UNIVERSITAS STUDIORUM MEDIOLANENSIS

Project: Hate speech detection – Use of racism and sexism in German and English language

Course: Text Mining and Sentiment Analysis

Name: Xaver Brückner

MAT: 989819

Xaver.brueckner@studenti.unimi.it

June 2024

1. Introduction

The following report summarizes the analysis and comparison of hate speech usage in German and English using an english Twitter dataset and a reader comment dataset from a German newspaper. The aim of this analysis is building a model to correctly and efficiently solve the multiclass classification problem of flagging text as racist, sexist or neither using supervised learning techniques. As an evaluation metric the F1 score has been chosen, since it works well for imbalanced datasets and can easily be extended from binary to multiclass problems. Furthermore, comparisons have been carried out to see how differently sexism and racism manifests in the German and English language.

2. Dataset description

The foundation of this analysis are two datasets, one consisting of English Tweets and one of reader letters sent as comments to a German newspaper.

RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets (zenodo.org) – Containing 85.000 reader comments with 8,4% of them being labeled as abusive.

thefrankhsu/hate_speech_twitter · Datasets at Hugging Face – With 5679 tweets in the training set including 1516 hate speech and 4163 non-hate speech ones and 1000 tweets in the test set with 500 hate speech and 500 non-hate speech ones.

Both dataset are therefore highly unbalanced with the majority of the data points being labelled as non-hate speech / non abusive. To balance out the English dataset, training and test set have been combined and will be used as one in the subsequent analysis.

2.1 Data Pre-Processing

To better understand the datasets and make them more comparable and usable in terms of hate speech analysis we will take some pre-processing steps. At first columns have been added to both datasets indicating a binary classification for Sexism and Racism in both datasets. In the German dataset a reader comment has been classified as racist/sexist, when at least one annotator has done so. In the English dataset the multiclass Column label has been used to generate the Columns racist/sexist based on the corresponding label type. A Column label / Reject Newspaper has been kept to try binary classification independent of type of hate speech. The final evaluation has been carried out using a third added column “multi”, where the labels of racist and sexist texts have been encoded with 1 & 2 and non-offensive texts with 0.

Further pre-processing steps then included:

- Transforming all text to lower case
- Removing multiple spaces
- Removing all non-word characters (i.e. punctuation)
- Removing all stopwords using English and German stopword lists
- Removing all German Reader comments under 30 and over 200 characters to make them more comparable to tweets

I considered translating the German dataset to English to make the pre-processing with in terms of stopword removal and stemming easier, but in the end wanted to be able to evaluate compare the usefulness of those techniques in another language. And even though adaptions had to be made i.e. stopword lists combined/extended and a different stemmer being used it worked out well in the end.

