JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

AI-based Natural Language Processing (NLP) Systems

Aastha Gour

Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002

Abstract. The goal of artificial intelligence (AI) known as Natural Language Processing (NLP) is to instruct computers how to read, comprehend, and even create new varieties of human language. NLP has applications in many different fields, including healthcare, banking, education, and even customer service. In recent years, natural language processing (NLP) has seen major advancements as a result of the availability of enormous datasets, powerful technology, and complex algorithms such as neural networks. These factors have made it possible for machines to read and analyse unstructured input in the form of text, voice, and video. The problems of data bias and quality, interpretability and explainability, domain-specific language, support for several languages, contextual comprehension, and adversarial assaults are some of the issues that still need to be resolved. In order to solve these obstacles, further research has to be done in a variety of different areas, such as the collection and annotation of data, the explainability of models, and the robustness of attacks. Despite these obstacles, natural language processing (NLP) has a bright future; in the years to come, we can likely anticipate a great deal of development and innovation in this sector.

Keywords. Natural Language Processing, NLP, Artificial Intelligence, AI, machine learning, algorithms, models, text analysis, speech analysis, sentiment analysis, text classification, named entity recognition, machine translation, summarization, question-answering, chatbots, virtual assistants, healthcare, finance, education, data quality, bias, interpretability.

I. Introduction

Sometimes abbreviated as NLP, Natural Language Processing is a branch of AI that studies how language is used in interactions between humans and computers. Creating algorithms and models that allow computers to understand, interpret, and generate human language is at the heart of this process. Thanks to the exponential growth of data and developments in machine learning techniques, NLP has made significant progress in recent years. Natural language processing (NLP) has emerged as an important area of study due to its wide range of potential applications[1].

Computers can analyse and analyse unstructured data like text, speech, and video with the use of natural language processing (NLP) [2]. Machine translation, question answering, summarization, emotion analysis, text classification, naming entities, and a plethora of other tasks fall under this umbrella. Natural language processing (NLP) systems have improved in accuracy and sophistication in recent years thanks to the availability of large datasets, high-priced technology, and complicated algorithms like neural networks [3].

Natural language generation, or the production of text or speech that is indistinguishable from that produced by people, is one of the most significant applications of natural language processing (NLP) [4]. This is arguably the most

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

significant use case for natural language processing. Potential applications for this include chatbots, digital assistants, and the creation of algorithmic content. For instance, chatbots might be used in customer service to provide instant and efficient responses to queries and complaints. Virtual assistants like Siri and Alexa may be instructed to do a wide range of tasks by simply being spoken to. Making phone calls, playing music, and setting alarms are all instances of such activities [5].

