Hate speech detection

[source]

16 year high

Personal attacks motivated by bias or prejudice reached a 16-year high in 2018 reported by the F.B.I.

[source]

Problem Statement

In this digital age, online hate speech has increased over the past few years. Studies has shown that online hate speech can lead to offline violence towards a certain group. [1]

In some cases, social media can lead a more direct role, in this case the New Zealand shooting incident was broadcasted live on Facebook. [2]

Due to the societal concern and how widespread hate speech is becoming on the Internet and especially on social media, there is a strong need to classify online hate speech comments that are considered hate speech. [3]

What is Hate Speech?

- Hate speech is speech that attacks a person or a group on the basis of protected attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity.

[source]

Types of hate speech

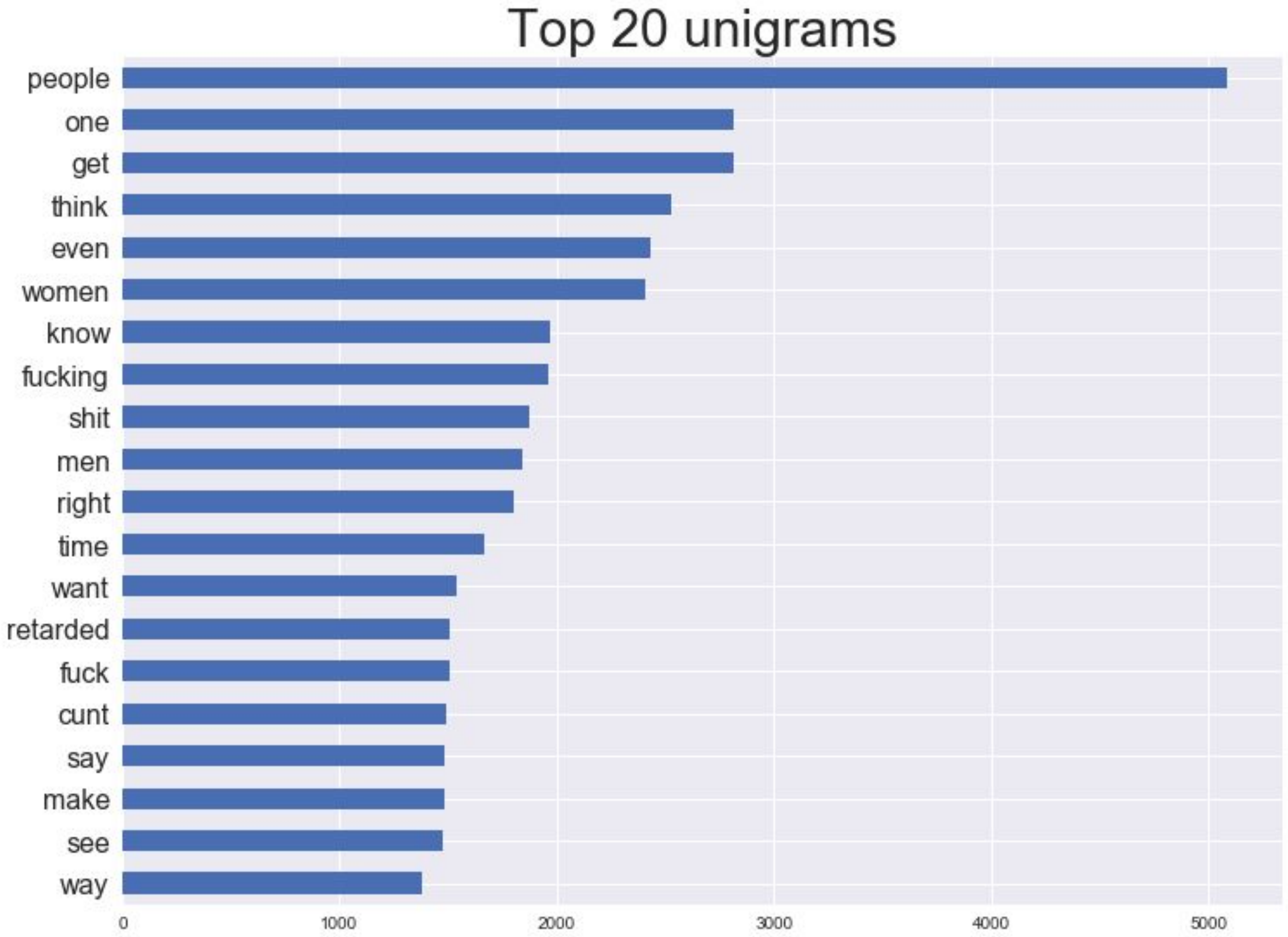
1. Misogyny → aimed at women
2. Misandry → aimed at men
3. Racism → aimed at a specific race
4. Sexual orientation
5. Religion
6. Disability
7. Ethnic origin