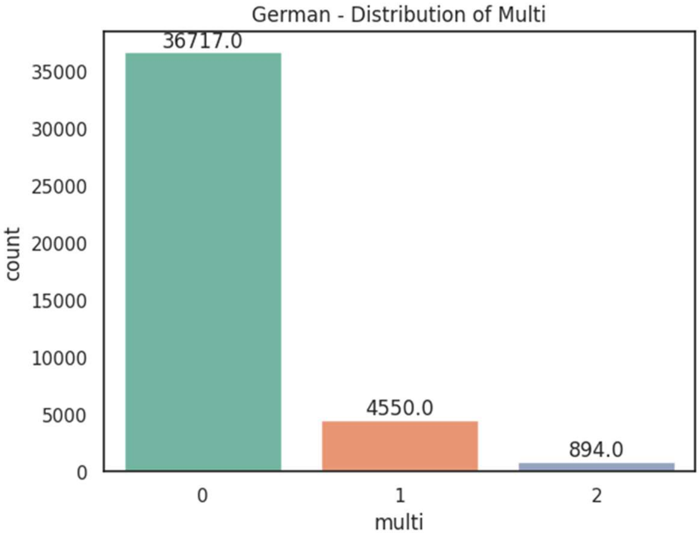
2.2 Data Augmentation

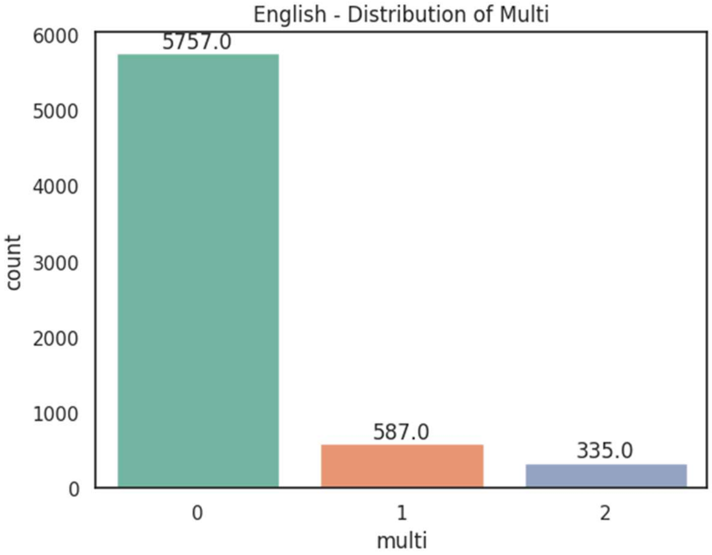
Data scarcity often poses a significant challenge in natural language processing (NLP) problems. The German dataset with its class imbalance, shows exactly this scarcity, concerning observations related to sexism and racism. To address this issue, various techniques have been explored in the past. For this project ultimately, the choice settled on data augmentation, which aims to create additional synthetic training data when the available data is insufficient. To ensure the validity of augmented data, diversity is essential, ensuring model generalization for future tasks. Among the effective methods for augmenting text data is machine translation. This involves translating the original text into other languages and then back-translating it to generate augmented text in the original language. So to combat the class imbalance in the German dataset, data augmentation has been used to synthetically generate additional Sexist and Racist training data. In this case the reader comments have been translated to 5 different languages [English, Spanish, Portuguese, French, Italian] and then translated back into German.

Final Datasets

The resulting Datasets had the following distributions:

The distribution of both datasets is still very unbalanced but significantly improved compared to the baseline. The non-hate speech group is by far the biggest for both datasets, followed by the racism and lastly sexism.





2.3 Stemming

Furthermore a Stemmer has been used on both datasets to make it easier to use them in the models. Stemming is a technique that simplifies words by removing prefixes and suffixes, transforming them into their fundamental or root form.

For example: “bikes” becomes “bike” “betrayal” becomes “betray”. Stemming is helpful for NLP pipelines with tokenized words obtained from dissecting a document.

It helps a lot in text normalization, by making it easier to process and analyze text. For the English dataset a PorterStemmer has been used and for the German one Cistem, which was the best performing one for German text.

2.4 Vectorization

Both texts of the datasets have then been TF-IDF (Term Frequency-Inverse Document Frequency) vectorized to make them processable for the machine learning models.

TF-IDF vectorization is a technique used to represent existing text documents as numerical feature vectors. The individual terms can be summarized as follows:

Term Frequency (TF):

- Measures how often a term appears in a text document.
- Calculated as the number of occurrences of a term divided by the total number of terms in the document.

$$TF(t,d) \ = \frac{total\ number\ of\ terms\ in\ document}{numer\ of\ times\ term\ t\ appears\ in\ document}$$

Inverse Document Frequency (IDF):

- Measures the rarity of a term across all documents in the collection.
- Helps give more weight to terms that are less common.

$$IDF(t) \ = \ln\left(\frac{total\ number\ of\ documents}{number\ of\ documents\ containing\ term\ t}\right)$$

TF-IDF Score:

- Combines both TF and IDF to create a feature vector for each term in a document.
- The TF-IDF score for a term in a document is the product of its TF and IDF values:

