Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Use of Natural Language Processing in Social Media Text Analysis

Badry Ali Mustofa[1*], Wawan Laksito Yuly Saptomo[2]

STMIK Sinar Nusantara
badryalimustofa88@gmail.com [1], wlaksito@gmail.com [2]*

Abstract

Social media generates enormous volumes of text data, creating both opportunities and challenges for analysis. Natural Language Processing (NLP) enables in-depth analysis of public opinion, identification of trends and language patterns from social media texts. However, texts from social media often face problems with informal language, slang, and spelling errors. This research discusses the application of NLP techniques, such as sentiment analysis, tokenization, and text classification, and compares classical machine learning models (Naive Bayes and SVM) with deep learning models (BERT). Results show deep learning-based models excel at understanding informal language contexts, producing more accurate analysis. This study makes an important contribution in the development of AI-based applications for social media analysis.

Keywords: *Natural Language Processing, social media, sentiment analysis, deep learning, text classification.*

1. Introduction

Social media has developed into one of the most significant communication platforms in the digital era. With billions of active users worldwide, platforms like Twitter, Facebook, Instagram, and Reddit have become incredibly rich and dynamic sources of data. Social media users produce various types of content, ranging from personal opinions, product reviews, to discussions of global issues. This creates a great opportunity for researchers to understand communication patterns, user behavior, and social trends. However, data from social media has unique characteristics, such as the use of informal language, slang, emoticons, abbreviations, and grammatical errors, which make it a challenge in text analysis [1], [2].

Natural Language Processing (NLP), as a branch of artificial intelligence, offers solutions to overcome these challenges. NLP allows computers to understand, analyze and generate text in natural human language. In the context of social media, NLP has been used for a variety of applications, including sentiment analysis, topic identification, text classification, and named entity detection. One example of its application is sentiment analysis, where user text is processed to identify the emotions or opinions contained in it, such as positive, negative, or neutral. This capability is an important tool in the fields of marketing, politics, and customer service, where understanding public opinion can provide a significant competitive advantage [3], [4].

However, although NLP has shown great potential, its application in social media text analysis faces various technical challenges. Text on social media tends to be unstructured and varies widely in writing style, so traditional algorithms often struggle to provide accurate results [4], [5].
Additionally, social media data is often very large and requires a scalable approach to data processing. In recent years, deep learning-based models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated superiority in understanding more complex and informal text contexts, making them increasingly popular tools for social media text analysis [6], [7]..
This research aims to explore the use of various NLP techniques in social media text analysis, including sentiment analysis, text classification, and named entity recognition. This study will compare the performance of classical machine learning models, such as Naive Bayes and Support Vector Machines (SVM), with deep learning-based models such as BERT in processing social media text data. Apart from that, this research also discusses the main challenges faced in processing social media texts, such as the diversity of language styles, grammatical errors, and the need to process data on a large scale [8], [9].

Through this research, it is hoped that a deeper understanding of the potential and limitations of NLP in social media text analysis can be obtained. It is also hoped that the results of this research can provide practical guidance for developing artificial intelligence-based applications that are more effective in analyzing public opinion, identifying trends, and monitoring user behavior patterns on social media [10].

Research Problem: Social media texts are usually unstructured, full of slang, informal language, and grammatical errors, which makes them challenging to analyze. Therefore, Natural Language Processing (NLP) techniques are important to extract meaningful information from social media texts.
Research purposes: The aim of this research is to explore the application of NLP techniques for social media text analysis, with a focus on comparing the performance of classical machine learning and deep learning models in overcoming informal language challenges.

2. Literature Review

Natural Language Processing (NLP): An overview of the basic concepts of NLP, including the processes of tokenization, stemming, lemmatization, and named entity recognition (NER). NLP has developed as an important tool for understanding naturally generated text.
Social Media Text Analysis: An explanation of how text from social media can be analyzed for business, social, and academic purposes. This includes the use of sentiment analysis to understand how users feel, as well as topic recognition to identify conversation trends.
Machine Learning vs Deep Learning Models in NLP: Comparison between classical models such as Naive Bayes and Support Vector Machines (SVM) with modern deep learning models such as BERT (Bidirectional Encoder Representations from Transformers). Previous research shows that deep learning models are more effective for handling complex informal language.

3. Research Methodoloy

3.1. Research Design

This research is quantitative with an experimental approach. Data is collected from several social media platforms and analyzed using NLP techniques.

