Esra Şekerci

2698215

COGS514 – Literature Review, due: 3 April 2024

The study conducted by Çabuk et al. (2024) aimed to comprehensively analyze linguistic features in Turkish-speaking individuals diagnosed with schizophrenia (SZ) compared to healthy controls (HC) using natural language processing (NLP) techniques. The researchers recruited a sample comprising 38 SZ patients and 38 well-matched HC individuals, meticulously documenting and analyzing their demographic and clinical characteristics. While no significant differences were found in gender distribution ($p = 0.81$) and age ($p = 0.65$), SZ patients displayed a higher prevalence of unmarried status ($p = 0.003$) and unemployment ($p = 0.004$) compared to the HC group. Furthermore, SZ patients exhibited significantly more severe thought and language disturbances, as indicated by lower scores on the Thought and Language Disorder Scale (TALD) compared to HC individuals ($p = 0.001$).

In the hypothesis-driven NLP analyses, significant differences emerged between SZ and HC speech patterns. SZ patients exhibited significantly shorter mean sentence lengths compared to HC individuals (4.681 vs. 6.571, $p = 0.001$), indicating reduced fluency. Additionally, SZ speech demonstrated lower lexical diversity, as evidenced by a lower moving average type-token ratio compared to HC speech (0.814 vs. 0.839, $p = 0.008$). Notably, SZ patients displayed higher usage of first-person singular pronouns (6 vs. 3, $p = 0.002$), suggesting a tendency towards self-referential language patterns. These findings underscored notable linguistic abnormalities in SZ speech, encompassing reduced fluency, lexical diversity, and distinctive patterns of self-reference.

Furthermore, correlations between TALD scores and linguistic variables revealed significant associations with symptom severity in SZ. Negative correlations were observed between TALD scores (Total, Objective Positive, Subjective Positive, Objective Negative, Subjective Negative) and measures of fluency, such as mean sentence length and moving average type-token ratio ($p < 0.05$). These findings highlight the clinical relevance of linguistic disturbances in schizophrenia, with more severe thought and language disturbances aligning with reduced fluency and lexical diversity.

In the exploratory NLP analyses, parts-of-speech tagging was employed to further characterize linguistic differences between SZ and HC speech. SZ patients exhibited significantly fewer coordinating conjunctions compared to HC individuals (7.17 vs. 9.14, $p = 0.007$), suggesting differences in sentence complexity. However, no significant differences were observed in other parts-of-speech categories. These findings collectively underscore the utility of NLP techniques in objectively assessing and characterizing language impairments in psychiatric populations, shedding light on the intricate interplay between linguistic abnormalities and symptomatology in schizophrenia.

Moreover, exploratory analyses revealed that SZ patients had higher semantic similarity than HC individuals ($p = 0.043$) and could be differentiated into two distinct groups with 86.84% accuracy using K-Means clustering based on Word2Vec embeddings. Overall, the study's findings suggest that SZ patients exhibit distinct linguistic patterns, including reduced syntactic complexity, increased self-referential language, and altered semantic structure, which are largely independent of language. These results underscore the potential of NLP methodologies in objectively evaluating linguistic features associated with schizophrenia. However, further research with larger sample sizes and in different linguistic contexts is warranted to validate and extend these findings.

The aim of the study "TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis" by Najafi and Varol (2024) was to assess the performance of TurkishBERTweet, a language model fine-tuned specifically for Turkish text, compared to other models in sentiment analysis and hate speech detection tasks. Trained on nearly 900 million tweets, TurkishBERTweet shares the architecture of the base BERT model but with a smaller input length, enhancing its speed and efficiency compared to BERTurk.

For hate speech detection, the study employed 5-fold cross-validation and evaluated models using the SIU 2023 hate speech detection competition dataset. TurkishBERTweet with LoRA fine-tuning achieved a macro-F1 score of 0.73167, surpassing the top-ranked submission in the competition (0.72167), showcasing its proficiency in identifying hate speech in Turkish text.

To evaluate model generalizability, an out-of-distribution evaluation was conducted. Despite a slight decrease in performance due to variations in testing datasets, TurkishBERTweet consistently outperformed BERTurk across various evaluation metrics, with statistically significant improvements observed on specific datasets such as Kemik-17bin, Kemik-3000, and TSATweets.

Furthermore, the study assessed TurkishBERTweet's efficiency in processing Turkish text data by comparing inference times statistically. Utilizing a sample of 1,000 Turkish tweets, TurkishBERTweet demonstrated significantly faster inference times than BERTurk and other baseline models, with p-values < 0.001. This efficiency was attributed to TurkishBERTweet's smaller model size and increased batch size capability, allowing it to process large volumes of data more rapidly. For instance, given Twitter's firehose data stream generates about 4,000 tweets per second, with less than 10% in Turkish, TurkishBERTweet can efficiently analyze such data streams in real-time.

In conclusion, the study provided valuable insights into TurkishBERTweet's performance and practicality for sentiment analysis and hate speech detection tasks in Turkish text. Its superior performance, especially when fine-tuned with LoRA, coupled with faster inference times, positions TurkishBERTweet as a valuable tool for researchers and practitioners in natural language processing and social media analytics. Additionally, by addressing the limitations of existing sentiment analysis models with a three-class classifier, the study contributes to advancing Turkish NLP research.

References:

Çabuk, T., Sevim, N., Mutlu, E., Yağcıoğlu, A. E. A., Koç, A., & Toulopoulou, T. (2024). Natural language processing for defining linguistic features in schizophrenia: A sample from Turkish speakers. *Schizophrenia Research*, *266*, 183-189. https://doi.org/10.1016/j.schres.2024.02.026.

Najafi, A., & Varol, O. (2023). TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis. arXiv e-prints. doi:10.48550/arXiv.2311.18063.