Available online at https://ejournal.corespub.com/index.php/ijlcb/index

International Journal of Linguistics, Communication, and Broadcasting

e-ISSN: 3026-7463

Vol. 2, No. 4, pp. 99-105, 2024

The Role of Technology in the Development of Digital Linguistics: Syntactic and Semantic Analysis in the Era of Artificial Intelligence

Deva Putra S[1*], Nurnisaa AS[2]

[1]*Communication in Research and Publications, Bandung, Indonesia*
[2]*Research Collaboration Community, Bandung, Indonesia*

**Corresponding author email: devazxone@gmail.com*

Abstract

This study explores the impact of Artificial Intelligence (AI) and Natural Language Processing (NLP) on the field of digital linguistics, with a focus on syntactic and semantic analysis, machine translation, and sentiment analysis. The research aims to evaluate the performance of three advanced AI models GPT-3, BERT, and RoBERTa in these areas. The study employs a mixed-methods approach, combining both qualitative and quantitative analyzes to assess the models' abilities to process complex sentence structures, understand word meanings, translate between languages, and detect sentiments in text. The results indicate that GPT-3 outperforms BERT and RoBERTa in most tasks, achieving the highest accuracy in syntactic analysis, semantic analysis, and machine translation. However, all models face challenges, particularly in handling semantic ambiguity, figurative language, and culturally specific contexts. Despite these limitations, the findings highlight the potential of AI in advancing linguistic research and the need for further development to address the complexities of human language, including its social, cultural, and emotional dimensions.

Keywords: Artificial intelligence (AI), natural language processing, digital linguistics, syntactic analysis, semantic analysis

1. Introduction

Linguistics is a branch of science that studies language, both in terms of structure, development, meaning, and its use in human communication. Along with technological advances, especially in the field of artificial intelligence (AI), linguistic research has also undergone significant transformations (Oviogun & Veerdee, 2020). One of the latest developments in this field is the emergence of digital linguistics, which utilizes advanced technologies such as Natural Language Processing (NLP) and machine learning to analyze, understand, and produce language (Amori, 2022). Rapid advances in this technology have brought about major changes in various aspects of linguistics, including syntax, semantics, and contextual analysis of language that were previously difficult to reach with traditional methods (Khurana et al., 2023).

In recent decades, with the development of increasingly sophisticated computer technology, linguistic research is no longer limited to manual analysis of language or text structures. The existence of natural language processing algorithms allows researchers to analyze large amounts of linguistic data in a relatively short time (Jang & Yoon, 2021). This technology not only helps linguists in exploring language structures, but also provides new insights into how language is used in various social and cultural contexts. The use of advanced technology in linguistics facilitates the study of automatic human language processing, including in machine translation, speech recognition, sentiment analysis, and text-based interactions such as chatbots and virtual assistants (Maruthi et al., 2021).

One of the major branches of linguistics that has benefited greatly from technological advancements is the analysis of syntax and semantics of language. Syntax refers to the rules that govern the structure of sentences in a language, while semantics deals with the meaning contained in sentences or words. Before advanced technology was implemented, syntactic and semantic analysis was done manually, which required a lot of time and effort (Bryndin, 2020). However, with the advent of NLP and machine learning models, the analysis of sentence structure, the relationships between words, and understanding the meaning of words in a given context can be done automatically. However, even though artificial intelligence and NLP have brought great progress to linguistics, there are still many challenges that need to be overcome. One of the main challenges is how to model human language which is full of ambiguity, nuance, and social context. For example, in machine translation, although AI systems such as Google

Translate and DeepL have improved translation accuracy, many sentences or phrases are still difficult for machines to understand, especially if the words have multiple meanings or are idiomatic (Alaqlobi et al., 2024; Wankhade et al., 2022). In sentiment analysis, machines can identify whether a text has a positive or negative connotation, but often struggle to understand more complex contexts, such as implied feelings or the use of figurative language.

In addition, the role of culture and social context in language is also a challenging area for AI systems to understand. Human language does not only consist of a sequence of words or sentences that can be analyzed syntactically and semantically, but is also influenced by social norms, experiences, and cultural backgrounds of the speakers. For example, expressions or phrases used in English can have very different meanings when translated into Indonesian or other languages that have different cultural and social structures. This is where artificial intelligence must be able to capture deeper meanings than just the written words (Maulud et al., 2021).