3.2. Data collection

Data Source
Data is taken from popular social media platforms such as:
1. Twitter: As a microblogging platform, Twitter provides rich short text data on opinions, conversations and reactions to various issues. Data is collected using the Twitter API (Twitter Developer Platform).
2. Facebook: Public comments and posts are used to understand user interactions in longer discussions.
3. Instagram: Comments and hashtags (#hashtags) explained to understand visual and text trends.
4. Reddit: A community-based discussion platform that provides data on opinions on a variety of topics.

Data Collection Techniques
1. Web Scraping: This technique is used to collect data from comments or public posts available on social media (for example, using libraries such as Beautiful Soup or Scrapy for Python).
2. Social Media APIs: Most platforms, such as Twitter and Reddit, provide official APIs for more structured data collection and in accordance with platform policies.
3. Already Available Datasets: Several popular datasets are used, such as:
4. Sentiment140 (Twitter)
5. Kaggle's Social Media Datasets
6. Reddit Comment Corpus
7. Facebook Large Page-Post Dataset

Data Preprocessing
Data collected via scraping or APIs often requires preprocessing steps, such as:
1. Data Cleaning: Removes URLs, hashtags, emoticons and special symbols.
2. Tokenization: Breaking down text into individual words for further analysis.
3. Normalization: Changing text to a standard form, for example changing "u" to "you".
4. Duplication Removal: Eliminates identical text entries to avoid bias.

Data Collection Ethics; Data collection is carried out in accordance with the platform's privacy policy. The data accessed comes from posts or comments that have been set as "public" by the user. No personal information is collected or used without explicit permission.

Dataset Source
Here are some links to relevant datasets:
1. Sentiment140 Dataset: https://www.kaggle.com/kazanova/sentiment140
2. Reddit Comment Dataset: https://www.reddit.com/r/datasets
3. Twitter US Airline Sentiment Dataset: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

3.3. Analysis Process:

Analysis Process:
1. Data Preprocessing: Tokenization, stemming, lemmatization, and data cleaning.
2. NLP techniques: Sentiment analysis, named entity recognition (NER), and text classification.
3. Machine Learning Models: Naive Bayes, Support Vector Machines (SVM), and deep learning models such as BERT will be applied to the processed data.

3.4. Model Performance Evaluation

Models are evaluated based on metrics such as accuracy, precision, recall, and F1-score. A comparison was made between classical and deep learning models to determine the best model for handling social media text.

4. Result and Discussion

4.1. Sentiment Analysis Results

The social media text data analyzed includes 10,000 entries taken from platforms such as Twitter and Reddit. Sentiment analysis techniques are applied to identify user emotions in three main categories: positive, negative, and neutral. Models compared include:
Naive Bayes: 68% accuracy on raw data without deep preprocessing.
SVM: 75% accuracy with tokenization and stopwords removal.
BERT: Accuracy reached 89% after training with the same dataset, showing superiority in understanding complex contexts and informal language use.
The results show that deep learning-based models, such as BERT, are superior in recognizing emotional patterns in text that contains slang, abbreviations, or other non-standard forms. Classical models such as Naive Bayes tend to have difficulty processing informal data.

4.2. Model Performance in Text Classification

Text classification is done to separate data into categories, such as political topics, product reviews, and entertainment. The following methods are used:
TF-IDF with SVM: Accuracy reaches 77% after data normalization.
Word Embeddings with LSTM: Accuracy increases up to 85%, but training time is longer.
BERT: With fine-tuning, BERT achieved 92% accuracy, showing that the model is able to capture more complex relationships between words than traditional approaches.
BERT's superiority in text classification can be attributed to its transformer-based architecture, enabling deep contextual analysis.

4.3. Use of Named Entity Recognition (NER) Technique

Named Entity Recognition (NER) is used to extract the names of people, locations, organizations, and other entities from social media text. For example:
CRF (Conditional Random Field): 65% accuracy on text containing many grammatical errors.
spaCy Pre-trained Model: 78% accuracy, with best performance on text with a more formal structure.
BERT NER Model: Accuracy reaches 88%, even on text containing errors or abbreviations.
These results show that deep learning-based models have a better ability to recognize entity patterns even in text with imperfect structure.

4.4. Discussion and Interpretation of Results

Model Performance: Deep learning-based models, especially BERT, consistently outperform classical models. This is due to his ability to understand the relationships between words in full context.
Social Media Data Challenges:
Language Variations: Social media texts are often unstructured, using emoticons, slang, or abbreviations, which is a barrier to traditional models.
Data Volume: Social media generates large amounts of data that require scalable techniques for processing.
Efficiency and Resources: Deep learning-based models require more training time and computing resources than classical models, which is an important consideration for real-world implementation.