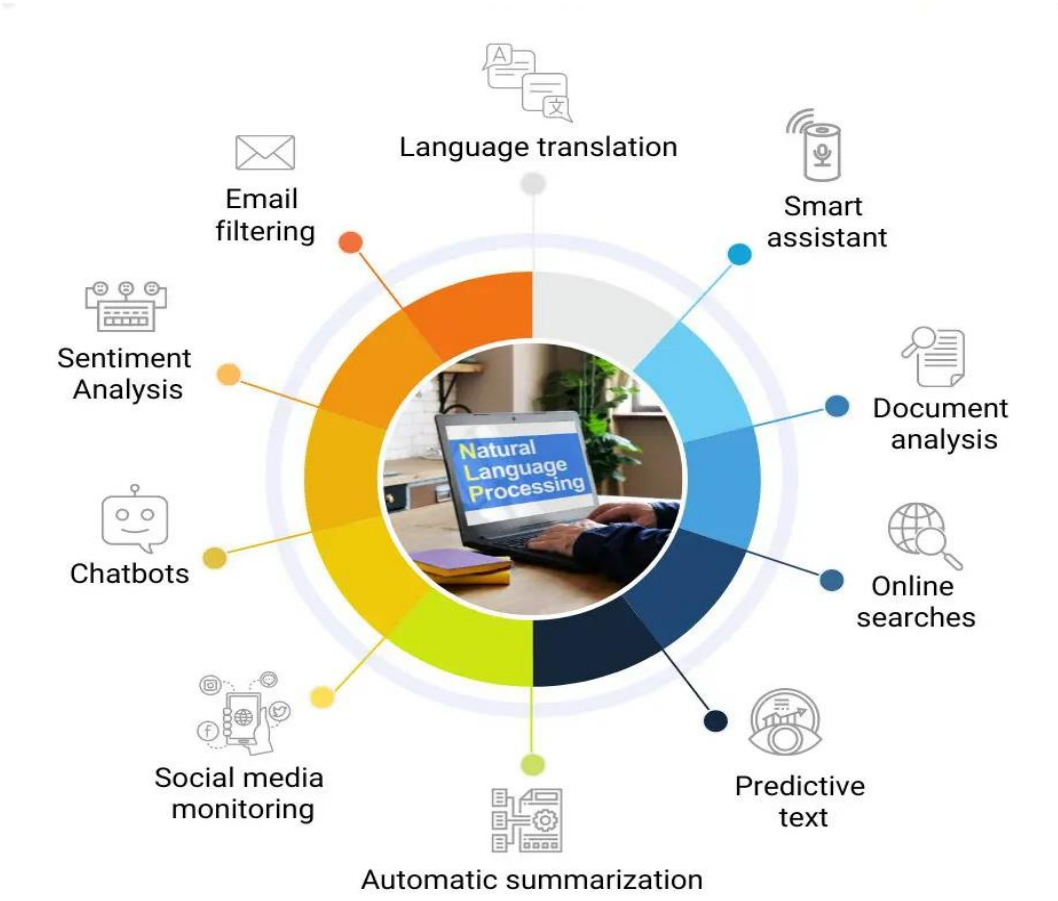

Figure.1 AI-based Natural Language Processing (NLP) Systems

Natural language processing may also find applications in other domains, such as teaching, business, and medicine. Natural language processing (NLP) has the potential to be used in the medical industry to extract useful information from patient records and clinical notes, allowing doctors to make better diagnosis and treatment decisions [6]. Natural language processing (NLP) might be used in the financial sector to analyse the tone of online discussions and media reports in order to better predict market movements and guide investment decisions. Natural language processing (NLP) has several applications in the field of education, one of which is the development of smart tutoring systems that can provide students with personalised feedback based on their linguistic abilities [7].

Although natural language processing (NLP) has come a long way, there are still open questions that require answering [8]. Some

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

examples of such difficulties are data bias and quality, interpretability and explainability, domain-specific language, support for many languages, contextual awareness, and adversarial attacks. Solving these issues will need constant R&D in a wide range of areas, such as data collection and annotation, model explainability, and attack resilience. Furthermore, continual efforts are needed to assure that AI-based NLP systems are used in an ethical and responsible manner [9].

To sum up, NLP (natural language processing) has matured into a substantial field of study with several practical implications. Advances in machine learning methods have allowed NLP, or natural language processing, to make significant progress in recent years. These developments have allowed for the synthesis and interpretation of human language by machines. Despite the challenges, natural language processing (NLP) looks to have a promising future, and we may anticipate much more progress and innovation in this area in the years to come.

II. Literature Review

A software traceability mapping study was completed in 2014 [10]. Previous assessments and the quality of the evidence were given further consideration in the research. However, the study was conducted without using standard NLP methods like machine learning or semantic networks. Since the intricacy of their development and implementation rendered them outside the scope of the research, they were disregarded for this purpose [11]. Despite this, significant breakthroughs have been made in recent years in the field of NLP research, propelling a new generation of tools and applications tailored to the procedures of software engineering. Training word embeddings in the software engineering domain space, needs categorization using deep learning, and natural language text classification in software engineering text mining pipelines are all examples of applications. Just a few instances are shown above [12].

