Web Mining Project

UNIVERSITÄT MANNHEIM

Hate Speech Detection on Social Media

2034311	**Nguyen Anh-Nhat**	1869371	**Momin Abdullah**
2035819	**Lei Minghao**	1873027	**Wei Yiyi**
1961779	**Lin Yueyi**		**Team 1**

Why Hate Speech ?

- Hatred, discrimination, and hate-driven violence have unfortunately been common problems in societies across history.
- The rise of the internet and online platforms has allowed hateful speech spread widely and quickly

Agenda

UNIVERSITÄT MANNHEIM

01 Introduction

02 Data Preprocessing

03 Models

04 Result Analysis

01 Introduction

Introduction-**What is Hate Speech ?**

- **Hate speech** targets and attacks **specific** groups based on characteristics like race, religion, or sexual orientation, often inciting violence or hatred.

- **Offensive language** on the other hand, is insulting or abusive **without** targeting a protected group based on such characteristics.

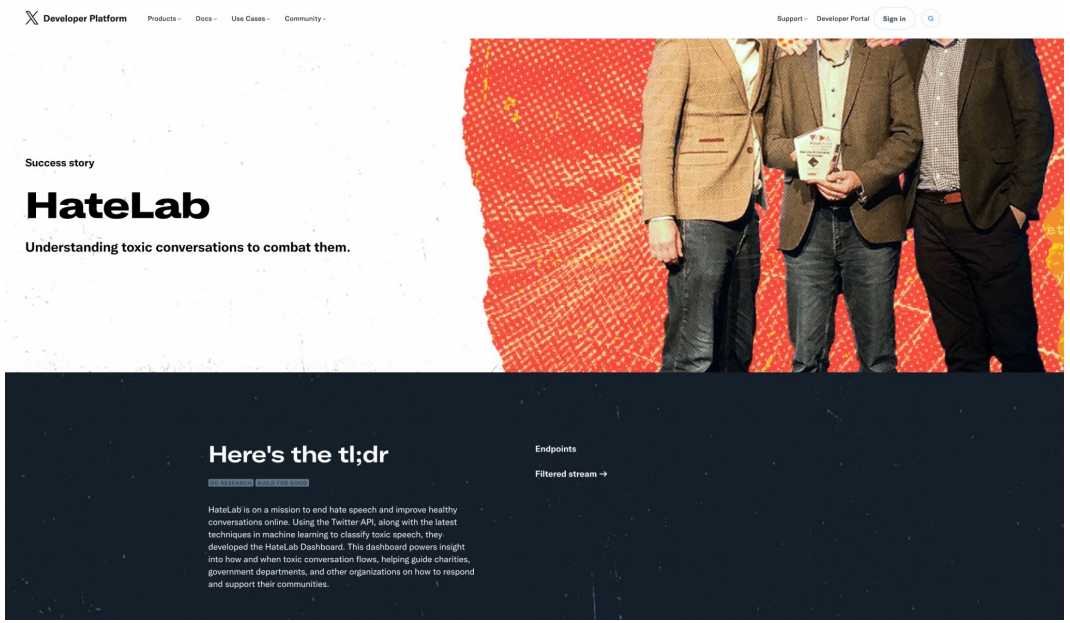


Figure: HateLab - A Lab for Hate Speech Researching on X.com

Introduction- **Hate Speech and Offensive Example**

- Offensive: "You're a complete **idiot** and a **waste of space**. Your ideas are downright stupid."
- Hate speech: "All **immigrants** are leeches on society and should be deported immediately. They contribute nothing and just breed **crime** and **violence**."

02 Data Preprocessing

Data Preprocessing-**Data Structure**

UNIVERSITÄT MANNHEIM

Overview

Attribute	Attribute Explanation
Count	Number of CrowdFlower users who coded each tweet (min is 3), sometimes more users coded a tweet when
Hate_Speech	Number of CF users who judged the tweet to be hate speech
Offensive_Language	Number of CF users who judged the tweet to be offensive
Neither	Number of CF users who judged the tweet to be neither offensive nor non-offensive
Class	Class label for majority of CF users
Tweet	The text content of the tweet or message collected

