Text Classification of Online Hate Speech

Created by Binh Bui, Pooja Vembu Rajan, Alexander Watkins

Agenda

- Project Background
- Technical Features
- App Demo

Project Background

Project Motivation

- Increasing number of ways to share voice online
  - Twitter/X, Instagram, TikTok, etc.
- Manual content moderation impractical due to number of users
  - Ex: > 1 billion active users on TikTok
- Hateful content incredibly harmful towards those targeted
- Use NLP methods to automatically detect presence of hate speech in content

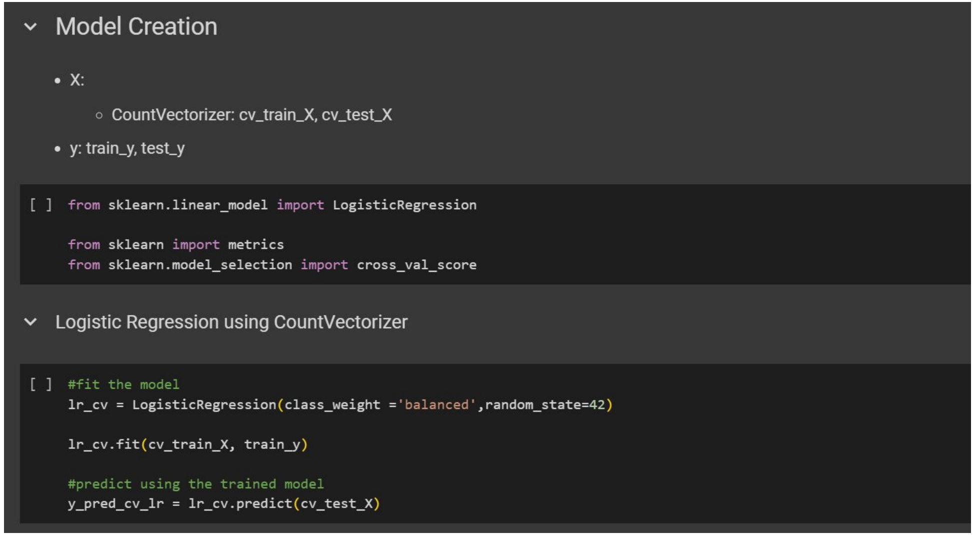
Hate Speech Definition

- The United Nations defines hate speech as:
    - **"Any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor"**
- How to translate this definition into a consistent classification model?

Technical Features

NLP Functionality

- Standard preprocessing techniques
  - Removed stopwords, tokenization, lemmatization
- Transformed text with CountVectorizer
- Trained logistic regression model to classify text as either:
  - 0 - No hate speech
  - 1 - Contains hate speech



Model Creation

- X:
  - CountVectorizer: cv_train_X, cv_test_X
- y: train_y, test_y

```
[ ]  from sklearn.linear_model import LogisticRegression

     from sklearn import metrics
     from sklearn.model_selection import cross_val_score
```

Logistic Regression using CountVectorizer

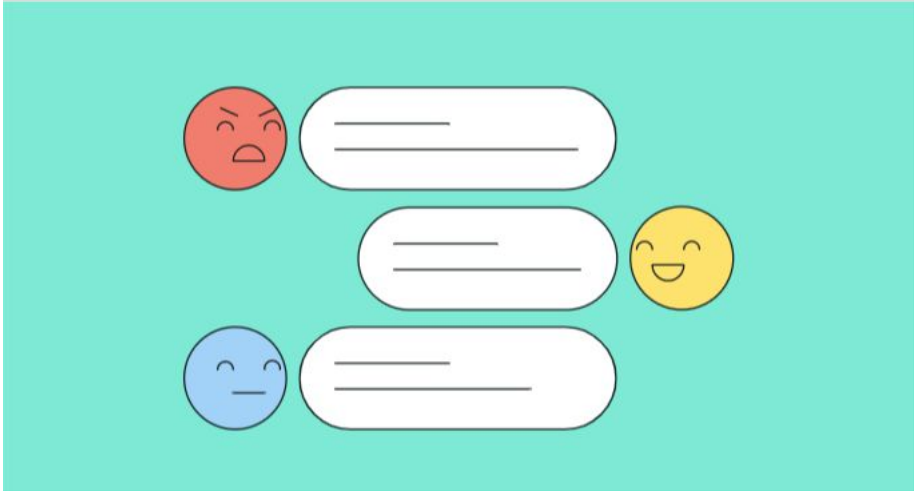
```
[ ]  #fit the model
     lr_cv = LogisticRegression(class_weight ='balanced',random_state=42)

     lr_cv.fit(cv_train_X, train_y)

     #predict using the trained model
     y_pred_cv_lr = lr_cv.predict(cv_test_X)
```

App Features

- Simulate a social media post
    - Present text to “comment” on
- Allow user to input text that may contain hate speech
- Run given text through LR model and inform user whether it contains hate speech or not
- Add text to “comment section” if it does not contain hate speech



Time for the demo!

References

Laub, Z. (2019, April 11). *Hate Speech on Social Media: Global Comparisons*. Indian Strategic Knowledge Online. Retrieved October 15, 2023, from https://indianstrategicknowledgeonline.com/web/Hate%20Speech%20on%20Social%20Media_%20Global%20Comparisons%20_%20Council%20on%20Foreign%20Relations.pdf

Malmasi, S., & Zampieri, M. (2017, December 26). *Detecting Hate Speech in Social Media*. arXiv. Retrieved October 15, 2023, from https://doi.org/10.48550/arXiv.1712.06427

United Nations. (n.d.). *Understanding Hate Speech*. United Nations. Retrieved October 15, 2023, from https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech