TOXICITY DETECTION

Tharun Komari, Qiji Xiang, Amaya Hijazi, Karl Juhl

01.

The Problem

An overview on the topic

02.

Modelling Process

Steps in the modeling process and project design

03.

Results

Evaluation of predictions

04.

CONCLUSIONS

The takeaway

01.

The Problem

111,350,250

US adults reduced internet usage after receiving abuse.

OVER 1/3

of the US population.

(Johnson, 2021)

Related Work

(Davidson, 2017)

Investigating hate speech within online comments

Origin of data set used

(Van Aken, 2018)

Error Analysis of toxic comment detection systems

Baseline

Modelling Process

02.

Data

01

24,802 tweets labelled hate speech, offensive language, or neither

02

F1 Score and Accuracy

03

Preprocessing and Modeling

Preprocessing

A

Removed usernames

B

Removed Punctuation

C

Stemming

D

Removed Numbers

F

Tokenization

G

Bag-of-Words word embedding

Supervised Learning

1

Logistic Regression

Multinomial configuration of classic logistic function

2

Linear SVC

A support vector classification model with a linear kernel

3

Decision Tree Classifier

Tree-like model of outcomes

4

Extra Tree Classifier

Aggregate result of many de-correlated trees

Supervised Learning

5

Random Forest Classifier

Aggregate result of many de-correlated trees

6

Ridge Classifier

Converts target values between [-1,1] and treats it as a regression to predict

7

Gradient Boosting Classifier

Combine weak learning models together in forward stage-wise order

Deep Learning

8

MLP Classifier

A multi-layer perceptron from sklearn

9

Neural Network

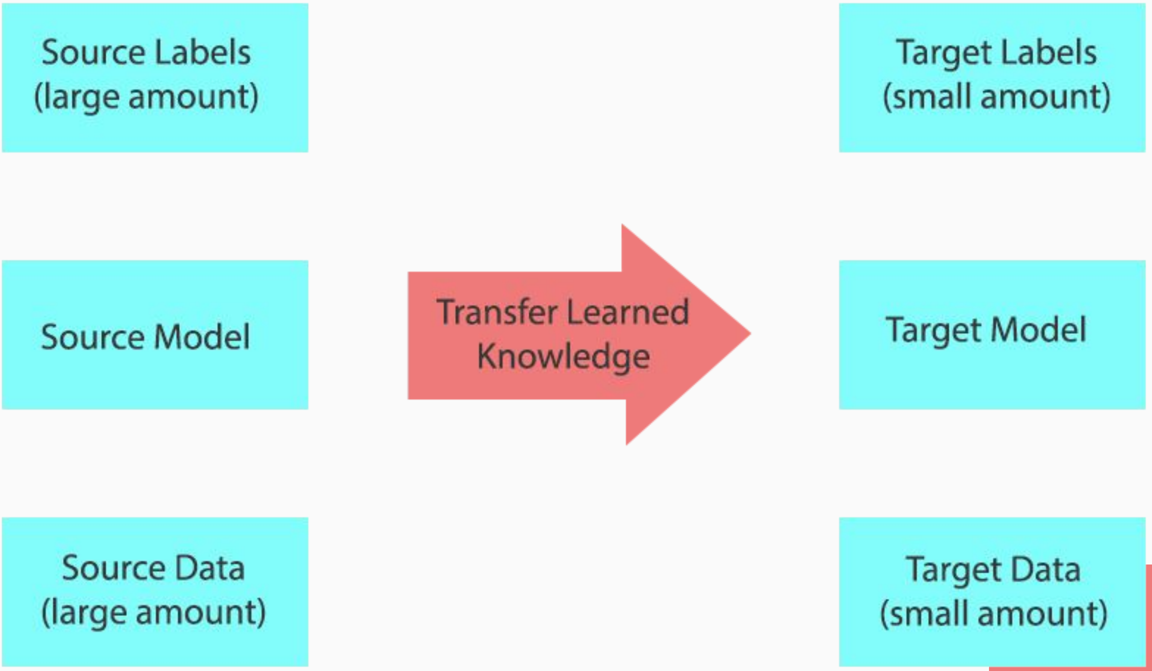
Embedding Layer, with 2 dense layers with relu activation function, dropout , a output dense layer with a softmax

10

FastAi ULMFiT

AWD-LSTM

Transfer Learning



FastAi ULMFit

Pretrained Model → Fine Tuning to our Dataset → Text Classifier

Results

03.

Supervised Modelling Results

Model	Accuracy	F1-Score	Recall	Precision
Logistic Regression	0.90	0.89	0.90	0.89
Linear SVC	0.89	0.88	0.89	0.87
Decision Tree	0.88	0.88	0.88	0.88
Extra Tree	0.83	0.83	0.83	0.83
Random Forest	0.89	0.88	0.88	0.87
Ridge	0.86	0.75	0.74	0.80
Gradient Boosting	0.87	0.76	0.70	0.83

Deep Learning Modelling Results

Model	Accuracy	F1-Score	Recall	Precision
MLP Classifier	0.86	0.77	0.75	0.80
Neural Network	0.81	0.80	0.81	0.80
ULMFiT	0.90	0.90	-	-

Conclusions

04.

CONCLUSIONS

01

The best performing model is FastAI's ULMFiT with a F1 score of 0.8961 and an accuracy score of 0.9029

02

The second best performing model is Logistic Regression with a F1 score of 0.8915 and an accuracy score of 0.9008

LIMITATIONS AND FUTURE WORK

LIMITATIONS

- The model is slightly biased to predict hate speech
- The model might run into problem with new words

FUTURE WORK

- Expand the dataset from different sources and implement more advanced models
- Explore different ways to vectorize the dataset
- Attempt stratified k folds for imbalanced label counts

Thank You!