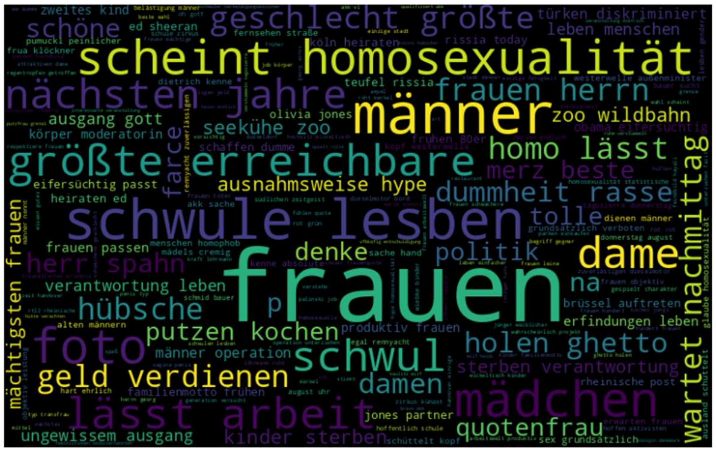
$$TF\_IDF(t,d) = TF(t,d) \times IDF(t)$$

3. Wordclouds

To understand how sexism and racism manifest differently across language and culture i.e. Germans versus American dominated Twitter space, we analyse the Word clouds of the text which were labelled sexist and racist respectively. It must be considered that a non-negligible part of the differences could also be attributed to the difference in audience of Twitter users and Newspaper reader / people who send reader comments to newspapers.

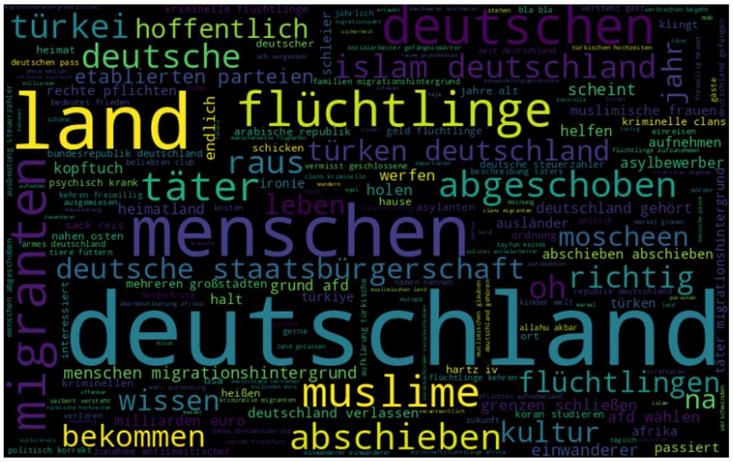
German - Sexism

The German Word cloud for sexism frequently mentions very prevalent words in the context of Sexism but surprisingly also a lot connected to homophobia as Frauen = Women, Männer = Men, Schwule & Lesben = Gays & Lesbians, scheint homosexualität = seems homosexual, cleaning & cooking, body [of] moderator, quota women, letting [us] work, waits [all] afternoon, dumbness race (as in women are the dumber gender) and jealous. This hints at sexism rooted in the perception that women are reduced to their looks and their work as housewives. Sexism in this case is not expressed through outright insults but more so opinions of inferiority and superficiality in different areas of life.



