Cyberbullying Classification

Project Proposal

Group 10: Pushyanth Damarapati, Sindhya Balasubramanian, Eileen Chang, Priyanka Padinam

Description

The rise of social media and the recent couple of years of covid-19 lockdown has led to a concerning increase in cyberbullying cases. In 2020, UNICEF even issued a warning in response to the increased cyberbullying compounded by social distancing and increased screen-time.

Those who bully others on the internet have the convenience of being able to hide anonymously behind a screen, but the people who are bullied are likely to develop mental-health issues that persist even after the bullying has ceased. Due to social media's ability to spread information quickly and anonymously, a single person can easily end up being targeted by a large number of people of various demographics.

We aim to create a model that will flag harmful tweets and, therefore, protect targets of cyberbullying.

Dataset

We will be using a kaggle dataset, Cyberbullying Classification, consisting of more than 47,000 tweets labeled according to 6 classes of cyberbullying: Age, Ethnicity, Gender, Religion, Other type of cyberbullying, and Not cyberbullying. Each row of the dataset will have a tweet and its class of cyberbullying. The dataset is meant to be used to create a multi-classification model to predict cyberbullying type, create a binary classification model to flag potentially harmful tweets, and examine words and patterns associated with each type of cyberbullying.

Methodology and Expected Results

The dataset on hand is granular to the detail of comment posted that was either considered cyber bullying or describes the event on hand and has been labeled as such in the following terms -

- Cyber Bullying (Further subdivided under Age, Ethnicity, Gender, Religion and Other)
- Not Cyber Bullying

This project can hence be scoped into the following following high level components -

- Classification of comment as Cyber Bullying vs Not Cyber Bullying
- Classification of Cyber Bullying comment into subcategories of -
    - Age
    - Ethnicity
    - Gender
    - Religion
    - Other

The project on hand will hence use the following methods of implementation -

Key Python Libraries - Scikit-learn, NLTK (For Preprocessing), Matplotlib/Seaborn (Visualization), Tensorflow (Text Classification))

1. *Two-ClassText Classification as Cyber Bullying vs Not Cyber Bullying* -
    1. Step one would be to import the text corpus and create a new dataset labeled as Not Cyber Bullying and Cyber Bullying (umbrella class to cover age, religion and other categories)
    2. Each document in the corpus will first undergo the following preprocessing steps:
        1. Remove hyperlinks using regular expression
        2. Clean out punctuation marks using punctuation component (NLTK)

3. Convert documents into lower case using string manipulation
4. Removal of stop words (NLTK)
5. Stemming using porter stemmer (NLTK)
6. Lemmatization of document (NLTK)

3. Data Visualization is key here to understand class imbalance
4. To ensure there no class imbalance, post visualization, the data will over/under sampled based on the majority and minority classes noted in the data
5. Post preprocessing and sampling, the documents will tokenized and converted to an appropriate vector format for model consumption using -
    1. Bag of Words (countvectorizer)
    2. TF-IDF (tfidfvectorizer) (Preferred method of vectorization)
6. Data will be divided into Training (60%), Validation(20%) and Test set (20%) using Scikit-learn's train test data separator method using random seed
7. The following models will be experimented with to accomplish text classification -
    1. Multinomial Naive Bayes Model will be implemented from scratch using K-fold cross-validation
    2. Artificial Neural Network (with different regularization and dropout layers)
    3. Convolutional Neural Networks (with different regularization and dropout layers)
    4. Graph Convolutional Neural Network (with different regularization and dropout layers)
8. Results of all the different models will be compared using the following metrics -
    1. ROC (AUC) curve visualization
    2. Accuracy
    3. Precision
    4. Recall
    5. F1 Score

2. *Multi-Class Text Classification as Age, Ethnicity, Gender, Religion and other within Cyber Bullying* -
    1. Step one would be to import the text corpus and create a new dataset containing only cyber bullying records retaining their original labels of age, religion and other categories
    2. Each document in the corpus will first undergo the following preprocessing steps:
        1. Remove hyperlinks using regular expression
        2. Clean out punctuation marks using punctuation component (NLTK)
        3. Convert documents into lower case using string manipulation
        4. Removal of stop words (NLTK)
        5. Stemming using porter stemmer (NLTK)
        6. Lemmatization of document (NLTK)
    3. Data Visualization is key here to understand class imbalance
    4. To ensure there no class imbalance, post visualization, the data will over/under sampled based on the majority and minority classes noted in the data
    5. Post preprocessing and sampling, the documents will tokenized and converted to an appropriate vector format for model consumption using -
        1. Bag of Words (countvectorizer)
        2. TF-IDF (tfidfvectorizer) (Preferred method of vectorization)
    6. Clustering of the data will be performed to understand if each of the classes are clustered as labeled
    7. Visualization of clustering will be used to make inferences on the labeled data

8. Data will be divided into Training (60%), Validation(20%) and Test set (20%) using Scikit-learn's train test data separator method using random seed
9. The following models will be experimented with to accomplish text classification -
    1. Multinomial Naive Bayes Model with K-fold cross-validation
    2. Artificial Neural Network (with different regularization and dropout layers)
    3. Convolutional Neural Networks (with different regularization and dropout layers)
10. Results of all the different models will be compared using the following metrics -
    1. ROC (AUC) curve visualization
    2. Accuracy
    3. Precision
    4. Recall
    5. F1 Score

Timeline

Week 9	Clean and preprocess the data
Week 10	Implement the models and train them
Week 11	Use clustering algorithms to group the data
Week 12	Test the models, try other models, and visualize the results
Week 13	Write the report and record the presentation

Responsibilities

Priyanka: Preprocessing, Clustering, ANN model, Presentation
Sindhya: Proposal Write-up, CNN Model, GCN Model
Pushyanth: Preprocessing, Naive Bayes model, Presentation
Eileen: Proposal Write-up, Final Report Write-up, Visualizations

Resources

Cyberbullying Classification
ArXiv Twitter Page of Recent NLP Papers