petrea & Galligan, Linda. (2023). A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis.
9. Sodhar, Irum Hafeez & Buller, Abdul Hafeez. (2020). Natural Language Processing: Applications, Techniques and Challenges. 10.22271/ed.book.784-.
10. Chowdhary, Prof. (2020). Natural Language Processing. 10.1007/978-81-322-3972-7_19.