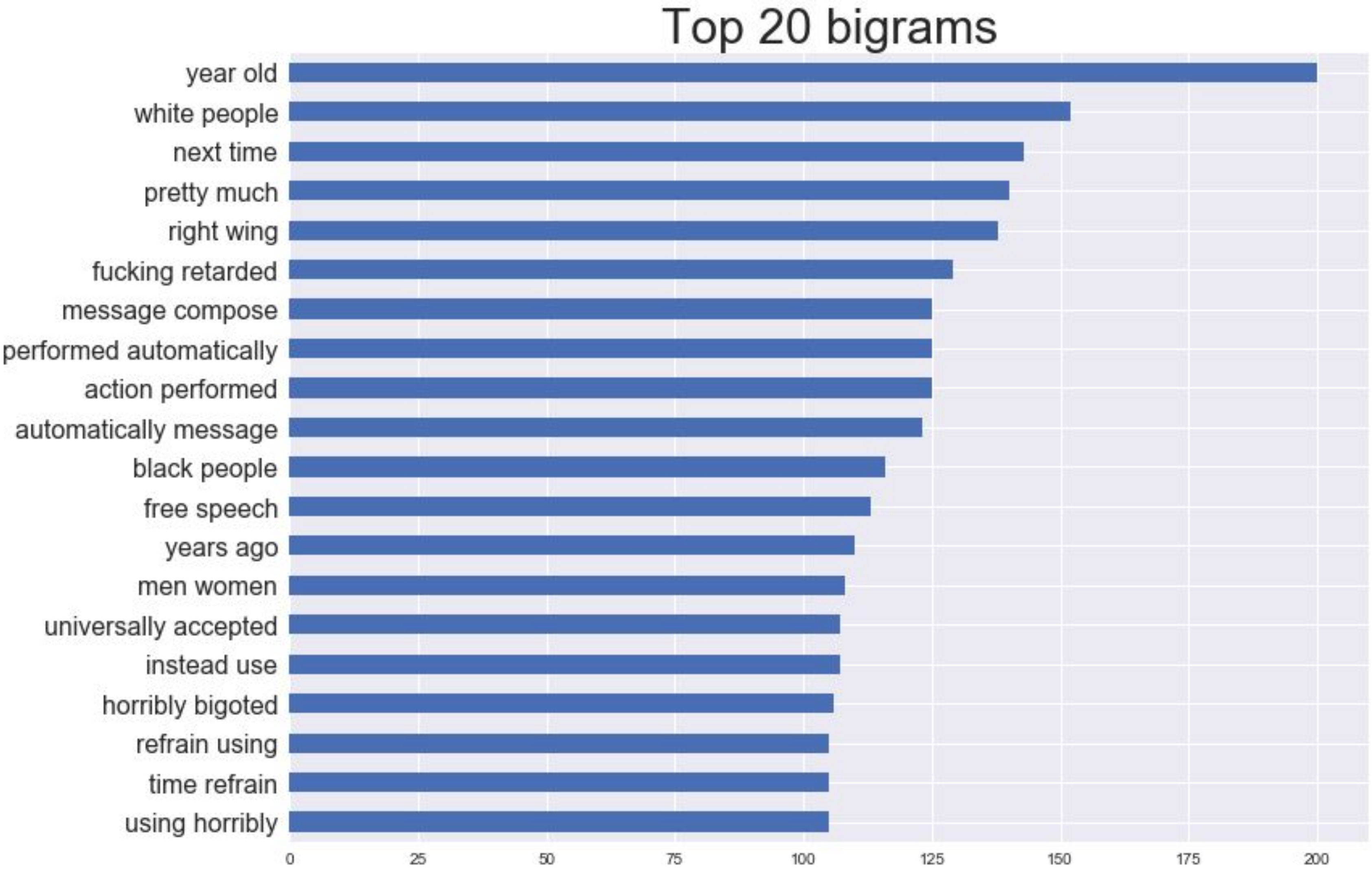
[source]

