INTISARI

DETEKSI KALIMAT SARKASME DENGAN ARSITEKTUR SENTENCE ENCODER BERT DAN PERBAIKAN KATA OUT-OF-VOCABULARY MENGGUNAKAN MODIFIKASI ALGORITMA FONETIK SOUNDEX

Oleh
Afiyati
17/422647/SPA/00643

Deteksi sarkasme merupakan salah satu topik penelitian di bidang Natural Language Processing yang banyak dibahas oleh peneliti. Sarkasme yang ditulis pengguna pada media sosial saat ini banyak digunakan sebagai bentuk ekspresi kebencian, kritik, atau ejekan dengan tujuan menyinggung seseorang, sebuah produk maupun organisasi. Kalimat sarkasme menjadi sulit dikenali karena mempunyai makna berlawanan antara makna tertulis dengan makna yang ingin disampaikan. Kesulitan mengenali kalimat sarkasme juga terjadi dengan belum adanya pola tertentu untuk kalimat sarkasme pada teks bahasa Indonesia yang dapat mencirikan kalimat sarkasme. Kesulitan mengenali kalimat sarkasme juga ditambah dengan kurangnya penggunaan struktur Bahasa Indonesia yang baku saat menulis pada media sosial menyebabkan meningkatnya pertumbuhan jumlah kata baru dan kalimat yang tidak terstruktur yang dikenal sebagai kata *out-of-vocabulary* (OOV) sehingga menambah tantangan dalam tugas deteksi kalimat sarkasme.

Penelitian ini bertujuan untuk membangun model deteksi sarkasme dalam bahasa Indonesia dengan penanganan terhadap kata-kata OOV. Dataset dibangun dan dikumpulkan dari media sosial Twitter berbahasa Indonesia. Normalisasi dilakukan dengan memodifikasi algoritma Soundex dengan aturan kode fonetik khusus untuk Bahasa Indonesia. Kode Soundex akan digunakan oleh algoritma menghitung jarak dan kemiripan kata untuk menghasilkan dataset kamus OOV. Model yang dikembangkan adalah arsitektur *sentence encoder* BERT dan digabungkan dengan konsep *cosine similarity* yang menunjukkan fitur bertolak belakang antara dua kalimat cuitan.

Metode yang diusulkan telah menunjukkan bahwa perbaikan kata OOV dan penggunaan arsitektur *sentence encoder* BERT dengan fitur kalimat bertolak belakang dapat meningkatkan performansi deteksi kalimat sarkasme pada data media sosial dengan teks berbahasa Indonesia. Performansi model yang diajukan terlihat dari percepatan waktu proses serta dari nilai akurasi yang diperoleh sebesar 0.77, presisi 0.76, recall 0.73 dan F1 0.74.

Kata kunci: ***Sarkasme, Media Sosial, Bahasa Indonesia, Out-of-Vocabulary, Sentence Encoder, Cosine Similarity***

ABSTRACT

SARCASM SENTENCE DETECTION WITH BERT SENTENCE ENCODER ARCHITECTURE AND OUT-OF-VOCABULARY WORD HANDLING USING SOUNDEX PHONETIC ALGORITHM MODIFICATION

By

Afiyati
17/422647/SPA/00643

The challenge of research using data sourced from Indonesian-language social media as an object of research is currently getting bigger due to the rapid development of vocabulary written by users. Vocabulary mixed between regional languages, foreign languages, slang, abbreviations, and typographical errors. The lack of use of standard Indonesian structure when writing on social media has led to a growing number of new words and unstructured sentences known as out-of-vocabulary (OOV) words. Sarcasm detection written by users on social media is currently widely used as a form of expression of hatred, criticism or ridicule with the aim of offending a person, a product or an organization.

Sarcasm sentences are difficult to recognize because they have opposite meanings between the written meaning and the meaning to be conveyed. Difficulty recognizing sarcasm sentences also occurs with the absence of a specific pattern for sarcasm sentences in Indonesian texts that can characterize sarcasm sentences. So this research builds a dataset collected from Indonesian-language Twitter social media by handling or normalizing the word OOV. Normalization is done by modifying the Soundex algorithm with special phonetic code rules for Indonesian. The Soundex code will be used by the algorithm to calculate the distance and word similarity which will eventually produce the OOV dictionary dataset.

The model developed is the BERT sentence encoder architecture combined with the concept of cosine similarity which shows contradictory features between two tweet sentences. The proposed method shows that improving OOV words and using the BERT sentence encoder architecture combined with contrary context feature may improve the performance of detecting sarcasm sentences on social media data with Indonesian text. The performance of the proposed model can be seen from the accelerated processing time and from the accuracy value obtained of 0.77, precision 0.76, recall 0.73, and F1 0.74.

Keyword: ***Sarcasm, Social Media, Indonesian Language, Out-of-Vocabulary, Sentence Encoder, Cosine Similarity***