Enhancing Hate Speech Detection : Leveraging Emoji Preprocessing With LSTM Models

Background

- X is a social media platform where users share short "tweets". Tweets often include emojis that convey emotion or intent.
- Training a model to recognize and interpret emojis, as they provide additional meaning that improves the classification process.
- Previously, many classifications ignored emojis, but this research focuses on the important role of emojis in the classification process.

Purpose

- Analyzing the impact of emoji descriptions on hate speech classification with evaluation metrics.
- Analyzing the impact of emoji embedding on hate speech classification with evaluation metrics.
- Comparing the performance of models using emoji descriptions, emoji embeddings, and models that remove emojis from the data.

Methods

Data Prepocessing

Data Aumentation → Data Cleaning → Drop Column

We use the random swap 2-word augmentation technique to enrich the information from the limited dataset. And the atribut that we use is Text and Label Gold

Text Preprocessing

Delete Emoji / Emoji Desc / Emoji Embedding → Remove Punctuation → Case Folding → Remove Stopword → Tokenization → Normalization

Emoji Description: Replacing emojis with text that represente.
Delete Emoji: Delete emojis in the text
Emoji Embedding: Do not make any changes to the emoji

Before Classification we split data into data train and data test. For training data we use K-Fold Cross Validation for split data into train and validation data

Classification Using LSTM

Embedding Layer → LSTM Layer → Dropout Layer → Dense Layer

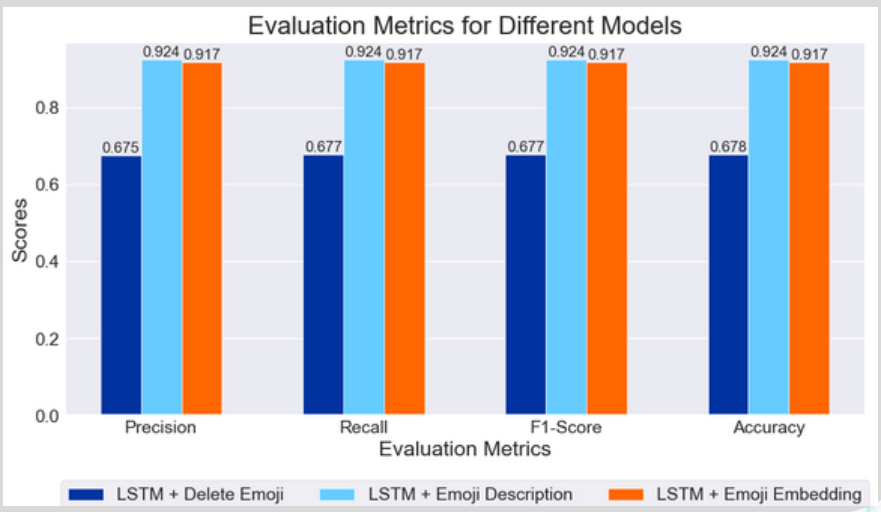
After classification is performed through the training and testing stages, the results will be analyzed using evaluation metrics such as accuracy, precision, recall, and F1 score.

Result

The experiment we conducted involved three different models at the text preprocessing stage. We also used two models for classifying text with emojis and one model for classifying text without emojis to observe the impact of emojis.

Evaluation matrix	LSTM + Delete Emoji	LSTM + Emoji Description	LSTM + Emoji Embedding
Precision	0,675	0,924	0,917
Recall	0,677	0,924	0,917
F1-Score	0,677	0,924	0,917
Accuracy	0,678	0,924	0,917

Adding emojis enriches information. Emojis can provide emotional context and clarify the meaning of a message.



Conclusion

- The use of emojis significantly enhances the accuracy of hate speech classification models. The LSTM + Emoji Description model achieved an accuracy of 0.924, and the LSTM + Emoji Embedding model achieved an accuracy of 0.917. In contrast, the model that deletes emojis (LSTM + Delete Emoji) only achieved an accuracy of 0.678.
- Emojis contribute additional context that is crucial for accurate predictions. Models that include emojis, either as embeddings or descriptions, perform better because they capture the semantic meaning conveyed by the emojis. Removing emojis eliminates important context, leading to lower model performance.
- For optimal results in hate speech classification involving emojis, it is recommended to incorporate emojis as embeddings or descriptions. Emojis are not merely decorative elements but significant semantic components in natural language analysis.

TASI-2324-105

- Yoga Sihombing
- Gabriel Panggabean
- Mares Siagian

Supervisior:

- Junita Amalia, S.Pd., M.Si
- Sarah Rosdiana Tambunan, S.Kom.,M.C

Bachelor of Information Systems
Faculty of Informatics and Electrical Engineering
Institut Teknologi Del

Enhancing Hate Speech Detection : Leveraging Emoji Preprocessing With LSTM Models

Latar Belakang

- X adalah platform media sosial tempat pengguna membagikan "tweet" pendek. Tweet sering menyertakan emoji yang menyampaikan emosi atau maksud.
- Melatih model untuk mengenali dan menginterpretasikan emoji, karena emoji memberikan makna tambahan yang memperbaiki proses klasifikasi.
- Sebelumnya banyak klasifikasi yang megabaikan emoji, namun pada penelitian ini memfokuskan pada peran penting emoji dalam proses klasifikasi.