Infos

24,783 pieces of Data

No Missing Value

No Duplicate Data

Data Preprocessing-**Original Text Data**

UNIVERSITÄT MANNHEIM

@Username

URL

Emoji

tweet
"@2015seniorprobs: I probably wouldn’t mind school as much if we didn’t have to deal with bitch ass teachers". Retweet
"@A7XDemery: I'm a fucking fag they said"
"@ARIZZLEINDACUT: Females think dating a pussy is cute now? http://t.co/VxBJg26Gsz" how does doing this stuff make him a pussy?
"@Addicted2Guys: -SimplyAddictedToGuys http://t.co/1jL4hi8ZMF" woof woof hot scally lad
"@AdoreBellaaa: Have ya ever asked your bitch for other bitches - kanye voice" Yes
"@AdoreZoey: How u gone bring ur side bitch to a game where You know Ya gf friends at ?! 😩😩😩😩" I SWEAR!!!!!
"@AllAboutManFeet: http://t.co/3gzUpfuMev" woof woof and hot soles
"@Allyhaaaaa: Lemmie eat a Oreo & do these dishes." One oreo? Lol
"@Almightywayne__: @JetsAndASwisher @Gook___ bitch fuck u http://t.co/pXmGA68NC1" maybe you'll get better. Just http://t.co/TPreVwfq0S
"@Almightywayne__: Fuck Red Malone man bitch ass niggah" could you please use complete sentences?
"@ArizonasFinest6: Why the eggplant emoji doe?"y he say she looked like scream lmao
"@AutoWorld: Hennessey Venom GT 🙈 http://t.co/i8eGMnKaJ9" that's one sexy bitch
"@BOSSBYTCHH: Him seh me pussy wetter then a shower curtain....#ahmesehwetness"<lmao!!
"@BRO_HEN314: #Eaglesnation and every #Eagles need to see that pic I just posted because that bitch just said the most racist shit"

Data Preprocessing-**Preprocessing Steps**

tweet
!!! RT As a woman you shouldn't complain about cleaning up your house. as a man you should always take the trash out...
!!!!! RT boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
!!!!!!! RT Dawg!!!! RT You ever fuck a bitch and she start to cry? You be confused as shit
!!!!!!!!! RT she look like a tranny
!!!!!!!!!!!!! RT The shit you hear about me might be true or it might be faker than the bitch who told it to ya

token
rt,woman,shouldnt,complain,cleaning,house,man,always,take,trash
rt,boy,dat,coldtyga,dwn,bad,cuffin,dat,hoe,1st,place
rt,dawg,rt,ever,fuck,bitch,start,cry,confused,shit
rt,look,like,tranny
rt,shit,hear,might,true,might,faker,bitch,told,ya

Vectorization

Input Data

UNIVERSITÄT MANNHEIM

03 Models

Logistic Regression Model

UNIVERSITÄT MANNHEIM

— a statistical technique, employs the logistic model to predict the likelihood of an event by evaluating the log-odds

```
%%time
from sklearn.linear_model import LogisticRegression

# Initialize Logistic Regression classifier
logreg = LogisticRegression(max_iter=1000, class_weight='balanced')

logreg.fit(tfidf_vectors.toarray(), y_train)
logreg_test_preds = logreg.predict(X_test_tfidf.toarray())

# Evaluate the Logistic Regression classifier
train_score_logreg = evaluate(y_train, logreg.predict(tfidf_vectors.toarray()))
test_score_logreg = evaluate(y_test, logreg_test_preds)
```

Performance:

Dataset	Accuracy	Precision	Recall	F1-Score
Train Set	89.31	93.36	89.31	90.36
Test Set	83.84	88.84	83.84	85.59

Multinomial Naive Bayes Model

UNIVERSITÄT MANNHEIM

— a variant of the Naive Bayes algorithm, which is a probabilistic classification algorithm based on Bayes' Theorem

```
%%time
from sklearn.naive_bayes import MultinomialNB, BernoulliNB

# Initialize Multinomial Naive Bayes classifier
mnb_count = MultinomialNB()
mnb_tfidf = MultinomialNB()

# Fit the data to MultinomialNB using CountVectorizer
mnb_count.fit(count_vectors.toarray(), y_train)

# Fit the data to MultinomialNB using TF-IDF vectorizer
mnb_tfidf.fit(tfidf_vectors.toarray(), y_train)

# Predictions on validation data
count_test_preds_mnb = mnb_count.predict(X_test_count.toarray())
tfidf_test_preds_mnb = mnb_tfidf.predict(X_test_tfidf.toarray())

# Calculate training scores for MultinomialNB
train_score_count_mnb = evaluate(y_train, mnb_count.predict(count_vectors.toarray()))
train_score_tfidf_mnb = evaluate(y_train, mnb_tfidf.predict(tfidf_vectors.toarray()))

# Calculate validation scores for MultinomialNB
test_score_count_mnb = evaluate(y_test, count_test_preds_mnb)
test_score_tfidf_mnb = evaluate(y_test, tfidf_test_preds_mnb)
```

Performance:

MultinomialNB using CountVectorizer				
Dataset	Accuracy	Precision	Recall	F1-Score
Train Set	90.05	90.29	90.05	90.14
Test Set	87.94	87.44	87.94	87.66