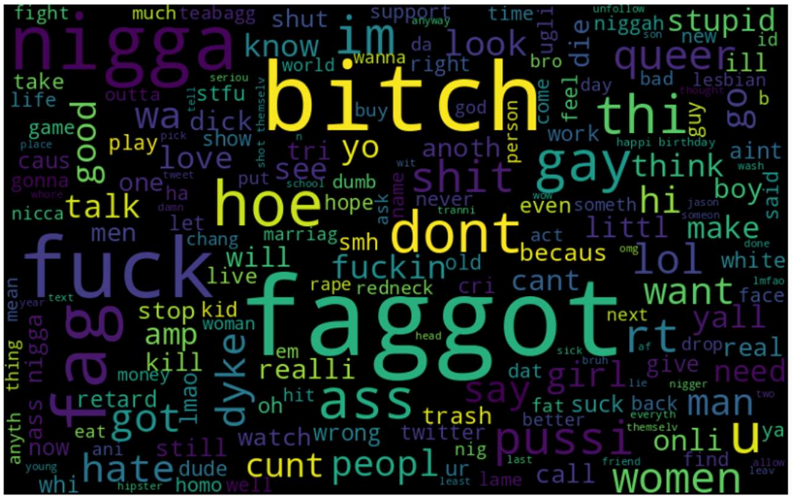
German – Racism

For racism on the other hand the focus lies on words like Deutschland = Germany, Flüchtlinge = Refugees, Einwanderer = Immigrants and more terms related to immigration from predominantly muslim countries as Turks, Islam, Headscarf and asylum. Related problems are also mentioned with the words criminal, perpetrator and money. The blame seems to go towards politics by mentioning Merkel, Grüne (green party), politicians, laws and police.



English – Sexism

The English word cloud for sexist tweets shows more explicit language and curse words directed towards women like Bitch, hoe, cunt , but also addressing sexuality in a negative way like the German dataset with words like faggot, fag, gay and dyke.



English – Racism

The is very cantered around discrimination of African Americans with shown by the very central term nigger in various ways of writing. The focus on white/black division is underlined by the frequent use of the two terms and more derogatory ones of describing either like crackers, white trash, whitey, coon and monkey.

end
right
white peopl
store
live
smell weird
lmao
america
tcot
everi
hi
shit
ass
nigga
aint
music
look
ugli
hey
today
pussi
never
cant
jew
refuge
thi hard
wit
onli
call
babi
coon
know
hoe
becaus
retard
ghetto
work
day
fuck
bad
one
real
darki
hood
happi
peopl
trash
immigr
away
show
mad
gyall
kill
dont
lol
foreign
boy
mean
negro
hang
wrong
ha
racist
time
redneck
still
uncl tom
got
talk
chicken
white
ask
alway
cuz
wetback
stupid
around
race
man
im
full
love
cracker
play
nation
da
school
losangel
think
will
isnt
make
ignor
say
whi
whitey
damn
countri
black
speak
wa
american
watch
dumb
even
dem
shoot
nigger
move
want
nigguh
amp
word
hous
activ
new
lot
spic
nicca
tri
take
color
thi
nig
ani
need
friend
way
mani
head
kid
someon
hate
disgust
gook
wash
happen
niggah
theyr
good
state
cop
back
go
ur
bitch
tell
trash
well
realli
see
discrimin
chink
stop
come
monkey
rt
die
ass nigga
shot themselv
dat
follow
shut african
u
beaner
white trash

4. Prediction

For prediction binary classification was first used but then switched to multiclass, containing 1 for a racist tweet/comment, 2 for a sexist one and 0 if it was neither.

As models Logistic Regression, Random Forest Classifier, Gradient Boost Classifier and SVM have been used on the TF-IDF Vectorized data frames.

5. Results

For simplicity purposes only the macro average F1 score is shown for each model in the following. A complete list of model performance metrics and confusion matrices is shown in the appendix. The used metrics are calculated as follows:

$$Precision \ = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 \ = \frac{2 * Precision * Recall}{Precision + Recall}$$

F1 Score Macro Average for both Datasets

Model	F1 Score – German	F1 Score - English
Logistic Regression	0.71	0.65
Random Forest	0.90	0.70
Gradient Boost	0.80	0.76
SVM	0.87	0.72

In this case the best performing model for the German dataset was the Random Forest one with a F1 Score of 0.90 and for the English dataset it was the Gradient Boosting Classifier with a F1 score of 0.76.

Gradient Boosting was less successful in the German dataset due to a high rate of false positives in the racism class. The false positive rate for the Gradient Boost Classifier was also relatively high in the English dataset, but less severe than in the German one.