Tujuan

- Menganalisis dampak deskripsi emoji terhadap klasifikasi ujaran kebencian dengan metrik evaluasi.
- Menganalisis dampak embedding emoji terhadap klasifikasi ujaran kebencian dengan metrik evaluasi.
- Membandingkan kinerja model yang menggunakan deskripsi emoji, embedding emoji, dan model yang menghapus emoji dari data.

Metode

Data Prepocessing

Data Aumentation → Data Cleaning → Drop Column

Kami menggunakan teknik augmentasi pertukaran 2-kata acak untuk memperkaya informasi dari dataset yang terbatas. Atribut yang kami gunakan adalah Teks dan Label Emas.

Text Preprocessing

Delete Emoji / Emoji Desc / Emoji Embedding → Remove Punctuation → Case Folding → Remove Stopword → Tokenization → Normalization

Emoji Description: Mengganti Emoji dengan teks aktualnya
Delete Emoji: Menghapus Emoji yang ada pada teks
Emoji Embedding: Mempertahankan Emoji pada teks

Sebelum klasifikasi, data dibagi menjadi data pelatihan dan data pengujian. Kami menggunakan K-Fold Cross Validation untuk membagi data pelatihan menjadi data pelatihan dan data validasi.

Classification Using LSTM

Embedding Layer → LSTM Layer → Dropout Layer → Dense Layer

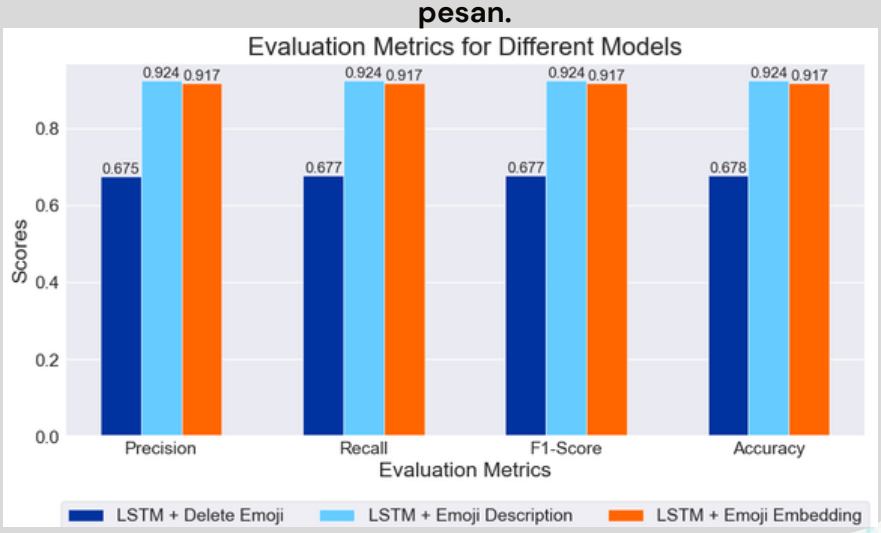
Setelah klasifikasi dilakukan melalui tahap pelatihan dan pengujian, hasilnya akan dianalisis menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan skor F1.

Hasil

Menambahkan emoji memperkaya informasi. Emoji dapat memberikan konteks dan memperjelas makna pesan.

Eksperimen yang kami lakukan melibatkan tiga model berbeda pada tahap preprocessing teks. Kami juga menggunakan dua model untuk mengklasifikasikan teks dengan emoji dan satu model untuk mengklasifikasikan teks tanpa emoji untuk mengamati dampak emoji.

Evaluation matrix	LSTM + Delete Emoji	LSTM + Emoji Description	LSTM + Emoji Embedding
Precision	0,675	0,924	0,917
Recall	0,677	0,924	0,917
F1-Score	0,677	0,924	0,917
Accuracy	0,678	0,924	0,917



Kesimpulan

- Penggunaan emoji secara signifikan meningkatkan akurasi model klasifikasi ujaran kebencian. Model LSTM + Deskripsi Emoji mencapai akurasi sebesar 0,924, sedangkan model LSTM + Emoji Embedding mencapai akurasi sebesar 0,917. Sebaliknya, model yang menghapus emoji (LSTM + Delete Emoji) hanya mencapai akurasi sebesar 0,678.
- Emoji memberikan konteks tambahan yang krusial untuk prediksi yang akurat. Model yang menyertakan emoji, baik sebagai embedding maupun deskripsi, menunjukkan kinerja yang lebih baik karena mereka menangkap makna semantik yang disampaikan oleh emoji. Menghapus emoji menghilangkan konteks penting, yang mengakibatkan kinerja model yang lebih rendah.
- Untuk hasil optimal dalam klasifikasi ujaran kebencian yang melibatkan emoji, disarankan untuk menggabungkan emoji sebagai embedding atau deskripsi. Emoji bukan hanya elemen dekoratif, melainkan komponen semantik yang signifikan dalam analisis bahasa alami.

TASI-2324-105

- Yoga Sihombing
- Gabriel Panggabean
- Mares Siagian

Supervisior:
- Junita Amalia, S.Pd., M.Si
- Sarah Rosdiana Tambunan, S.Kom.,M.Cs.

Program Studi Sistem Informasi
Fakultas Informatika
dan Teknik Elektro
Institut Teknologi Del