MultinomialNB using TF-IDF				
Dataset	Accuracy	Precision	Recall	F1-Score
Train Set	86.45	86.92	86.45	83.3
Test Set	84.49	85.75	84.49	80.71

Bernoulli Naive Bayes Model

— a subset of the Naive Bayes Algorithms. In contrast to the Multinomial Naive Bayes model, which works with term frequencies, the Bernoulli Naive Bayes model considers only the presence or absence of each feature (binary features) in the dataset

```
%%time
# Initialize Bernoulli Naive Bayes classifier
bnb_count = BernoulliNB()
bnb_tfidf = BernoulliNB()

# Fit the data to BernoulliNB using CountVectorizer
bnb_count.fit(count_vectors.toarray(), y_train)

# Fit the data to BernoulliNB using TF-IDF vectorizer
bnb_tfidf.fit(tfidf_vectors.toarray(), y_train)

# Predictions on validation data
count_test_preds_bnb = bnb_count.predict(X_test_count.toarray())
tfidf_test_preds_bnb = bnb_tfidf.predict(X_test_tfidf.toarray())

# Calculate training scores for BernoulliNB
train_score_count_bnb = evaluate(y_train, bnb_count.predict(count_vectors.toarray()))
train_score_tfidf_bnb = evaluate(y_train, bnb_count.predict(tfidf_vectors.toarray()))

# Calculate validation scores for BernoulliNB
test_score_count_bnb = evaluate(y_test, count_test_preds_bnb)
test_score_tfidf_bnb = evaluate(y_test, tfidf_test_preds_bnb)
```

Performance:

BernoulliNB using CountVectorizer				
Dataset	Accuracy	Precision	Recall	F1-Score
Train Set	90.06	90.14	90.06	90.04
Test Set	87.88	87.14	87.88	87.43

BernoulliNB using TF-IDF				
Dataset	Accuracy	Precision	Recall	F1-Score
Train Set	90.06	90.14	90.06	90.04
Test Set	87.88	87.14	87.88	87.43

Random Forest Model

UNIVERSITÄT MANNHEIM

— also known as Random Decision Forests, is an ensemble learning technique designed for classification, regression, and various other tasks

```
%%time
from sklearn.ensemble import RandomForestClassifier

random_forest = RandomForestClassifier(class_weight='balanced')
random_forest.fit(tfidf_vectors.toarray(), y_train)

# Predictions on test data
rf_test_preds = random_forest.predict(X_test_tfidf.toarray())

# Evaluate the Random Forest classifier
train_score_rf = evaluate(y_train, random_forest.predict(tfidf_vectors.toarray()))
test_score_rf = evaluate(y_test, rf_test_preds)
```

Performance:

Dataset	Accuracy	Precision	Recall	F1-Score
Train Set	99.67	99.68	99.67	99.67
Test Set	88.64	86.91	88.64	87.06

Logistic Regression

Multinomial Naive Bayes

Bernoulli Naive Bayes

Random Forest Classifier

Models

Hyperparameter Tuning

Cross Validation

UNIVERSITÄT MANNHEIM

- Experimented with models with a few different parameter values.
- Utilized RandomizedSearchCV with StratifiedKFold cross-validation (10 folds) for hyperparameter tuning.

Test Scores for Best Estimators:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83.84	88.84	83.84	85.59
MultinomialNB	86.75	86.38	86.75	86.47
BernoulliNB	87.55	87.67	87.55	87.6
Random Forest Classifier	86.89	88.13	86.89	87.04

Parameter values:

```
# Define classifiers and their respective parameter grids
classifiers = {
  'Logistic Regression': (LogisticRegression(max_iter=1000, class_weight='balanced'), {
      'C': [0.1, 1, 10],
  }),
  'MultinomialNB': (MultinomialNB(), {
      'alpha': [0.1, 0.5, 1.0, 2.0],
      'fit_prior': [True, False]
  }),
  'BernoulliNB': (BernoulliNB(), {
      'alpha': [0.1, 0.5, 1.0, 2.0],
      'binarize': [0.0, 0.5, 1.0]
  }),
  'Random Forest Classifier': (RandomForestClassifier(class_weight='balanced'), {
      'n_estimators': [200, 400, 600, 800],
      'max_depth': [10, 20, 30, 40]
  })
}
```

Observations:

- The models achieved high performance in classifying hate speech and offensive language.
- Bernoulli Naive Bayes achieved the highest F1-score, indicating its effectiveness in this task.