EDA

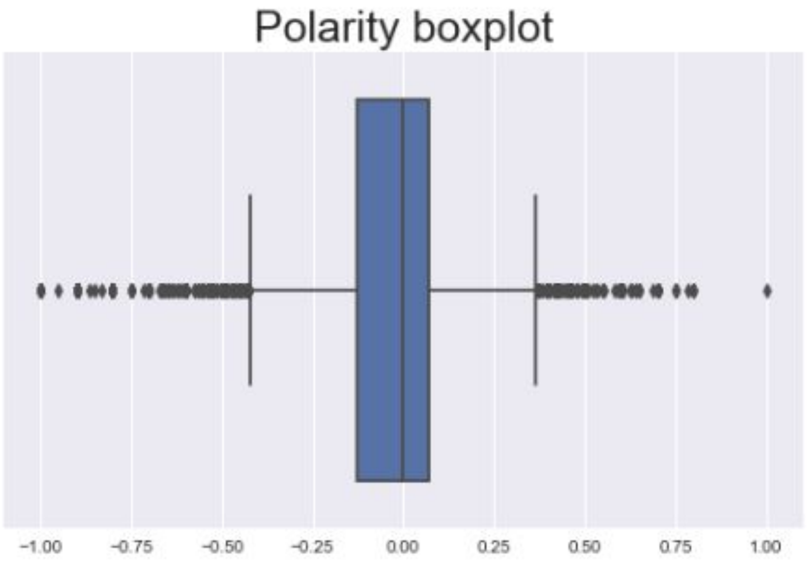
Disclaimer: You may find the following slides having offensive content.

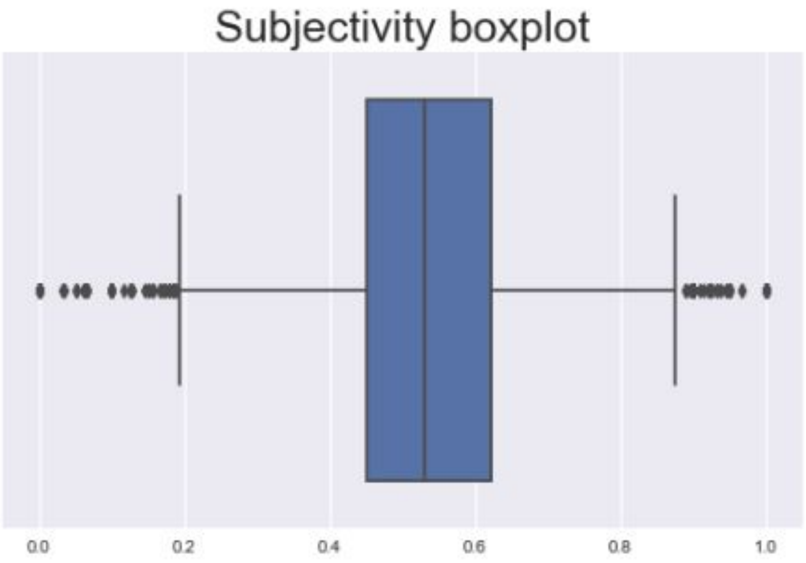

shit
think
fuck
right
women
say
know
one
even
fucking
mean
cunt
see
want
people
get
make
way
thing
men


Top 20 unigrams
people
one
get
think
even
women
know
fucking
shit
men
right
time
want
retarded
fuck
cunt
say
make
see
way
0
1000
2000
3000
4000
5000


Top 20 bigrams
year old
white people
next time
pretty much
right wing
fucking retarded
message compose
performed automatically
action performed
automatically message
black people
free speech
years ago
men women
universally accepted
instead use
horribly bigoted
refrain using
time refrain
using horribly
0
25
50
75
100
125
150
175
200

Polarity and Subjectivity





Modelling

How my model will help

- It is to aid the moderator in efficient hate speech detection
- It is not to replace the moderators job

Two approaches of Modelling

- Machine learning
    - BOW model
- Deep learning

Vectorizers used

- TF-IDF
- Count Vectorizer
- Word2Vec
    - Trained on dataset
    - Pre-trained by Google