A more recent evaluation focused on how natural language processing (NLP) may be used to exploit the unstructured data found in software repositories. In order to complete the analysis, we looked at mining repositories as a whole, with a particular emphasis on their use in traceability projects [13]. This action was taken so that we could wrap up the review. It was mentioned that, before adding NLP applications to the SDLC, a study of how NLP is used (at each step) should be conducted. The potential benefits of this integration, known as multidisciplinary research, are highlighted, and they include a more rounded approach to Computer Science and Engineering, more opportunities for automation, and a nudge towards universal programmability, wherein programming can be done informally and in a natural language [14]. Given the current state of tools, it is crucial to have well-defined semantics for tracking dependencies across various systems in the context of artefact traceability. Given that the characteristics of trace data may vary by sector, by organization, or even by the details of a given project, this evaluation highlighted the need to define a taxonomy for trace links [14].

To examine the efforts in automatic traceability reconstruction using machine learning classifiers to discover tactic-related classes, one of the earliest comprehensive literature reviews of traceability techniques

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

was conducted [16]. These are the core classes that allowed the tactical design choices to be put into action. This was one of the earliest comprehensive literature evaluations of traceability methods, published despite the fact that machine learning (including NLP) has received enormous attention just in the past several years. This article [17] extends the investigation on NLP application for traceability that was conducted in 2014 to include more current work. By making advantage of the inherent natural language semantics in the studied artifacts, our research aims to learn more about modern applications of NLP [18]. This is a perennial focus of research in the field of information retrieval, especially in light of recent developments in computing power and the introduction of enormous amounts of linguistic data. This is a major research topic in the field of information retrieval [19]. There is an urgent need to centralise and study these scattered efforts across various global platforms in order to analyse trends in the techniques and tools employed, trends in traceability across the various stages of the software development life cycle (SDLC), and open challenges pertaining to the application of NLP for traceability [20]. This direction will be the focus of our future contributions; it will serve as a point of reference from which we can assess the success of our continuing work and make adjustments as necessary [21].

Research	**Main Focus**	**Techniques and Methods**	**Applications**	**Challenges and Future Directions**
"Foundations of Statistical Natural Language Processing" (1999)	Statistical NLP	Hidden Markov models, probabilistic context-free grammars, maximum entropy models	Language modeling, parsing, information retrieval, machine translation	Dealing with rare events, modeling long-range dependencies, integrating multiple levels of linguistic knowledge
"Speech and Language Processing" (2008)	NLP and speech processing	Text classification, named entity recognition, sentiment analysis, speech recognition, synthesis	Information retrieval, machine translation, spoken dialogue systems, voice-enabled interfaces	Dealing with noise, variability, and ambiguity, modeling discourse and context, improving performance on low-resource languages

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

"Efficient Estimation of Word Representations in Vector Space" (2013)	Word embeddings	Word2vec, skip-gram, continuous bag-of-words	Language modeling, sentiment analysis, named entity recognition	Handling out-of-vocabulary words, improving quality and interpretability of word embeddings, incorporating contextual information
"Deep Learning for Natural Language Processing: Theory and Practice" (2017)	Deep learning for NLP	Convolutional neural networks, recurrent neural networks, attention mechanisms	Text classification, sentiment analysis, machine translation, question answering	Dealing with data sparsity, overfitting, and bias, improving interpretability and explainability of models, developing more efficient and scalable methods
"Text Mining: Concepts, Implementation, and Big Data Challenge" (2015)	Text mining	Text preprocessing, feature selection, clustering, classification	Opinion mining, sentiment analysis, text summarization, entity recognition	Handling noisy and unstructured data, dealing with language variation and ambiguity, scaling up to big data
"Natural Language Processing: State of The Art, Current Trends and Challenges" (2018)	NLP overview	Language modeling, parsing, named entity recognition, sentiment analysis	Information retrieval, machine translation, dialogue systems, question answering	Dealing with ambiguity, language variation, and context, improving the quality and diversity of training data, developing more robust and flexible models
"A Survey of Deep Learning Techniques for Natural Language Processing"	Deep learning for NLP	Recurrent neural networks, convolutional neural networks, autoencoders	Text classification, sentiment analysis, machine translation,	Handling long-range dependencies and complex structures, developing more efficient and scalable methods, integrating multiple