Ensemble Model: Stacking Classifier

UNIVERSITÄT MANNHEIM

```python
from sklearn.ensemble import StackingClassifier

estimators = [(k, v) for k, v in trained_models.items()]
final_estimator = LogisticRegression(max_iter=1000)

ensemble = StackingClassifier(estimators=estimators, final_estimator=final_estimator, cv=5)
ensemble.fit(tfidf_vectors.toarray(), y_train)
ensemble_preds = ensemble.predict(X_test_tfidf.toarray())

# Evaluate the Ensemble model
train_score_en = evaluate(y_train, ensemble.predict(tfidf_vectors.toarray()))
test_score_en = evaluate(y_test, ensemble_preds)
```

Estimators: Trained models from the previous step

- Logistic Regression
- Multinomial Naive Bayes
- Bernoulli Naive Bayes
- Random Forest Classifier

Final Estimator: Logistic Regression

Model Training:

- StackingClassifier aggregates predictions from base estimators and uses Logistic Regression as the final estimator.
- Trained on TF-IDF vectors of text data.

Ensemble Model: Stacking Classifier

UNIVERSITÄT MANNHEIM

Performance on the test data:

Metric	Train Score	Test Score
Accuracy	94.63%	89.43%
Precision	94.6	88.15
Recall	94.63	89.43
F1-Score	94.44	88.4

Observations:

- The ensemble model achieved high performance on both training and test sets, demonstrating its effectiveness in combining the strengths of individual classifiers.
- The ensemble approach enhances the robustness and generalization ability of the model, leading to improved classification accuracy.

RNN Model with Pre-trained GloVe Embeddings

UNIVERSITÄT MANNHEIM

Model Architecture:

- LSTM (Long Short-Term Memory) neural network architecture;
- Utilizes pre-trained GloVe embeddings for word representation

```
LSTMModel(
  (embedding): Embedding(400000, 200)
  (lstm): LSTM(200, 64, num_layers=2, batch_first=True, dropout=0.2)
  (fc_1): Linear(in_features=64, out_features=64, bias=True)
  (fc): Linear(in_features=64, out_features=3, bias=True)
  (dropout): Dropout(p=0.2, inplace=False)
)
```

Performance on the test data:

Metric	Value
Accuracy	86.75%
Precision	85.28
Recall	86.75
F1-Score	85.82

Observations:

- The LSTM model with pre-trained GloVe embeddings achieved competitive performance in classifying hate speech and offensive language.
- Leveraging pre-trained word embeddings enhances the model's ability to capture semantic information from text data.

Fine-tuning DistilBERT Model

UNIVERSITÄT MANNHEIM

Model Architecture: DistilBERT, a lightweight version of BERT (Bidirectional Encoder Representations from Transformers), trained by Hugging Face

Experimental Setup:

- Fine-tuned the DistilBERT model on hate speech and offensive language detection task;
- Utilized the same train-test split as other models for consistency;
- Trained for 3 epochs

Evaluation Results:

Dataset	Metric	Value
Training	Accuracy	90.50%
	F1-Score	90.01
Test	Accuracy	90.32%
	F1-Score	90.09

Fine-tuning DistilBERT Model

UNIVERSITÄT MANNHEIM

Observations:

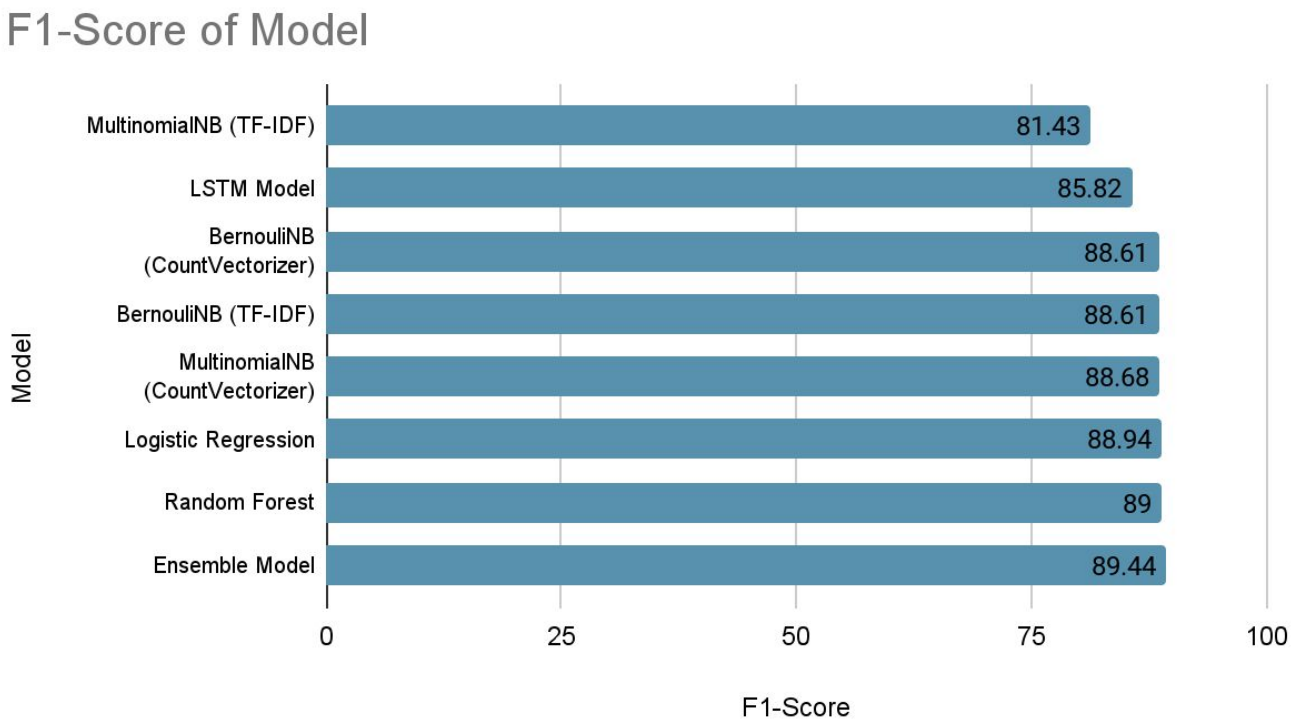
- **The fine-tuned DistilBERT model achieved high accuracy and F1-score on both training and test sets**
- **Leveraging pre-trained language models like DistilBERT can effectively capture complex linguistic patterns in hate speech and offensive language detection tasks**