Vectorizers used

- TF-IDF
- **Count Vectorizer**
- Word2Vec
    - Trained on dataset
    - Pre-trained by Google

Model Comparison - Machine Learning

Model	Test F1
Logistic Regression	87.9%
Multinomial NB	86.03%
Extra Trees Classifier	81.52%
SVM Classifier	86.28%
Random Forest Classifier	86.3%

Model Comparison - Machine Learning

Model	**Test F1**
Logistic Regression	**87.9%**
Multinomial NB	86.03%
Extra Trees Classifier	81.52%
SVM Classifier	86.28%
Random Forest Classifier	86.3%

Model Comparison - Machine Learning

Model	**Test F1**	**Test recall**
Logistic Regression	**87.91%**	**87.35%**
Logistic Regression with Balanced Class	87.64%	87.33%

Deep Learning Models

Model	Word Embeddings
LSTM with 8 Dense layer (LSTM 1)	Pre-trained Word embeddings
LSTM with 32 Dense layer (LSTM 2)	Pre-trained Word embeddings
LSTM	Word embeddings on Dataset
CNN + LSTM	Word embeddings on Dataset
BERT	Pre-trained Word embeddings

BERT - Bidirectional Encoder Representations for Transformers

Consists of:

- Transformer layer
    - an attention mechanism that learns contextual relations between words (or sub-words) in a text
    - the Transformer encoder reads the entire sequence of words at once, learning context left and right of a word

BERT - Bidirectional Encoder Representations for Transformers

Learning context using two strategies:

1. Masking
   - Mask 15% of the words in the input, run the entire sequence through a deep bidirectional Transformer encoder, and then predict only the masked words, based on context of non-masked words.

```
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
```

[source]

BERT - Bidirectional Encoder Representations for Transformers

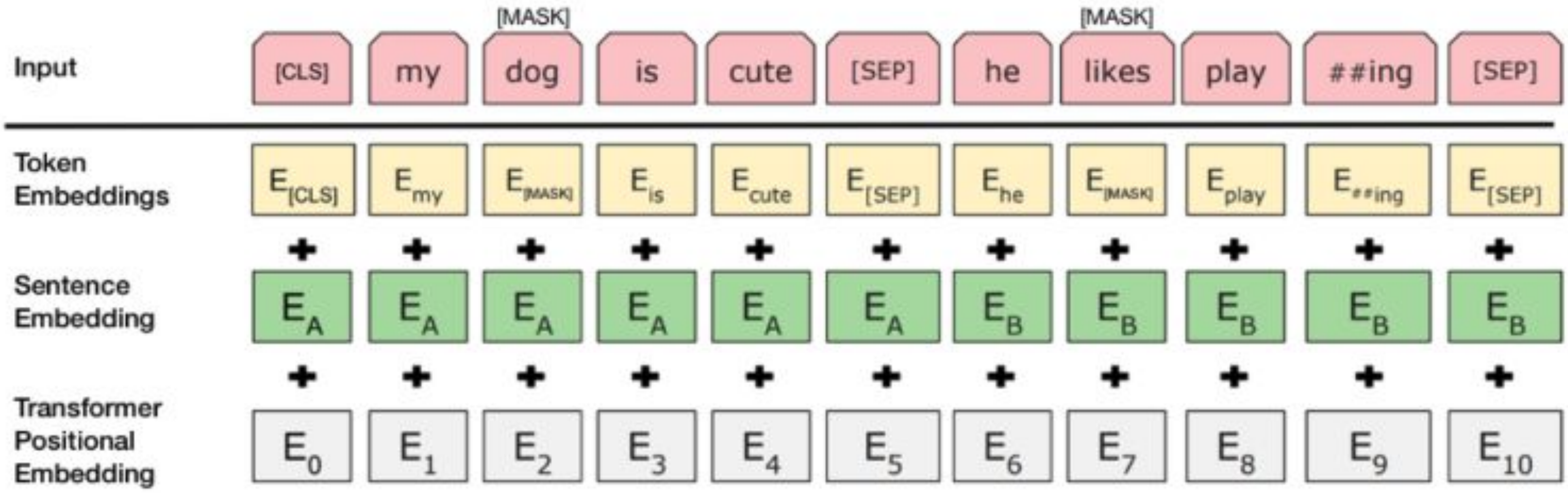
Learning context using two strategies:

