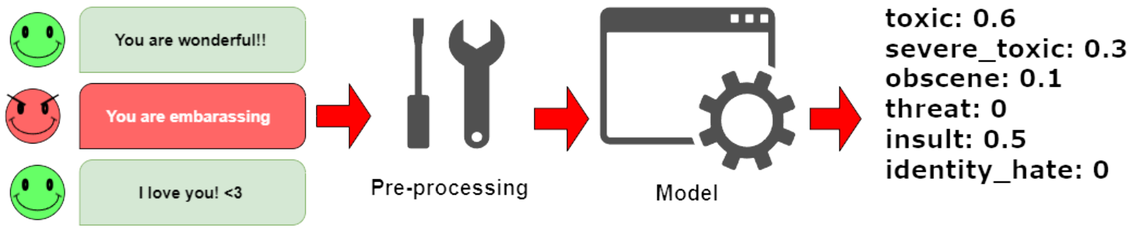
Toxic-Sentiment Analysis for live Chat

Victoria Grosu - Chiara Pesce - Francesco Perencin - Tommaso Talamo

Introduction

GOALS:
- Real-time detection
- User safety
- Automated moderation
- Flexibility

CHALLENGING:
- Understanding context
- Balanced language and Adaptability
- Mitigation of bias
- False positives and negatives



Related Works

Overlapping Toxic Sentiment Classification using Deep Neural Archit. [1]:
1. Three different Convolutional Neural Networks (CNNs)
2. Bidirectional LSTM/GRU

Challenges for Toxic Comment Classification [2]:
1. CNN (FastText/ Glove)
2. LSTM (FastText/ Glove)
3. BiLSTM/BiGRU (FastText/ Glove)

INNOVATION: The CLS token in BERT usually carries aggregate information about the sequence. Concatenating it with the context vector could provide a mix of both global (CLS) and local (context vector) information, potentially improving the model's ability to handle diverse tasks.

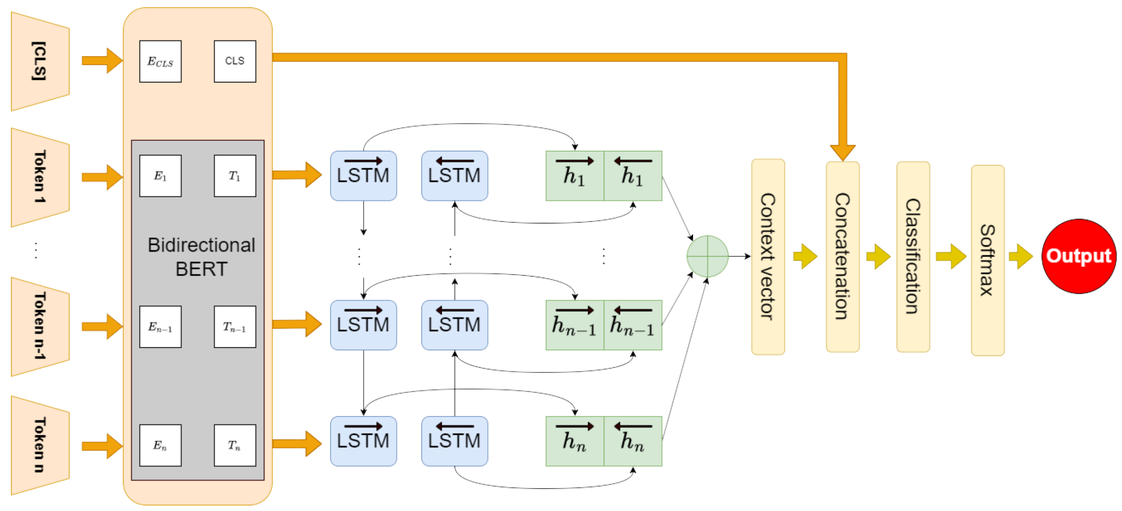
The proposed method

STEPS:
1. Pre-processing
2. BERT embedding and CLS token
3. LSTM/GRU with following attention mechanism
4. Concatenation between CLS token and context vector
5. Classification

When conducting sentiment analysis, the majority of research papers initially focused on binary classification (positive or negative sentiment) due to its simplicity and ease of comparison. However, the real-world sentiment is often more nuanced and can encompass a spectrum of emotions or opinions.