4.5. Implications and Relevance

This research underscores the importance of selecting an appropriate model for the task of social media text analysis. Deep learning models, although requiring greater resources, provide much more accurate and relevant results for understanding user behavior patterns on social media. This research also opens up opportunities for the development of practical applications, such as:

1. Real-time monitoring of public opinion.
2. Detection of social trends and crisis issues.
3. Sentiment analysis-based marketing personalization.

5. Conclusion

This research explores the application of Natural Language Processing (NLP) in social media text analysis for various purposes, such as sentiment analysis, text classification, and named entity recognition. The study results show that NLP has great potential in extracting insights from rich but complex social media text data. Classic machine learning-based techniques, such as Naive Bayes and Support Vector Machines (SVM), offer simple and fast approaches, but have limitations in understanding the context of informal language and slang often found on social media platforms.
In contrast, deep learning-based models, such as BERT, show significant advantages in understanding text with more complex contexts. The model is able to better capture emotional and linguistic nuances, resulting in superior performance in sentiment analysis and entity

recognition tasks. However, implementing deep learning-based techniques requires larger computing resources and longer training time, which is one of the main challenges in real-world scenarios.

The research also highlights other challenges associated with social media text analysis, such as stylistic variations, grammatical errors, and the need to handle data on a large scale. Nevertheless, advances in NLP techniques and the availability of more efficient pretrained models provide a great opportunity to overcome these obstacles.

Overall, this study makes an important contribution to understanding how NLP can be used to analyze social media texts and aid the development of artificial intelligence-based applications. It is hoped that the results of this research can provide guidance for researchers and practitioners in implementing NLP technology for social media analysis, both in the context of academic research and practical applications.

References

[1] J. Camacho-Collados *et al.*, "TweetNLP: Cutting-Edge Natural Language Processing for Social Media," *EMNLP 2022 - 2022 Conf. Empir. Methods Nat. Lang. Process. Proc. Demonstr. Sess.*, no. April 2022, pp. 38–49, 2022, doi: 10.18653/v1/2022.emnlp-demos.5.

[2] A. Sandu, L. A. Cotfas, A. Stănescu, and C. Delcea, *A Bibliometric Analysis of Text Mining: Exploring the Use of Natural Language Processing in Social Media Research*, vol. 14, no. 8. 2024. doi: 10.3390/app14083144.

[3] Y. A. Telaumbanua, A. Marpaung, C. Putri, D. Gulo, D. K. Wijaya, and U. Nias, "An Analysis of Two Translation Applications : Why is DeepL Translate more accurate than Google Translate ?," vol. 4, no. 1, 2024.

[4] T. Architecture, "TRANSLI : a Case Study for Social Media Analytics and Monitoring," 2018.

[5] A. Nur Oktavia, M. Iqbal, R. W. Saputra, M. I. Zulfikar, and A. Saifudin, "Implementasi Metode Natural Language Processing Dalam Studi Analisis," *J. Ilm. Ilmu Komput. dan Multimed.*, vol. 2, no. 1, pp. 154–159, 2024, [Online]. Available: https://jurnalmahasiswa.com/index.php/biikma

[6] I. Huda, "Implementasi Natural Language Processing (Nlp) Untuk Aplikasi Pencarian Lokasi," *J. Nas. Teknol. Terap.*, vol. 3, no. 2, p. 15, 2021, doi: 10.22146/jntt.35036.

[7] R. Khoirunisa, "Penggunaan Natural Language Processing Pada Chatbot Untuk Media Informasi Pertanian," 2020. doi: 10.20961/ijai.v4i2.38688.

[8] J. A. Putra and A. Budi, "Penerapan Natural Language Processing dalam Aplikasi Chatbot Sebagai Media Pencarian Informasi Dengan Menggunakan React (Studi Kasus: Institut Bisnis dan Informatika Kwik Kian Gie)," *J. Inform. dan Bisnis*, vol. 9, no. 2, pp. 1–12, 2020.

[9] A. Puspitasari, A. N. Paradhita, Y. W. Tineka, V. Sulistyowati, N. K. S. Noriska, and Haryanto, "Natural Language Processing (NLP) Technology for Chatbot Website," *J. Penelit. Pendidik. IPA*, vol. 10, no. SpecialIssue, pp. 319–324, 2024, doi: 10.29303/jppipa.v10ispecialissue.8241.

[10] D. Radhian, I. Afrianto, P. Studi, and T. I. Komputer, "Pembangunan Aplikasi Chatbot Sebagai Media Pencarian Informasi Dalam Bidang Peternakan," *Progr. Stud. Tek. Inform. Komput. Indones.*, 2019.