2. Next-Sentence Prediction
   - 50% of input are pairs if subsequent sentences and the other 50% a pair of sentences randomly selected from the corpus.
   - Predicts if next sentence is indeed the next sentence

```
Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence
```

```
Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence
```

BERT - Bidirectional Encoder Representations for Transformers

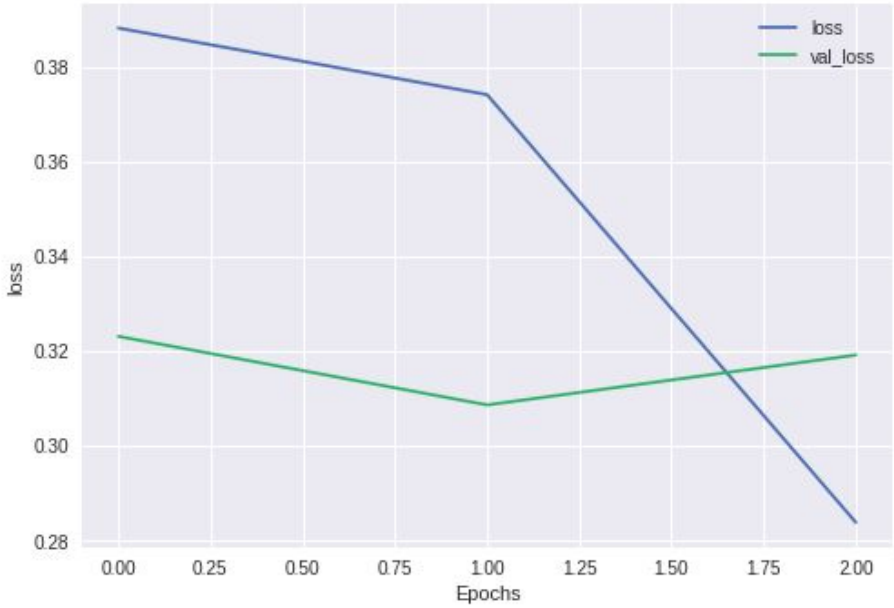


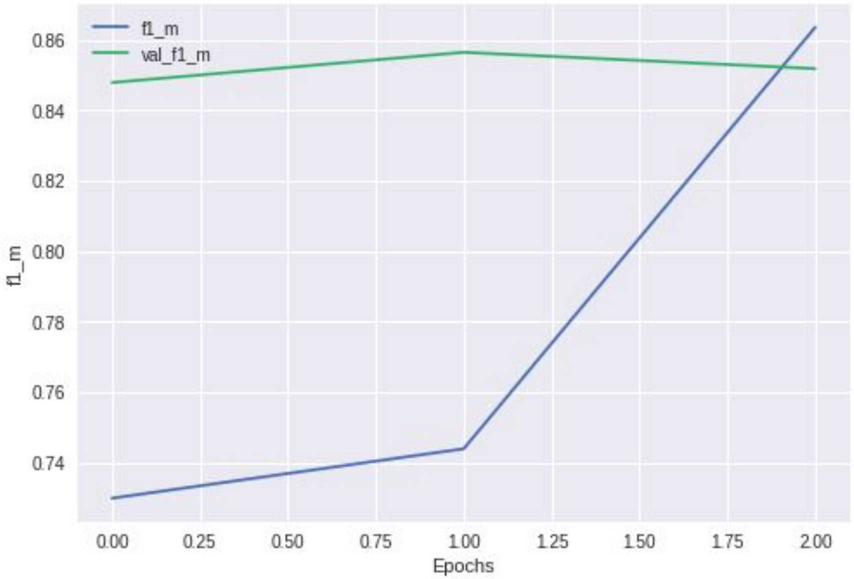
[source]