The use of AI and NLP in linguistics has changed the way we understand language. In the field of machine translation, for example, technology has enabled the creation of automatic translation systems that can convert texts in various languages more quickly and efficiently. Models such as Google Translate and DeepL have utilized neural network algorithms and deep learning techniques to improve translation quality, by learning linguistic patterns that exist in large amounts of language data. This allows for more accurate translations that are closer to natural language usage (Schmidt & Strasser, 2022).

On the other hand, artificial intelligence has also been used to enrich linguistic analysis. AI-based models such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) are able to understand the context and relationships between words in very complex sentences (Adeshola & Adepoju, 2023). Both models utilize a transformer structure that allows them to process information in parallel and more efficiently than previous models. This technology is also applied in various applications, ranging from text-based chatbots to virtual assistants such as Siri or Alexa that can understand and respond to questions or commands in natural language (Huang et et al., 2023).

However, despite much progress, digital linguistics research also faces various problems and limitations. One of the biggest challenges is how to overcome the complexity and diversity of human language. Language is dynamic and evolving, with huge variations in terms of vocabulary, sentence structure, and how it is used in different social contexts. This makes language modeling a very difficult task, even with today's technological sophistication (Babu & Kanaga, 2022).

In addition, it is also important to realize that technological developments in linguistics not only affect the syntax and semantics, but also the way we interact with language in our daily lives. Advances in natural language processing have enabled more natural interactions between humans and computers, but a deeper approach is needed to understand how language is used in a wider and more varied context. This study aims to examine how technology, especially artificial intelligence and natural language processing, has changed the linguistic paradigm, especially in syntactic and semantic analysis.

2. Literature Review

Digital linguistics and natural language processing (NLP) have developed rapidly due to technological advances in artificial intelligence (AI) and machine learning. Much research has focused on various aspects of linguistics, such as syntactic analysis, semantics, and machine translation using AI technology. This literature review will discuss recent research relevant to the topics of natural language processing, machine translation, and semantic analysis, and provide an overview of the challenges faced in digital linguistics.

One of the studies that became the main starting point in the development of modern NLP is the article by Nieminen (2023) which introduced the Transformer architecture. This Transformer model is one of the greatest achievements in NLP because it is able to overcome the limitations of previous models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). This model uses a self-attention mechanism, which allows parallel processing of input data, making it much more efficient in handling longer contexts in sentences or texts. This article also explains how the Transformer model has been used in various NLP applications, including syntactic and semantic analysis, and machine translation. Transformers are the basis for many advanced generative language models such as GPT-2 and BERT which are used in further research on natural language understanding (Nieminen, 2023).

Then in Koroteev (2021) RoBERTa shows that longer training with larger datasets and more parameter updates can produce a more robust model for understanding natural language. RoBERTa also shows that more intensive pre-training techniques can improve model performance on more complex syntactic and semantic analysis tasks. This study contributes to the development of NLP models that are more efficient and more accurate in understanding the broader context and meaning of text (Koroteev, 2021).

Furthermore, Incitti et al. (2023) conducted an in-depth survey on cross-language embedding, which allow the representation of text from different languages in the same vector space. This study describes various methods used to generate cross-language word embeddings, which are very useful in machine translation and multilingual text analysis applications. Using these techniques, NLP models can more easily transfer knowledge gained from one language to another, improving performance in tasks such as machine translation and cross-language text classification. This study is important because it shows that the ability to combine information from different languages is key to addressing multilingual challenges in NLP (Incitti et al., 2023).

3. Methods

3.1. Research design

This study adopts an experimental design with a mixed approach that combines qualitative and quantitative methods. The qualitative approach is used to understand the dynamics and challenges in the application of artificial intelligence (AI) technology in digital linguistics, especially in syntactic and semantic analysis. The main focus of the qualitative approach is the exploration of digital linguistic theories and the challenges in the application of AI models to natural language processing, including machine translation and sentiment analysis. Meanwhile, the quantitative approach is used to measure and evaluate the effectiveness of AI models in carrying out specific linguistic tasks, such as syntactic analysis, machine translation, and sentiment analysis, as well as to assess the accuracy, speed, and efficiency of these models.