04 Result Analysis

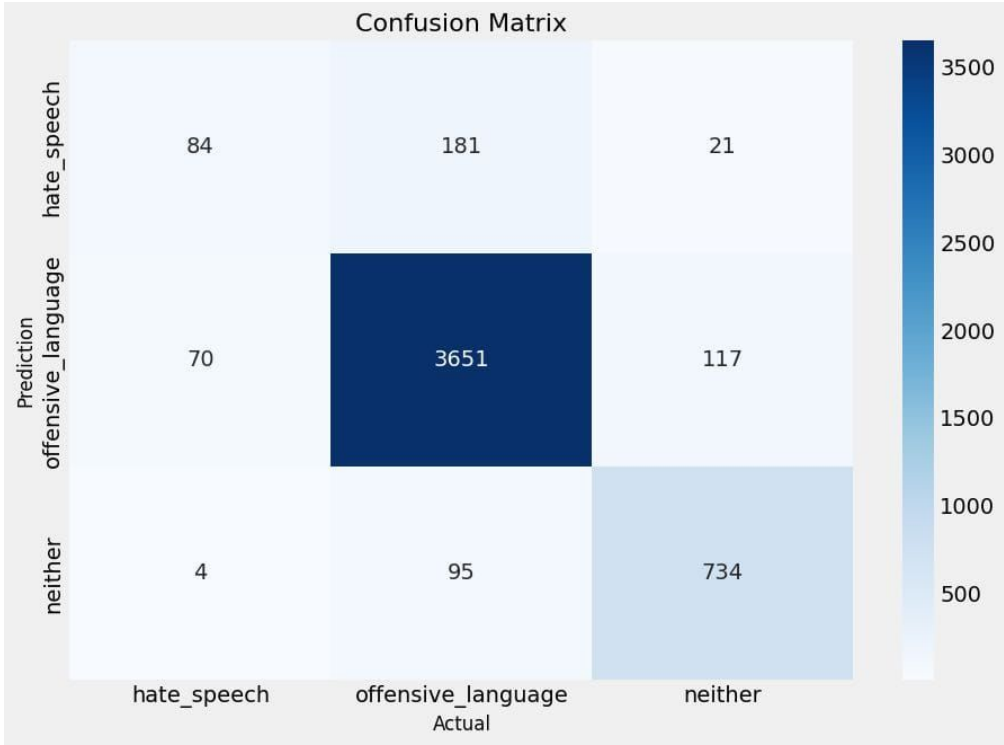
Result Analysis-**Evaluation**

UNIVERSITÄT MANNHEIM

*** Model Performance in F1 Score**



*** Ensemble Model Confusion Matrix**



The BEST Performance(F1 Score: 89.44)