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

(2017)			language modeling	modalities and sources of information
"A Survey of Text Mining Techniques and Applications" (2018)	Text mining	Text classification, sentiment analysis, opinion mining, information extraction	Marketing research, social media analysis, customer feedback analysis, competitive intelligence	Handling unstructured and diverse data, dealing with language variation and ambiguity, improving the interpretability and explainability of models
"Natural Language Processing with Python" (2009)	NLP with Python	Text classification, information extraction, sentiment analysis	Named entity recognition, machine translation, speech recognition	Developing practical and scalable methods, dealing with language variation and ambiguity, integrating multiple modalities and sources of information
"A Survey of Machine Translation: Its History, Current Trends and Future Directions" (2018)	Machine translation	Rule-based, statistical, and neural machine translation	Human translation assistance, cross-lingual information retrieval, international communication	Handling language and domain adaptation, improving the quality and fluency of translations, developing more efficient and scalable methods

Table.1 Literature Review

III. Challenges

The Bias and Quality of Data In order to train accurate NLP models, a large amount of high-quality data is required. However, this data is often biassed or incomplete. This can lead to inaccurate or unfair predictions and outcomes.

- The capacity of AI-powered NLP systems to be understood and explained: The intricacy and opacity of many of these systems makes it challenging to understand how they reach their findings or assessments. As a result of their opaque nature, these systems may be hard to trust or audit.
- Natural language processing models trained on general language data may have trouble distinguishing jargon or other kinds of domain-specific language, which can lead to inaccurate results.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

- Multiple language support Many NLP systems provide this capability, however there are still challenges in processing and understanding languages with complex grammar or syntax.
- Inaccurate predictions or misinterpretations might emerge from NLP models' inability to fully understand the context in which a given piece of language is utilised.
- adversarial attacks Adversaries may be able to fool Natural Language Processing algorithms into giving inaccurate predictions by manipulating the data used to train the model.

Solving these issues will need constant R&D in a wide range of areas, such as data collection and annotation, model explainability, and attack resilience. Furthermore, continual efforts are needed to assure that AI-based NLP systems are used in an ethical and responsible manner.

IV. Datasets

Datasets of General Language: The text in these datasets can be used to train models for various NLP tasks requiring the usage of natural language processing. News stories, books, and social media posts are all examples of broad language text. Here are a few widespread ones:

- The term "Common Crawl Corpus" describes this collection.
- Wikipedia's body of work.
- The homepage of BookCorpus.

Collections of Texts for Emotion Analysis: These datasets may be used to train models for sentiment analysis, as they contain text that has been annotated with a sentiment such as positive or negative. Here are a few widespread ones:

- The IMDB Movie Review Dataset
- Collection of Yelp Ratings and Comments
- The Amazon.com review dataset

Datasets for Named Entity Recognition are used to train models for named entity recognition, and contain text annotated with names of people, places, and other entities. Named entity recognition dataset based on the CoNLL-2003 registry is one such well-known example.

- OntoNotes's Named Entity Recognition Dataset
- Dataset for Recognizing Named Entities Created by AC Staff

These datasets can be used to train machine translation models, as they include parallel text in two or more languages. Here are a few widespread ones:

- WMT Machine Translation's Dataset

Machine Translation Dataset Provided by the International Workshop on Spoken-to-Speech Translation (IWSLT OPUS) Dataset

Question-Responding Datasets: These question-and-answer datasets may be used to teach machines to provide accurate replies to queries. Here are a few widespread ones:

- The Stanford Question Answering Dataset (SQuAD) is the dataset used by TriviaQA's system.
- MS MARCO's Data Set

These are only a handful of the numerous public datasets available for use with NLP systems based on artificial intelligence. The

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431
right dataset must be selected for the task at hand and the research field being tackled. Data quality and bias should also be taken into account before reaching a final conclusion.