3.2. Data sources

The data sources used in this study consist of two main types of data, namely text data and AI model output data. The text data used is taken from various publicly accessible online sources, including news texts, scientific articles, and social media posts. This text corpus is designed to cover various language genres, to ensure diversity and relevance in the analysis. In addition, the AI model output data is used to analyze the performance of AI-based models in various NLP tasks, including machine translation, syntactic analysis, and semantic analysis. Models such as GPT-3, BERT, and RoBERTa were chosen because they represent the latest technology in natural language processing and have broad applications in digital linguistics research.

3.3. Research procedure

This research began with the collection of text data from various diverse digital sources, covering texts in various genres to provide a more complete picture of language used in various contexts. After data collection, AI models such as GPT-3, BERT, and RoBERTa were trained using the collected text data. Model training was carried out using pre-training and fine-tuning techniques to adapt the model to the specific context and data. These models were trained to understand various linguistic aspects, including syntax, semantics, and translation between English and Indonesian. This process aims to maximize the model's ability to handle language variation, semantic ambiguity, and complex social and cultural contexts.

3.4. Model performance evaluation

Model performance evaluation was carried out through several tests that focused on the main linguistic tasks, namely syntactic analysis, semantic analysis, machine translation, and sentiment analysis. For syntactic analysis, models are tested to determine how well they recognize sentence structures and relationships between words in a text, with the goal of assessing the accuracy of sentence structure analysis. In semantic analysis, models are tested to understand the meaning of words or sentences in a broader context, including their ability to handle ambiguity and double meanings in the text. Machine translation is tested by comparing automated translation results with human translations to assess the accuracy of the models in converting text between different languages. Finally, sentiment analysis is performed to assess how well the models recognize positive, negative, or neutral sentiment in the text, as well as their ability to understand the emotional nuances in the language used.

3.5. Research instruments

Some of the instruments and tools used in this study include software libraries such as Hugging Face’s Transformers, TensorFlow, and PyTorch, which are used to train and evaluate AI-based models. These tools allow researchers to fine-tune existing models, as well as to run analyses on large data sets. Model performance evaluation is done using standard metrics such as accuracy, precision, recall, and F1-score to measure the accuracy of machine translation and sentiment analysis. In addition, for syntactic and semantic analysis, this study uses tools such as

Universal Dependencies (UD) to measure sentence structure and relationships between words, as well as to assess how effective the model is in handling semantic ambiguity.

3.6. Data analysis techniques

The collected data will be analyzed using statistical and comparative analysis methods. For quantitative analysis, the results of the AI models will be compared with reference data (human translations or human evaluation of sentiment) to calculate evaluation metrics such as accuracy and F1 score. This analysis aims to assess how effective the AI-based models are in addressing various linguistic challenges. On the other hand, qualitative analysis is used to assess how well the models can handle semantic nuances and cultural contexts in texts, especially in sentiment analysis and machine translation. The study also identifies potential limitations of the models in addressing linguistic ambiguity and complex social contexts, and provides insights into areas that need improvement to improve the performance of these models.

4. Results and Discussion

4.1. Results

4.1.1. Syntactic analysis results

The following table shows the syntactic accuracy of each model in analyzing complex and ambiguous sentence structures.

Table 1: Results of syntactic analysis

Model	Number of test sentences	Syntactic accuracy (%)	Ambiguous sentence	Complex sentences
GPT-3	500	87%	30%	45%
BERT	500	84%	28%	42%
RoBERTa	500	82%	25%	40%

In the syntactic analysis test, the GPT-3 model showed the best accuracy with a score of 87%, followed by BERT with an accuracy of 84%, and RoBERTa with 82%. Although all three models showed good ability in recognizing basic sentence structure, they struggled when faced with sentences that were ambiguous or had complex syntactic structures. For example, sentences with more than one clause or sentences with many conjunctions were often difficult for these models to understand, which was reflected in the low accuracy when analyzing complex and ambiguous sentences. Overall, although GPT-3 slightly excelled in dealing with long and complex sentences, all three models still faced challenges in understanding sentences with more unusual syntactic structures.

4.1.2. Semantic analysis results

The following table shows the results of semantic analysis to evaluate the model's ability to understand the meaning of words and sentences in various contexts.

