uOttawa

Project Proposal
Group ID : 7

DTI5125[EG] Data Science Applications [LEC] 20235

Submitted By:

Abdelrahman Ahmed Mansour Badran
Hussien Tarek Ismail Abdelrazik
Mohamed Ahemd Sayed Mohamed
Mohamed Magdy Mahmoud Elasmar

Instructor:

Professor: Arya Rahgozar

University of Ottawa ,Ca

July 16, 2023

Main Article Used

N. von Boguszewski, S. Moin, A. Bhowmick, S. M. Yimam, and C. Biemann, "How Hateful are Movies? A Study and Prediction on Movie Subtitles." arXiv, Aug. 19, 2021. Accessed: Jul. 14, 2023. [Online]. Available: http://arxiv.org/abs/2108.10724

1 Title

Detecting Hate Speech in Movie Subtitles: A Machine Learning Approach

2 Problem Formulation

The aim of this research proposal is to develop a machine learning-based approach to detect hate speech in movie subtitles. Hate speech is a significant societal issue, and its detection can contribute to creating safer and more inclusive environments, including in the realm of entertainment media. By analyzing movie subtitles, we can identify instances of hate speech and offensive language, allowing for targeted interventions and content moderation. The proposed methodology will leverage natural language processing techniques and transfer learning to adapt existing hate speech detection models to the domain of movie subtitles.

3 Methodology

- Pre-processing: Apply text pre-processing techniques to clean and normalize the movie subtitle data. This may include removing noise, special characters, and punctuation, as well as lemmatization and tokenization.
- Model Selection: Explore different machine learning models suitable for hate speech detection, such as Bag-of-Words, Bi-LSTM, and BERT-based models. Assess their performance on social media datasets (e.g., Twitter) to establish baseline results. [1]
- Domain Adaptation: Train and evaluate the selected models on publicly available hate speech datasets from social media platforms (e.g., Twitter and FoxNews). Fine-tune the models to classify hate speech, offensive language, and normal speech effectively.
- Movie Subtitle Dataset: Collect a novel dataset of movie subtitles, including a diverse set of movies across different genres. Annotate the dataset for hate speech, offensive language, and normal speech using crowd-sourcing platforms like Amazon Mechanical Turk.
- Transfer Learning to Movie Subtitles: Apply domain adaptation techniques to transfer the knowledge learned from the social media datasets to the movie subtitle dataset. Train and evaluate the models on the movie subtitle dataset, considering different classification metrics.
- Ensemble Learning: Investigate the potential of ensemble learning techniques to improve hate speech detection in movie subtitles. Combine the predictions of multiple models to enhance the overall performance and robustness of the system.
- Evaluation Method: Measure the performance of the models using evaluation metrics such as precision, recall, and F1-score. Conduct statistical analyses to compare the results of different models and identify the best-performing model for hate speech detection in movie subtitles.
- Packages and libraries : re, pandas, numpy, matplotlib, seaborn, string, nltk, warnings, wordcloud, PIL, urllib, requests, keras_preprocessing, keras, sklearn, tensorflow, os, transformers, torch, collections

4 Data Description and Data Sources

The data for this research will consist of two main sources:

1. For our research, we will utilize publicly available hate speech datasets from social media platforms, specifically Twitter and FoxNews. These datasets have been annotated to classify text into categories of hate speech, offensive language, and normal speech, providing a valuable foundation for training and fine-tuning our hate

speech detection models. The Twitter dataset[2] consists of a large collection of tweets that have been labeled with three categories: normal, hate, and offensive. This dataset offers a diverse range of text content from social media, allowing us to capture different forms of hate speech prevalent on the platform. On the other hand, the FoxNews dataset[3] contains comments from discussion threads on the FoxNews website. These comments have been labeled as either normal or hate speech, focusing specifically on the hate speech aspect. This dataset provides insights into hate speech prevalent in news-related online platforms. By leveraging these social media datasets, we can train our hate speech detection models to effectively identify hate speech in movie subtitles. The inclusion of multiple labels in the Twitter dataset (normal, hate, offensive) and the hate-specific focus in the bi-labeled FoxNews dataset (normal, hate) allows for a comprehensive analysis of hate speech in different contexts.

2. Movie Subtitle Dataset: Collect a novel dataset of movie subtitles from a diverse set of movies across various genres. Annotate the dataset using crowdsourcing platforms to label instances of hate speech, offensive language, and normal speech.

5 Evaluation Method

In our evaluation, we will assess the performance of the proposed hate speech detection models using standard evaluation metrics such as precision, recall, and F1-score. These metrics will provide insights into the models' effectiveness in accurately identifying hate speech in movie subtitles. To ensure the generalizability and robustness of the models, we will employ cross-validation and train-test splits. By using these techniques, we can evaluate how well the models perform on different subsets of the data and assess their ability to handle new instances.

Furthermore, as a benchmark, we will include the evaluation of another state-of-the-art BERT-based classification model called HateXplain [4]. While we acknowledge that fine-tuning the HateXplain model is possible, we will focus on reporting the performance of the "off-the-shelf" classification system on new domains, specifically movie subtitles. This will allow us to compare the results obtained from different models and identify the most effective approach for hate speech detection in movie subtitles.

6 Expected Results

We anticipate that our proposed approach will yield promising results in detecting hate speech and offensive language in movie subtitles. We expect the models trained on social media data and fine-tuned using the movie subtitle dataset to demonstrate improved performance in identifying hate speech instances accurately. By leveraging transfer learning and domain adaptation techniques, we aim to achieve high precision and recall values, leading to a higher F1-score. Additionally, we expect the evaluation results to provide insights into the strengths and weaknesses of different machine learning models for hate speech detection in movie subtitles.In addition, considering the potential benefits of ensemble learning techniques, we anticipate that combining the predictions of individual models through ensemble methods will further enhance the accuracy and effectiveness of hate speech detection in movie subtitles, surpassing the performance achieved by individual models alone. Overall, this research proposal aims to contribute to the development of effective hate speech detection methods in the context of movies, fostering a safer and more inclusive media environment.

References

[1] N. von Boguszewski, S. Moin, A. Bhowmick, S. M. Yimam, and C. Biemann, "How hateful are movies? a study and prediction on movie subtitles," Aug 2021. arXiv:2108.10724 [cs].

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pp. 512–515, 2017.

[3] L. Gao and R. Huang, "Detecting online hate speech using context aware models," May 2018. arXiv:1710.07395 [cs].

[4] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," Apr 2022. arXiv:2012.10289 [cs].