→ **Ensemble Model**

Result Analysis-**Benchmark Comparison**

UNIVERSITÄT MANNHEIM

Nr.	tweet	true_label	pred_label	pred_prob	true_class	pred_class	gpt3.5-turbo-0125	llama3-70b-instruct	GPT True?	llama-3 True?
1	RT @obeyyreggie: stfu hoe RT “@tedddydgaf: "I'm proud to be African American" - Proud African American "I'm proud to be white" - Racist …	0	1	0.791	hate_speech	offensive_language	1	0	0	1
2	@jacksparenohoe let's hang out nigger	1	0	0.591	offensive_language	hate_speech	0	0	0	0
...	I openly admit to being the level of white trash that will drive across town to the gas station with free hot dogs & half price drinks.	2	1	0.402	neither	offensive_language	1	2	0	1
24	Hitler didn't finish it. Can u. If a nigger ur Jew confronts u in the street what then.	0	1	0.616	hate_speech	offensive_language	0	0	1	1
25	RT @StonerBoii2cold: “@TreVaughnLG: Moma said no pussy cats inside my dog house” that's what got bro nem locked inside the dog pound !!	2	1	0.925	neither	offensive_language	1	1	0	0
								Result	8	10

llama 3 exhibited superior predictive accuracy by furnishing correct prognostications for 10/25

GPT 3.5 evinced accuracy in 8/25

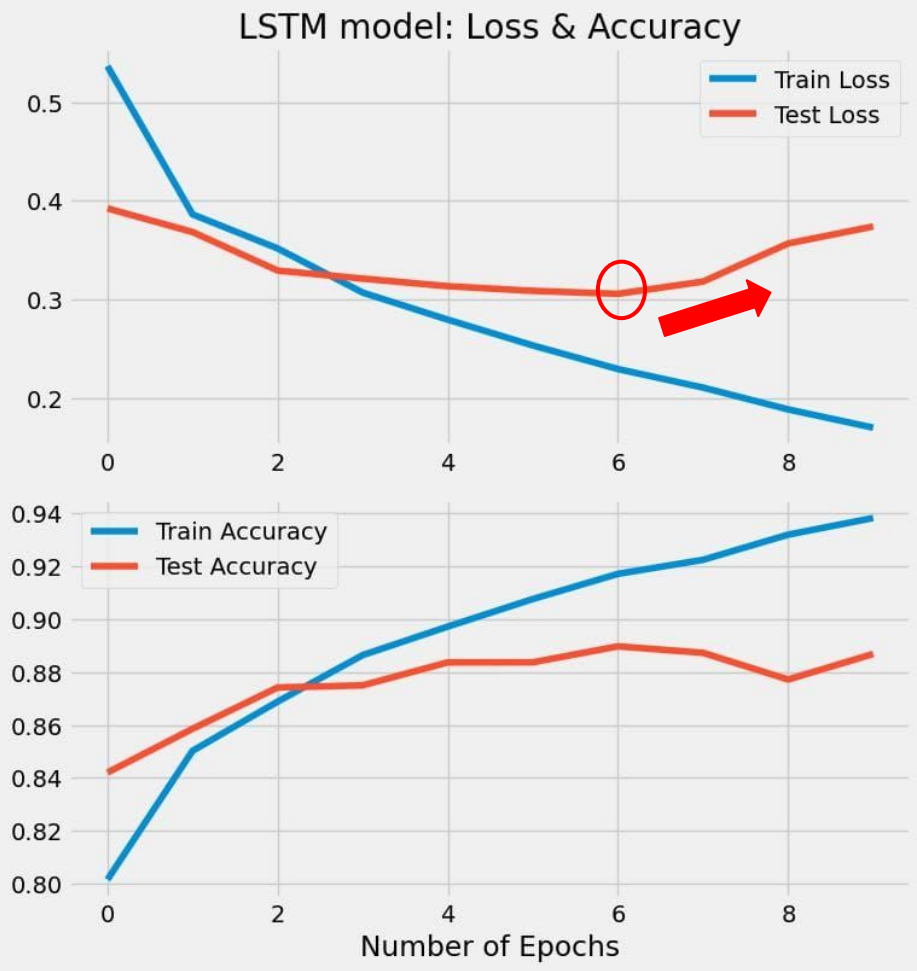
Result Analysis-**Challenge**

UNIVERSITÄT MANNHEIM

Overfitting

We encountered a notable challenge pertaining to overfitting, notably observed in the LSTM Model

The discernible rise in test loss subsequent to the sixth epoch



*** Overfitting in LSTM Model**

Result Analysis-**Challenge**

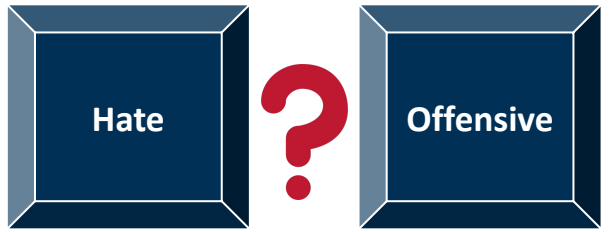
UNIVERSITÄT MANNHEIM

Difficulties in detection （due to similarity)