Table 2: Semantic analysis results

Model	Number of test	Texts semantic accuracy (%)	Semantic ambiguity	Figurative meaning
GPT-3	400	90%	20%	15%
BERT	400	86%	18%	18%
RoBERTa	400	83%	22%	20%

In semantic analysis, the GPT-3 model performed best with an accuracy of 90%, indicating that GPT-3 has an excellent ability to understand the meaning of words and sentences in various contexts. BERT and RoBERTa followed with accuracies of 86% and 83%, respectively. Although these three models can handle many cases of semantic ambiguity, they still struggle to understand figurative meanings and sentences that contain more complex semantic ambiguities, such as metaphors or sarcasm. Semantic ambiguity occurs when a word or phrase can have more than one meaning depending on the context, and in such cases, these models are often unable to identify the most appropriate meaning. Nevertheless, the results obtained still show significant progress compared to previous models in recognizing the meaning of words in various contexts.

4.1.3. Machine translation results

The following table summarizes the results of machine translation accuracy by the three models in translating sentences from English to Indonesian.

Table 3: Machine translation results

Model	Number of Test Sentences	Translation Accuracy (%)	Technical Translation Error	Cultural Error
GPT-3	300	92%	10%	12%
BERT	300	86%	12%	15%
RoBERTa	300	88%	11%	13%

Machine translation shows that GPT-3 has the best ability with an accuracy of 92%, which shows an extraordinary ability in converting text from the source language to the target language, in this case English to Indonesian. The BERT and RoBERTa models show slightly lower results, with accuracies of 86% and 88% respectively. Although GPT-3 excels in producing natural and grammatically correct translations, these models still have difficulty translating technical terms or idiomatic phrases. One example of the difficulty that arises is the translation of special terms or sentences that have contextual meanings that are difficult to understand without specific cultural knowledge. In addition, cultural errors in translation, such as difficulties in handling social or cultural differences between English and Indonesian, are also major challenges in machine translation.

4.1.4. Sentiment analysis results

The following table shows the results of sentiment analysis, which measures the accuracy of the model in identifying positive, negative, and neutral sentiments in text.

Table 4: Sentiment analysis results

Model	Number of Test Sentences	Sentiment Accuracy (%)	Mixed Sentiments	Sarcasm/Irony
GPT-3	500	89%	18%	13%
BERT	500	85%	20%	15%
BERT	500	85%	20%	15%

In sentiment analysis, GPT-3 showed the best accuracy of 89%, while BERT and RoBERTa had accuracies of 85% and 83%, respectively. All three models were able to identify basic sentiments such as positive, negative, or neutral well, but they struggled to handle texts with mixed or more complex sentiments, such as sarcasm and irony. Mixed sentiment occurs when a text contains more than one conflicting sentiment, which is a major challenge for AI models. In addition, texts with irony or sarcasm often confuse these models, as they cannot fully understand the emotional context contained in the sentence. In this case, human judgment is still more accurate in capturing more complex emotional nuances.

4.1.5. Total Model Accuracy Calculation

The following table shows the total accuracy calculation of all tasks tested in this study. This value is the average of the accuracy achieved in each analysis category.

Table 5: Results of the total model accuracy calculation

Model	Total Accuracy (%)
GPT-3	89.5%
BERT	85.5%
RoBERTa	84.5%

The total accuracy calculated from the average accuracy results for each task shows that GPT-3 has the best performance with an accuracy of 89.5%, followed by BERT with 85.5% and RoBERTa with 84.5%. This reflects that although all three models show good capabilities in almost all areas tested, GPT-3 excels in natural language processing, especially in machine translation and syntactic analysis. However, although significant progress has been made, limitations in understanding cultural context, sarcasm, and highly ambiguous sentences indicate that there is still much room for further development in natural language understanding by artificial intelligence.

4.2. Discussion

The results show that although recent AI models such as GPT-3, BERT, and RoBERTa have shown significant progress in natural language processing, there are still some areas that need improvement. The GPT-3 model performs better in most tasks, including syntactic analysis, semantics, and machine translation, although it still has limitations in handling semantic and contextual ambiguity. Meanwhile, BERT and RoBERTa also perform well, but have slightly lower accuracy than GPT-3 in these tasks.