6. Conclusion

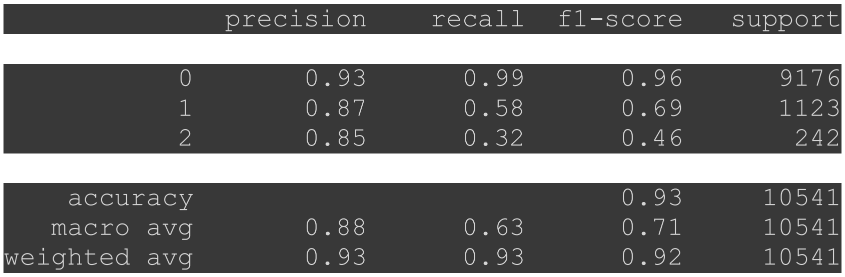
The goal of this analysis was to compare the use of German and English language in sexism and racism and to build a model to correctly flag instances of such use. Through pre-processing and augmentation techniques a F1 score of 0.90 for the German data set and 0.76 for the English dataset could be achieved. This in itself is a very good result, especially for the German dataset, but needs to be taken with a grain of salt considering the synthetically generated data shows higher similarity and to the existing one and will therefore be easier to classify by a model resulting in a better F1 score.

The comparison of language used shows that in German the insults were used much less than in the English dataset and sexism and racism were expressed more subtly as concepts or latently. A limitation of this observation is that it also has to do with the nature of the data set and the socioeconomic background of the authors of the texts. Whereas the English one consisted of Tweets by a variety of different people with the anonymity of the internet, the German one did of comments to a newspaper written as letters or E-Mails.

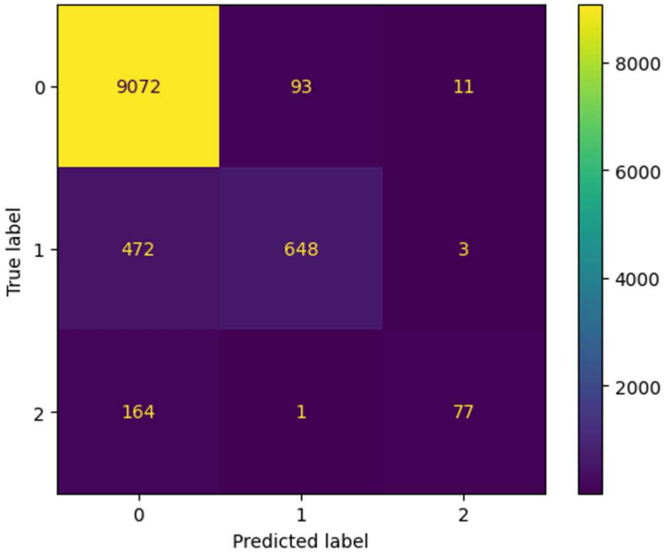
7. Appendix

German Dataset

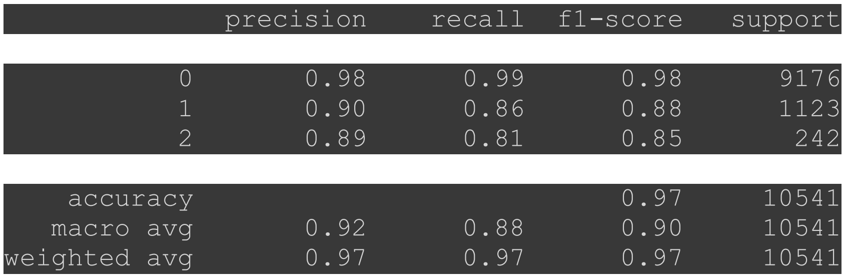
Logistic Regression



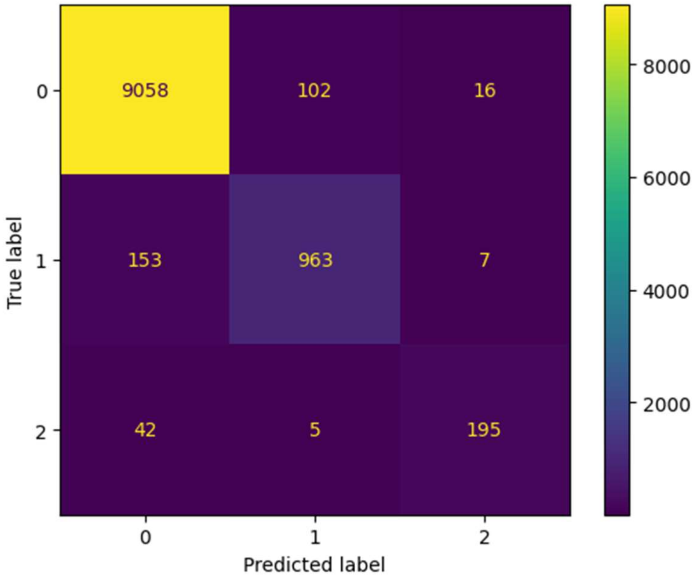
```
              precision    recall  f1-score   support

           0       0.93      0.99      0.96      9176
           1       0.87      0.58      0.69      1123
           2       0.85      0.32      0.46       242

    accuracy                           0.93     10541
   macro avg       0.88      0.63      0.71     10541
weighted avg       0.93      0.93      0.92     10541
```



