Arabic Toxic Tweet Classification: Leveraging the AraBERT Model

Paper Link

1. Summary:

1.1 Motivation :

The authors aim to address the problem of toxic content in Arabic social media platforms, which can cause online abuse and harassment. Social media platforms have emerged as the principal channels for information sharing and communication, enabling users to engage in interactive conversations. Regretfully, hate speech and insults are among the offensive and harmful content that is shared on these platforms. They noted that most existing studies focus on the English language and do not consider the linguistic and cultural diversity of Arabic. By creating a standardized Arabic dataset especially for the classification of toxic tweets, this study fills this gap. With the help of three linguists and native Arabic speakers, Google's Perspective API is automatically used to annotate the dataset. We use seven models in a series of experiments to compare the performance of the models.

This work sheds light on the significance of addressing toxicity in social media platforms while taking into account diverse languages and cultures, and it represents a significant advancement in the classification of toxic tweets in Arabic.

1.2 Contribution:

The authors make the following contributions:

- ➢ They construct a novel dataset of 31,836 Arabic tweets, annotated as toxic or non-toxic using Google's Perspective API and the expertise of three native Arabic speakers and linguists.
- ➢ They conduct a series of experiments using seven models: LSTM, BiLSTM, CNN, GRU, BiGRU, mBERT, and AraBERT. They also employ word embedding techniques such as AraBERT and mBERT.
- ➢ They demonstrate that the fine-tuned AraBERT model outperforms other models, achieving an accuracy of 0.9960. They also show that their accuracy value surpasses similar approaches reported in recent literature.
- ➢ They advance the field of Arabic toxic tweet classification, highlighting the importance of addressing toxicity in social media platforms while considering diverse languages and cultures.

1.3 Methodology:

The authors follow a three-phase methodology:

a. Dataset creation
b. Preprocessing
c. Classification.

In the dataset creation phase, they collect Arabic tweets containing words from a list of toxic keywords, and filter them using Google's Perspective API. They then manually annotate the tweets as toxic or non-toxic using three experts.

In the preprocessing phase, they clean, normalize, and remove stop words from the tweets.

In the classification phase, they train and fine-tune different models using the preprocessed tweets as features and the corresponding labels as targets. They calculate and compare the performance metrics of each model, such as accuracy, precision, recall, and F1 score.

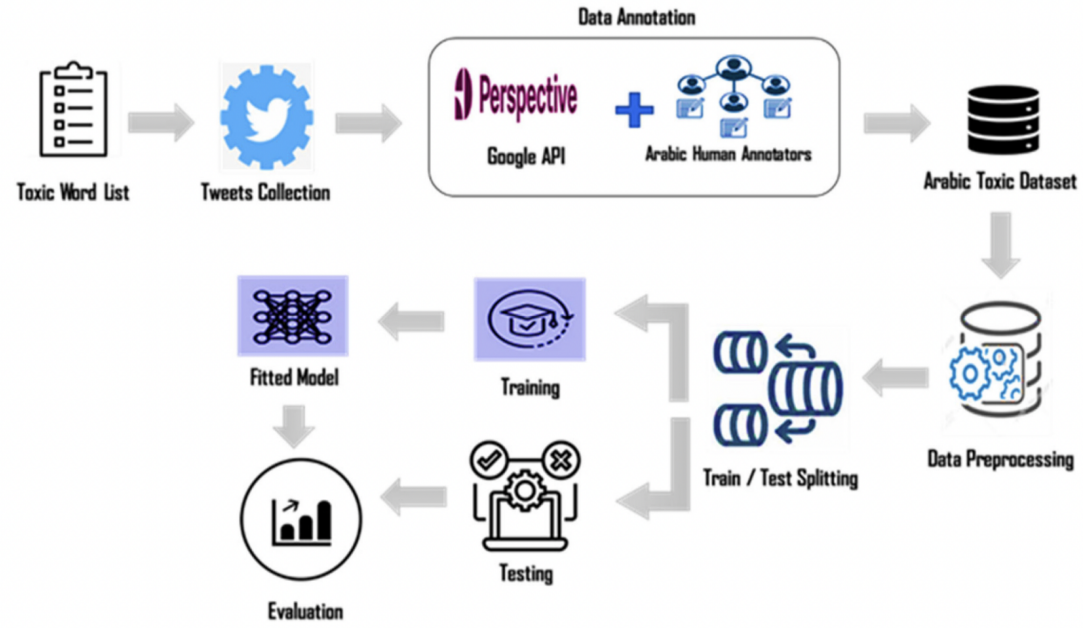

Figure 1. Summary of methodology steps: Dataset creation, preprocessing, and classification.

1.4 Conclusion:

The authors conclude that their proposed dataset is a valuable resource for Arabic toxic tweet classification, and that their fine-tuned AraBERT model is the most effective in identifying toxic content in Arabic tweets. They also suggest some directions for future work, such as expanding the dataset to include more dialects and categories of toxicity, and exploring the impact of toxicity on user behavior and engagement.

2. Limitations:

2.1 First Limitation :

Dataset Size and Diversity: The authors acknowledge that their dataset is relatively small compared to other datasets for toxic content analysis in English. They also note that their dataset may not cover all the variations and nuances of Arabic language and culture, as it is mainly based on a list of toxic keywords and filtered by Google's Perspective API,

which may not capture all the aspects of toxicity. Moreover, their dataset may not reflect the temporal and spatial dynamics of toxicity, as it is collected from a specific period and location.

2.2 Second Limitation :

Model Generalization and Interpretability:

The authors admit that their fine-tuned AraBERT model may not generalize well to other domains and tasks, as it is tailored to their specific dataset and task. They also point out that their model may not be easily interpretable, as it is based on a complex neural network architecture that relies on attention mechanisms and hidden layers. They suggest that future work could explore ways to improve the generalization and interpretability of their model, such as using explainable AI techniques and incorporating human feedback.

3. Synthesis:

The ideas in the paper relate to potential applications or future scopes in the following ways:

- ➢ The paper provides a valuable resource for researchers and practitioners interested in studying and combating toxic content in Arabic social media platforms, as it offers a standardized and publicly accessible dataset and a comprehensive evaluation of different models and techniques.
- ➢ The paper contributes to the advancement of Arabic NLP, as it demonstrates the superiority of the AraBERT model, which is a pre-trained language model specifically designed for Arabic texts and dialects, over other multilingual and monolingual models.
- ➢ The paper opens up new avenues for further research and development in the field of toxic content detection and prevention, as it suggests possible ways to improve the quality and diversity of the dataset, the generalization and transferability of the models, and the incorporation of other factors and features, such as context, intent, and sarcasm, into the classification task.