The subtle differentiations between offensive language and outright hate speech blurred boundaries, complicating the task of identification and classification

*** Sample for Similarity between offensive and hate**

	tweet	true_label	pred_label	pred_prob	true_class	pred_class
189	Lmao let these hoes be hoes ain't no Savin nem	0	1	0.991680	hate_speech	offensive_language
2747	RT @_iHATEMOON: All these bitches & niggaz...	0	1	0.989366	hate_speech	offensive_language
1509	Black bitches don't be kickin up in our school...	0	1	0.988873	hate_speech	offensive_language
2634	@bonnoxxx haha bitch ima draw a webb in bullet...	0	1	0.988282	hate_speech	offensive_language
4880	He ain't shit girl, 💯he a bitch made n...	0	1	0.987566	hate_speech	offensive_language
327	RT "@_ThatGAPeach: & alla my niggas hot bo...	0	1	0.986541	hate_speech	offensive_language
4177	RT @dirtyimage: @Tronkitty not just cause of h...	0	1	0.985885	hate_speech	offensive_language
3407	RT @JHazeThaGod: You other niggas a call up a ...	0	1	0.985340	hate_speech	offensive_language
4590	It's so shady when you bitches talk to guys w/...	0	1	0.985332	hate_speech	offensive_language



UNIVERSITÄT MANNHEIM

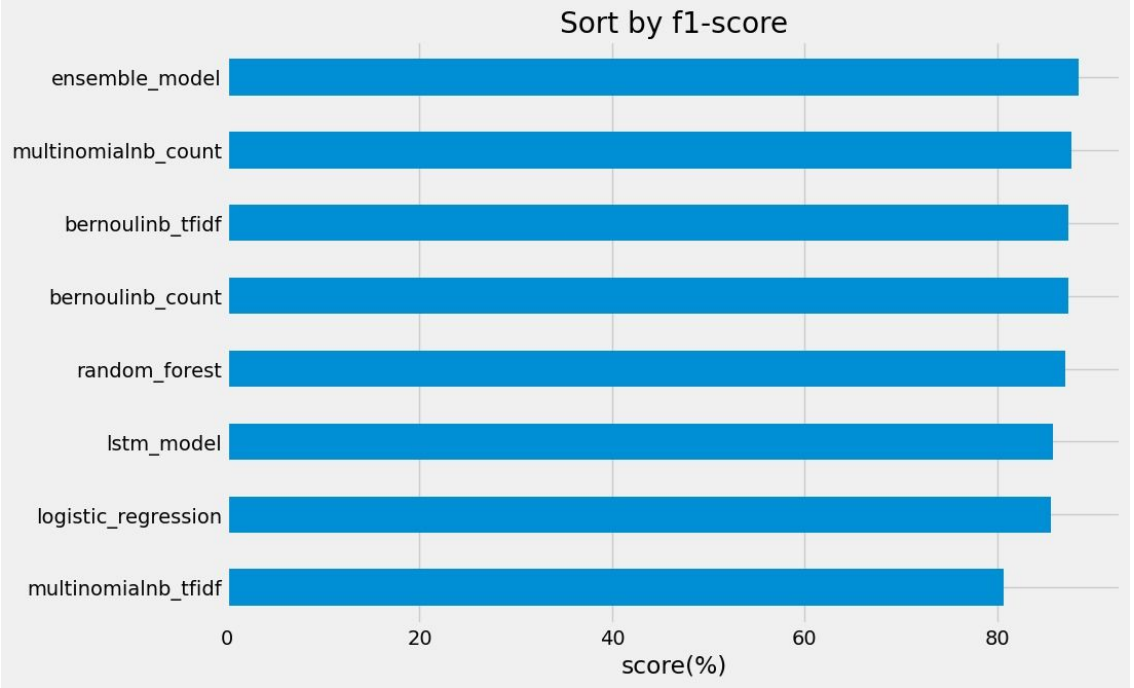
Thanks !