Model Comparison - Deep Learning

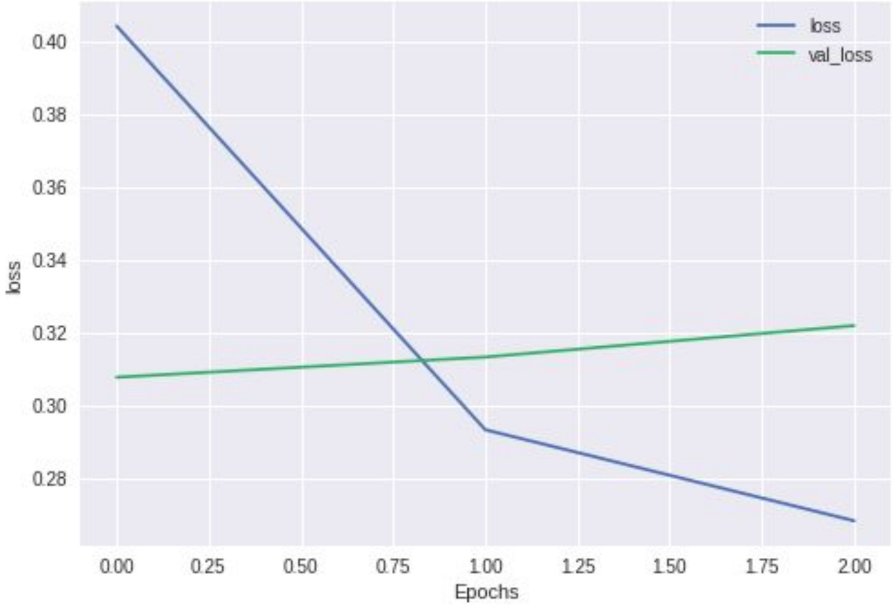
Model	Word Embeddings	Test F1	Epoch
LSTM with 8 Dense layer (LSTM 1)	Pre-trained Word embeddings	81.75%	4
LSTM with 32 Dense layer (LSTM 2)	Pre-trained Word embeddings	84.36%	10
LSTM	Word embeddings on Dataset	85.18%	3
CNN + LSTM	Word embeddings on Dataset	84.95%	1-2
BERT	Pre-trained Word embeddings	**87%**	2