- **toxic**: level of toxicity detected inside the text
- **severe toxic**: additional score for exceptionally inappropriate comments
- **obscene**: inclination towards use of sensitive language
- **threat**: score that detects the presence of terms that threaten an individual
- **insult**: level of offensiveness expressed through personal slurs
- **identity hate**: aggressiveness towards an individual race, gender or religion.

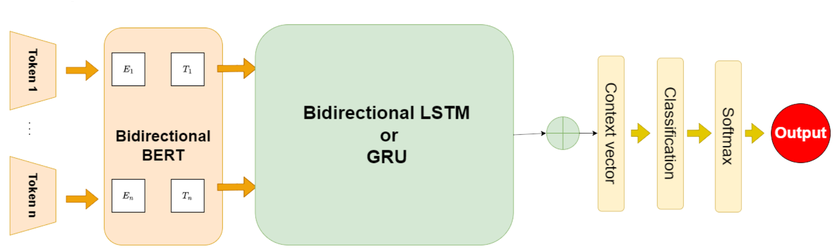
Architecture Pipeline

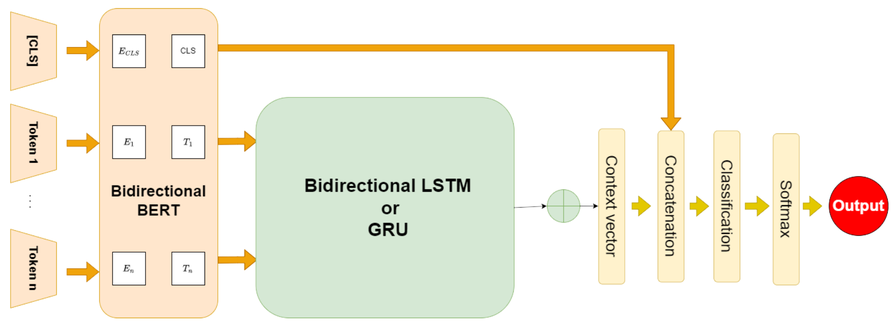


BERT + LSTM/GRU with Attention Mechanism Drawbacks:
1. computational complexity
2. susceptibility to overfitting
3. sensitivity to noise
4. challenges in interpretation
5. hyperparameter tuning
6. dependencies on the pretrained models

Experiments

Evaluate and **compare** four distinct architectures to see if the innovation is useful for our problem.





Results & Conclusions

In conclusion we expect to return an architecture capable of solving multi label classification problems when it comes to sentiment analysis. With the following results:
- expect accuracy metrics between 85% and 90%
- expect real-time performance, our goal is an average response time per message of less than 100 milliseconds

Model	Precision	Recall	F1
Bidirectional LSTM (Glove)	.74	.84	.777
Bidirectional GRU (Glove)	.73	.85	.772
Bidirectional GRU Attention (Glove)	.73	87	.779
Bidirectional LSTM Attention(BERT)	.78	.84	.779
Bidirectional GRU Attention (BERT)	.79	.87	.782
Bidirectional LSTM Attention + CLS (BERT)	.82	.89	.790
Bidirectional GRU Attention + CLS (BERT)	**.85**	**.90**	**.801**

References

[1] Hafiz Hassaan Saeed, Khurram Shahzad, and S. Chokshi. Overlapping Toxic Sentiment Classification using Deep Neural Architectures. *IEEE computer society*, 2018.

[2] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.

Università Ca'Foscari Venezia

CV ML LAB Computer Vision & Machine Learning