Happy to Questions

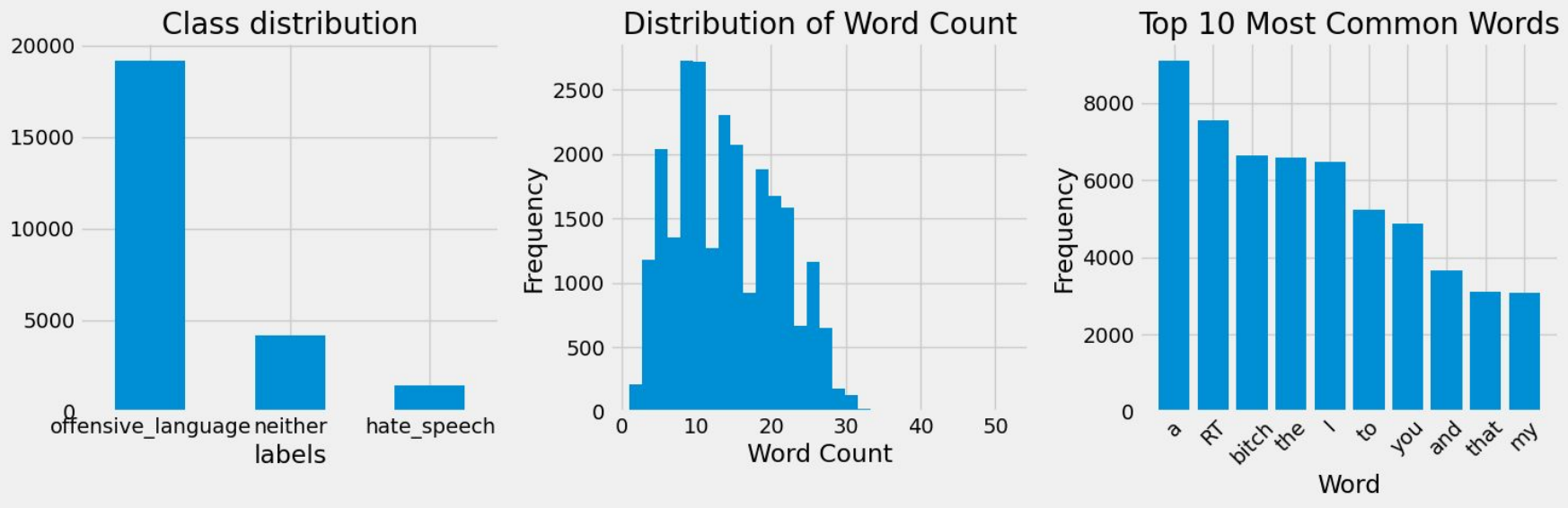
Appendix

Table 1: Data Preprocessing: Operation And Reason

Operation	Reason
Lowercasing	Transforming all text to lowercase ensures uniformity and prevents the model from treating words with different cases as distinct entities.
URL Removal	Eliminating irrelevant information that may not contribute to the analysis.
Number Removal	Eradicating numerical digits from the text, reducing noise and focusing analysis on textual content.
Word Tokenization	Utilizing the NLTK library, the text is tokenized into individual words for futher training.
Handling Emojis	Converting them into text representations to get more information from emojiis.
Stopword Removal	Stop words, which are common words that often do not contribute significant meaning to the text, are removed from the tokenized text.
Lemmatization	Lemmatization reduces words to their base or dictionary form, ensuring that different inflected forms of a word are treated as the same token.



Appendix



Prompt: You are an expert linguist specializing in social media discourse analysis. Your expertise is in identifying and classifying comments on Twitter based on their content and intent. You have developed a nuanced understanding of the differences between hate speech, offensive language, and neutral expressions. Your task is to help classify Twitter comments into one of the following categories:
Hate Speech (label=0): Comments that involve hostility or prejudice against a particular group based on race, ethnicity, nationality, religion, gender, sexual orientation, disability, or similar grounds.
Offensive Language (label=1): Comments that include profanity, vulgarity, or other language that may be considered disrespectful or rude, but do not necessarily target a specific protected group.
Neither (label=2): Comments that do not contain hate speech or offensive language and are generally neutral or benign in nature.
Return only the label (0, 1, or 2) without any explanation.
Please classify the following sentence:

Appendix

Nr.	tweet	true_label	pred_label	pred_prob	true_class	pred_class	gpt3.5-turbo-0125	llama3-70b-instruct	GPT True?	llama-3 True?
1	RT @obeyyreggie: stfu hoe RT “@tedddydgaf: "I'm proud to be African American" - Proud African American "I'm proud to be white" - Racist …	0	1	0.791	hate_speech	offensive_language	1	0	0	1
2	@jacksparenohoe let's hang out nigger	1	0	0.591	offensive_language	hate_speech	0	0	0	0
...	I openly admit to being the level of white trash that will drive across town to the gas station with free hot dogs & half price drinks.	2	1	0.402	neither	offensive_language	1	2	0	1
24	Hitler didn't finish it. Can u. If a nigger ur Jew confronts u in the street what then.	0	1	0.616	hate_speech	offensive_language	0	0	1	1
25	RT @StonerBoii2cold: “@TreVaughnLG: Moma said no pussy cats inside my dog house” that's what got bro nem locked inside the dog pound !!	2	1	0.925	neither	offensive_language	1	1	0	0
								Result	8	10