Random Forest

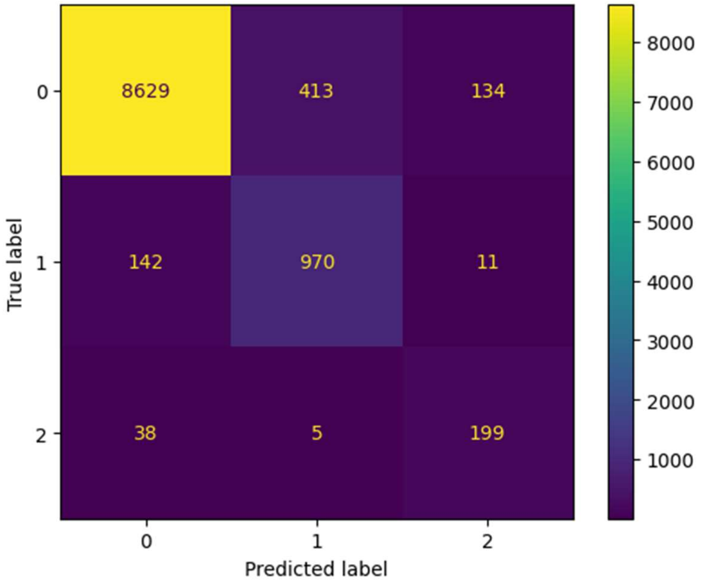


```
              precision    recall  f1-score   support

           0       0.98      0.99      0.98      9176
           1       0.90      0.86      0.88      1123
           2       0.89      0.81      0.85       242

    accuracy                           0.97     10541
   macro avg       0.92      0.88      0.90     10541
weighted avg       0.97      0.97      0.97     10541
```



Gradient Boosting

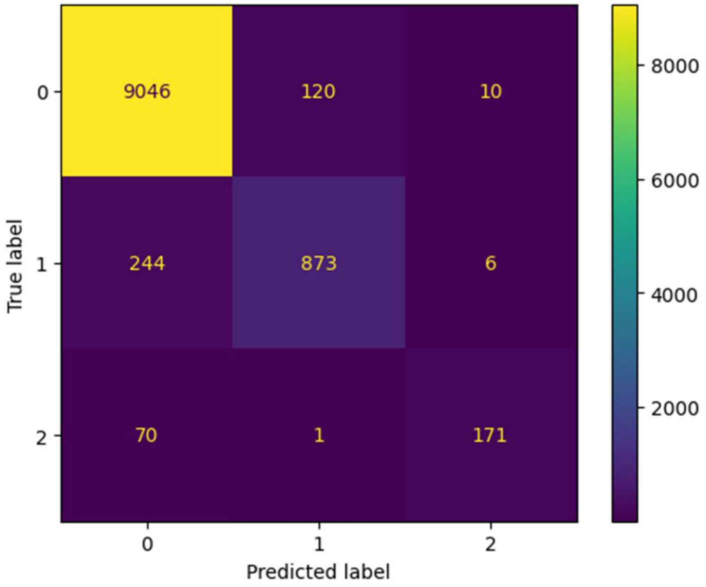
```
              precision    recall  f1-score   support

           0       0.98      0.94      0.96      9176
           1       0.70      0.86      0.77      1123
           2       0.58      0.82      0.68       242

    accuracy                           0.93     10541
   macro avg       0.75      0.88      0.80     10541
weighted avg       0.94      0.93      0.93     10541
```



