Hate Speech Detection: Methods and Challenges

Akshaya Basani
Anoushka Gupta
Sanya Garg

May 16, 2024

What is Hate Speech?

- Hate speech is any communication that belittles a person or a group based on characteristics such as race, religion, ethnicity, gender, sexual orientation, disability, or other traits.
- It can incite violence, spread misinformation, and contribute to social polarization.

Importance of Detecting Hate Speech

- Protects individuals and communities from harm.
- Maintains social harmony and public safety.
- Complies with legal and platform-specific regulations.

Challenges in Hate Speech Detection

- Ambiguity and context-dependence of language.
- The evolving nature of hate speech and slang.
- Balancing freedom of speech with regulation.

Data Collection and Preprocessing

- **Data Sources:** Social media platforms, news articles, forums.
- **Preprocessing:** Tokenization, removing stop words, stemming, and lemmatization.
- **Annotation:** Manual labeling by human annotators or using predefined dictionaries.

Feature Extraction

- **Text-Based Features:** Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF).
- **Linguistic Features:** Sentiment analysis, part-of-speech tagging.
- **Deep Learning Features:** Word embeddings (Word2Vec, GloVe), contextual embeddings (BERT, GPT).

Modeling Techniques

- **Traditional Machine Learning:** Naive Bayes, SVM, Random Forest.
- **Deep Learning:** CNN, RNN, LSTM, Transformers.
- **Ensemble Methods:** Combining multiple models to improve performance.

Performance Metrics

- **Accuracy:** Proportion of correctly classified instances.
- **Precision, Recall, F1-Score:** Evaluates the balance between precision (accuracy of positive predictions) and recall (completeness of positive predictions).
- **ROC-AUC:** Measures the ability of the model to distinguish between classes.

Real-World Applications

- **Social Media Monitoring:** Facebook, Twitter, YouTube.
- **Law Enforcement:** Detecting and preventing hate crimes.
- **Content Moderation:** Automated filtering of harmful content.

Case Study: Social Media Platform

- **Problem:** Identifying and removing hate speech from user posts.
- **Approach:** Combining machine learning with human moderation.
- **Outcome:** Improved detection rates and user safety.

Conclusion

- Hate speech detection is a critical task in maintaining online safety and social harmony.
- Advances in NLP and machine learning have significantly improved detection capabilities.
- Ongoing research is needed to handle evolving language and new forms of hate speech.

Bibliography I