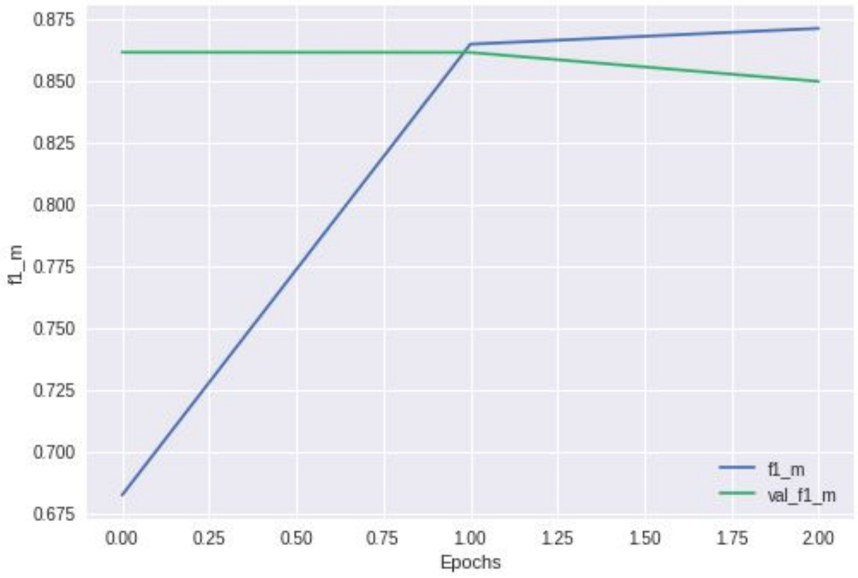
LSTM Word2Vec on Dataset



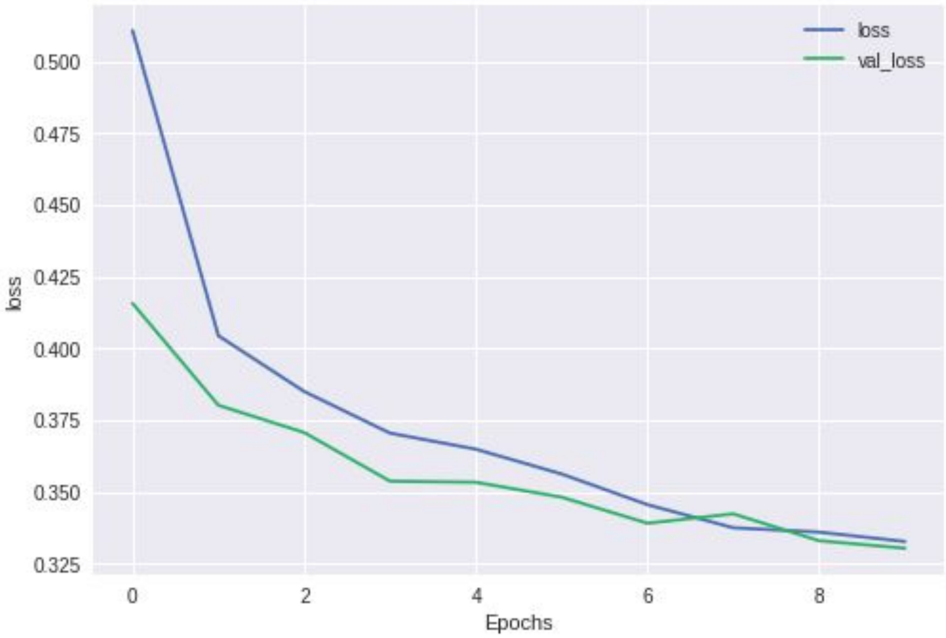


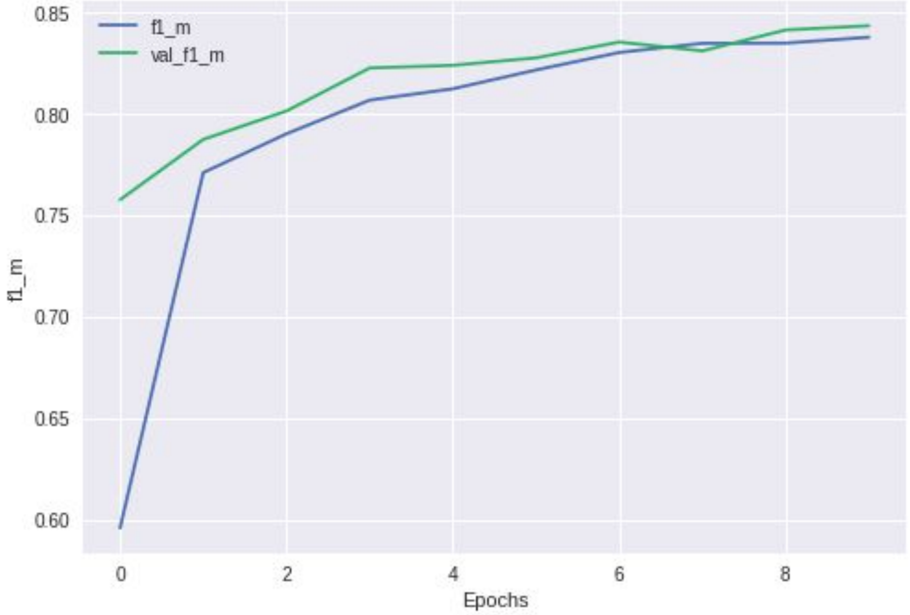
CNN + LSTM Word2Vec on Dataset





Pre-trained LSTM - 32-neuron Dense layer



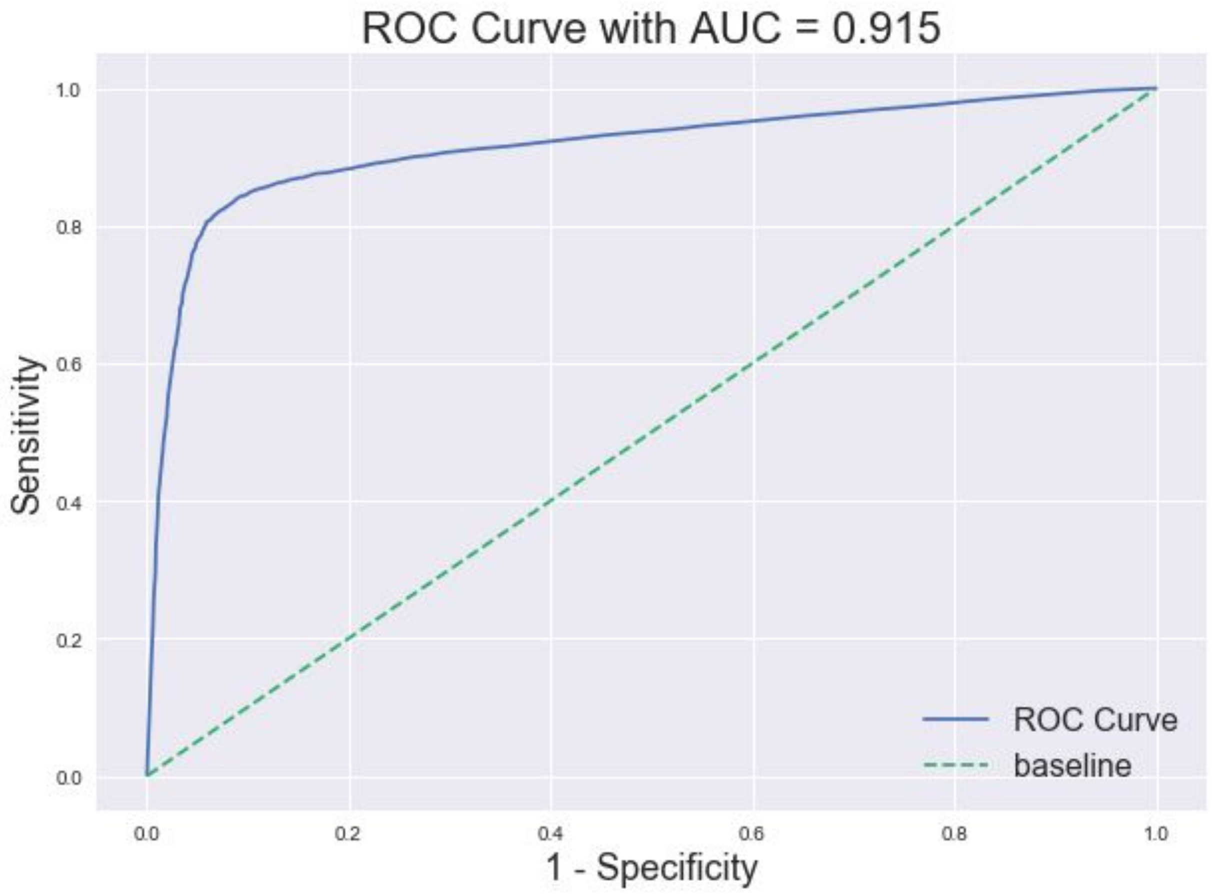


All together

Classifier	F1 score	Recall
Logistic Regression	**87.91%**	**87.35%**
Logistic Regression with balanced class	87.64%	87.33%
LSTM - word embeddings on dataset	85.18%	-
LSTM & CNN - word embeddings on dataset	84.95%	-
LSTM 1 - pre-trained word embeddings	81.75%	-
LSTM 2 - pre-trained word embeddings	84.36%	-
BERT - pre-trained	87%	87%

Logistic Regression

The best model of 87.91%!


ROC Curve with AUC = 0.915
Sensitivity
1 - Specificity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve
baseline

Misclassifications - False Negatives

- Subjective
- Mispelled derogatory terms
- Mislabelling
- Contextual to the conversation
    - “Exacly my point, and thats why we have the second amendment so if any of those monkeys try and give me consequences for my speech i can blow them away”

Misclassifications - False Positives

- Mislabelling
- Sensitive to strong words
    - “That is a fair concern. However, I am a hillbilly stuck in Denver. Everytime I hear one of these harpies try to to act like we are evil men because we are white and straight, I feel like reminding them just how dangerous we really are. How easy would it be for you and your buddies to leave many leftist hats on the ground? I know it would not be a challenge on my end.”

Misclassifications - False Positives

- Mislabelling
- Sensitive to strong words
- Many derogatory words in one comment does not mean hate speech
    - "There are a lot of women who are f***ing c**** too, but I still love the women in my life and I know there are a majority out there who aren't dumb c****. It doesn't mean all women have to apologize for the c**** out there. F***ing shit man, f*** progressives."

Limitations and further work

- Subjectivity of hate speech
- New urban words coined every few years or decades
- Detecting sarcasm
- Context

Conclusions

- Rise of social media and anonymity, still a need to continue exploring the best ways of detection
- Presented current approaches and state-of-the-art NLP model
- Still no better way than simple modelling at the moment
- Much more research needs to be done on context based modelling

Let’s eradicate hate speech!