SVM

```
              precision    recall  f1-score   support
```

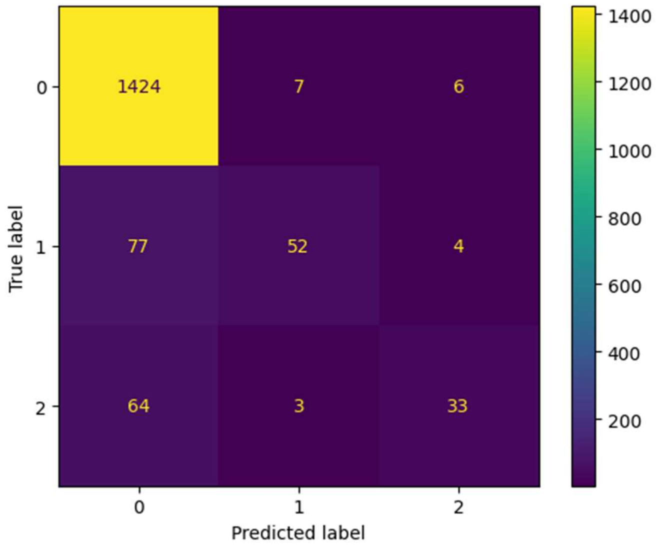
```
           0       0.97      0.99      0.98      9176
           1       0.88      0.78      0.82      1123
           2       0.91      0.71      0.80       242

    accuracy                           0.96     10541
   macro avg       0.92      0.82      0.87     10541
weighted avg       0.96      0.96      0.96     10541
```



English Dataset

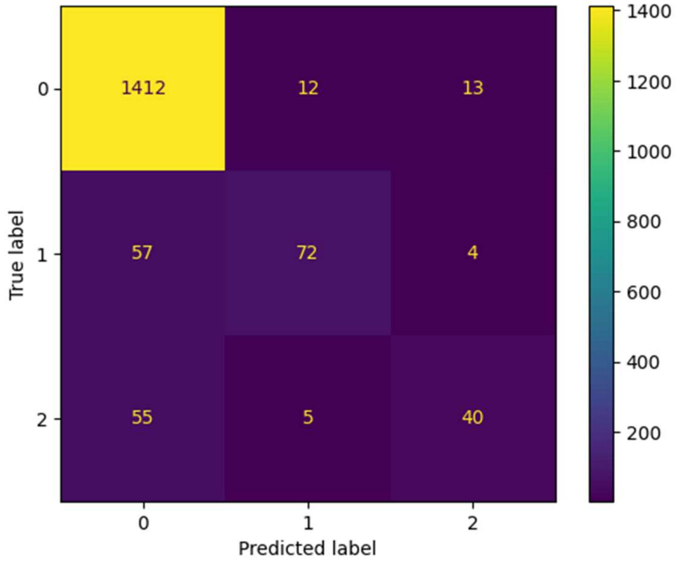
Logistic Regression

```
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      1437
           1       0.84      0.39      0.53       133
           2       0.77      0.33      0.46       100

    accuracy                           0.90      1670
   macro avg       0.84      0.57      0.65      1670
weighted avg       0.90      0.90      0.89      1670
```



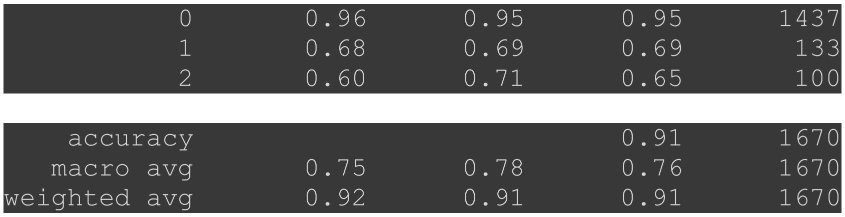
Random Forest

```
              precision    recall  f1-score   support

           0       0.93      0.98      0.95      1437
           1       0.81      0.54      0.65       133
           2       0.70      0.40      0.51       100

    accuracy                           0.91      1670
   macro avg       0.81      0.64      0.70      1670
weighted avg       0.90      0.91      0.90      1670
```