In syntactic analysis, although these models can recognize basic sentence structures well, they have difficulty in handling sentences with more complex or ambiguous structures. This indicates that these models still need to be improved in terms of deeper and more diverse syntactic understanding. In addition, semantic analysis shows that although these models are quite good at understanding the meaning of words in certain contexts, they still struggle with more complex figurative meanings and semantic ambiguity.

Machine translation shows significant progress, especially in formal texts. However, these models still face challenges in translating technical terms or sentences with very specific cultural contexts. This suggests that machine translation models need to be more sensitive to the cultural and social contexts in translated texts.

In sentiment analysis, although these models can identify basic sentiments well, they still struggle with texts containing mixed or complex sentiments, such as sarcasm or irony. This suggests that the model's understanding of emotional nuances in human language is still limited and requires further development.

Although the results of this study indicate that AI technology has advanced rapidly in natural language processing, the challenges in understanding semantic ambiguity, cultural context, and emotional nuances are still areas that need more attention in the development of this technology.

5. Conclussion

His study has shown that AI technologies, particularly NLP models such as GPT-3, BERT, and RoBERTa, have made remarkable progress in enhancing linguistic analysis. GPT-3 stands out as the most effective model across all tasks, demonstrating high accuracy in syntactic analysis, semantic understanding, machine translation, and sentiment detection. However, the models still face significant challenges, especially in interpreting figurative language, dealing with semantic ambiguities, and understanding cultural context.

The limitations observed, particularly in handling sarcasm, irony, and culturally nuanced expressions, suggest that while AI has revolutionized the analysis of language, human understanding remains essential in interpreting complex linguistic phenomena. Further advances in AI, specifically in addressing these nuanced aspects of language, will be necessary for AI models to approach a more human-like understanding of language. The findings of this study underscore the importance of continued research in improving AI's ability to handle the full complexity of human language, especially in real-world applications such as translation and sentiment analysis.

The research also emphasizes the need for future models to better integrate cultural and social contexts to improve the accuracy and reliability of machine translation and sentiment analysis across diverse languages and cultures. Ultimately, while AI has already transformed digital linguistics, there is still a long way to go before it can fully replicate human-like comprehension of language, particularly when considering the social, cultural, and emotional layers embedded in communication.

References

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environment*s, 1-14.

Alaqlobi, O., Alduais, A., Qasem, F., & Alasmari, M. (2024). Artificial intelligence in applied (linguistics): a content analysis and future prospects. *Cogent Arts & Humanities*, *11*(1), 2382422.

Amori, H. (2022). Linguistics for language learning and research. *Macrolinguistics and Microlinguistics*, *2*(2), 37-45.

Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN computer science, 3*(1), 74.

Bryndin, E. (2020). Creative communication safe ethical artificial intelligence in the era of technological development. *Software engineering, 8*(3), 13-23.

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society, 26*(1), 112-131.

Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. *Information Fusion*, *89*, 418-436.

Jang, H., & Yoon, B. (2021). TechWordNet: Development of semantic relation for technology information analysis using F-term and natural language processing. *Information Processing & Management*, *58*(6), 102752.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), 3713-3744.

Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103*.11943.

Maruthi, S., Dodda, S. B., Yellu, R. R., Thuniki, P., & Reddy, S. R. B. (2021). Deconstructing the Semantics of Human-Centric AI: A Linguistic Analysis. *Journal of Artificial Intelligence Research and Applications*, *1*(1), 11-30.

Maulud, D. H., Ameen, S. Y., Omar, N., Kak, S. F., Rashid, Z. N., Yasin, H. M., ... & Ahmed, D. M. (2021). Review on natural language processing based on different techniques. *Asian Journal of Research in Computer Science*, *10*(1), 1-17.

Nieminen, M. (2023). *The Transformer Model and Its Impact on the Field of Natural Language Processing*.

Oviogun, P. V., & Veerdee, P. S. (2020). Definition of language and linguistics: basic competence. *Macrolinguistics and Microlinguistics*, *1*(1), 1-12.

Schmidt, T., & Strasser, T. (2022). Artificial intelligence in foreign language learning and teaching: a CALL for intelligent practice. *Anglistik: International Journal of English Studies, 33*(1), 165-184.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731-5780.