V. Conclusion

In conclusion, computers now have the ability to understand and generate human language thanks to artificial intelligence-based natural language processing (NLP) technologies, which have revolutionised the way people engage with technology. Large datasets, robust computing resources, and cutting-edge methods like neural networks have all contributed to rapid advancements in the field of natural language processing (NLP) in recent years. Natural language processing has many potential uses in fields as diverse as healthcare, finance, education, and customer service. Nonetheless, there are still problems that need to be fixed, such as poor data quality and bias, insufficient interpretability and explainability, a lack of language support, difficulty understanding context, and vulnerability to adversarial attacks. Solving these issues will need constant R&D in a wide range of areas, such as data collection and annotation, model explainability, and attack resilience. Furthermore, continual efforts are needed to assure that AI-based NLP systems are used in an ethical and responsible manner. Although these challenges exist, the future of natural language processing (NLP) remains promising. The field has the potential to change a broad number of areas and improve the quality of life for people all over the world, so we should expect to witness further development and innovation in the coming years.

VI. Future Work

Natural language processing (NLP) systems are a rapidly developing area of artificial intelligence with several promising research directions. Here are a few points to think about:

The state-of-the-art NLP models now available mostly focus on textual inputs; hence, multimodal NLP has lagged behind. Multimodal natural language processing (NLP) models are in great demand, but this need is only expected to grow. Scientists might look at developing NLP models with multimodal support for natural language processing. These models may be used in a wide range of contexts, from bettering image and video captioning to enhancing chatbots and voice assistants.

Language Processing that can Be Explained Despite the fact that NLP systems powered by AI have shown to be highly effective in a wide range of applications, they are not without their detractors, who point to the fact that understanding how the system is making predictions can be a challenge. This is something that Explainable NLP hopes to alter. Researchers can zero down on building NLP models with transparent decision-making processes and explainability. This can boost the models' credibility and win over the confidence of the people who will be using them.

The current crop of NLP models is frequently trained on large, general-purpose datasets that may not be suitable for more niche applications. If researchers are interested in improving the accuracy and effectiveness of NLP models in a particular domain, like healthcare, finance, or law, they can focus on building domain-specific models.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

Recent years have seen great advancements in natural language processing (NLP) models, although most of the focus has been on high-resource languages like English and Chinese. Building natural language processing (NLP) models for low-resource languages can have a significant impact on the information access and social functioning of individuals who speak those languages.

There is a growing need to ensure the resilience and security of NLP systems as natural language processing (NLP) models become more widely employed in practical applications. Researchers may decide to focus on building NLP models that are immune to common adversarial attacks, such as those that introduce noise or modify the input to manipulate the output. Also, scientists may look into the feasibility of creating secure NLP systems, which can protect sensitive information and prevent unauthorised access.

References

[1] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13(3), 55-75.

[2] Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

[3] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), 2493-2537.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[5] Zhang, Y., Wallace, B., & Barzilay, R. (2017). Designing deep learning models for summarization. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2045-2055).

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[8] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[9] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 1-5).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[11] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

language understanding by generative pre-training.

[12] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 2227-2237).

[13] Goldberg, Y., & Elhadad, M. (2017). Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies, 10(1), 1-309.

[14] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (pp. 1631-1642).

[15] Chollet, F. (2018). Deep learning with Python. Manning Publications.

[16] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[17] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[18] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751).

[19] Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5754-5764).

[20] Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 1243-1252). JMLR. org.

[21] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[23] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 1480-1489).

[24] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

[25] Johnson, R., & Zhang, T. (2015). Semi-supervised convolutional neural networks for text categorization via region

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 48-58
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1431

embedding. In Advances in neural information processing systems (pp. 919-927).

[26] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[28] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.