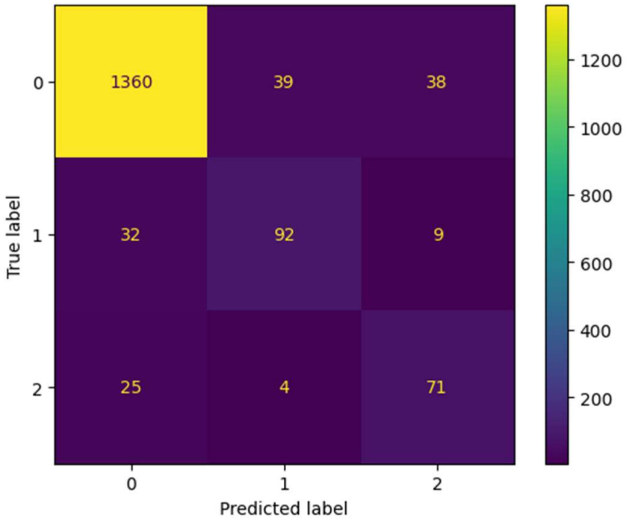


Gradient Boosting

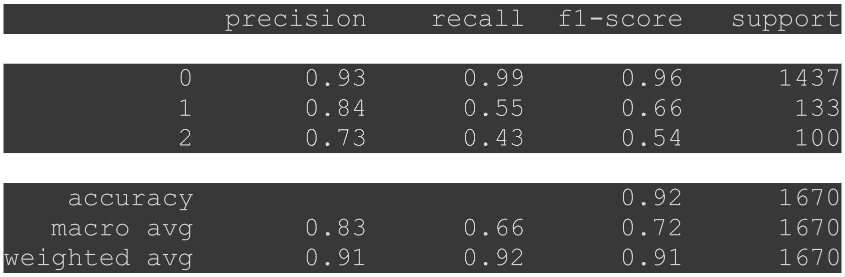
```
              precision    recall  f1-score   support
```



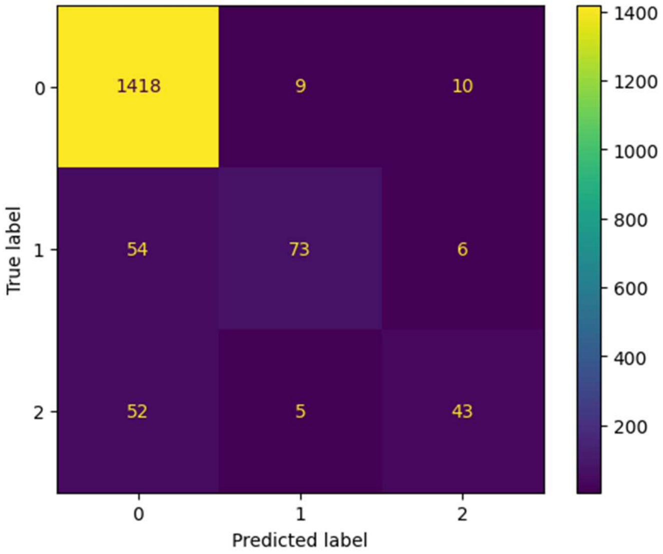
```
           0       0.96      0.95      0.95      1437
           1       0.68      0.69      0.69       133
           2       0.60      0.71      0.65       100

    accuracy                           0.91      1670
   macro avg       0.75      0.78      0.76      1670
weighted avg       0.92      0.91      0.91      1670
```



SVM



```
              precision    recall  f1-score   support

           0       0.93      0.99      0.96      1437
           1       0.84      0.55      0.66       133
           2       0.73      0.43      0.54       100

    accuracy                           0.92      1670
   macro avg       0.83      0.66      0.72      1670
weighted avg       0.91      0.92      0.91      1670
```


1418
9
10
54
73
6
52
5
43
0
1
2
True label
0
1
2
Predicted label
1400
1200
1000
800
600
400
200