Hate detection by NLP and deep learning methods ...

Presented By: Rami Abulfadl
Advisor: Isabel Segura-Bedmar
Masters in Big Data | September 2020

Content

- Definition and Motivation
- Objectives
- Approach: dataset and methods
- Evaluation
- Conclusions and Future work

What is hate?

“any kind of communication that attacks or uses discriminatory language 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. ”

UNITED NATIONS STRATEGY AND PLAN OF ACTION ON HATE SPEECH

why is it important to detect ?

Facebook engineers quit in protest, accusing the company of 'profiting off hate'

“When the looting starts, the shooting starts” Donald J. Trump

why is it important to detect ?

- Remove abusive content for a good cause.
- Business reputation

Objectives

Acquire

the basic knowledge in the field of NLP and deep learning

Analyse

the progression of methods and techniques till reaching the state of art performance in hate detection.

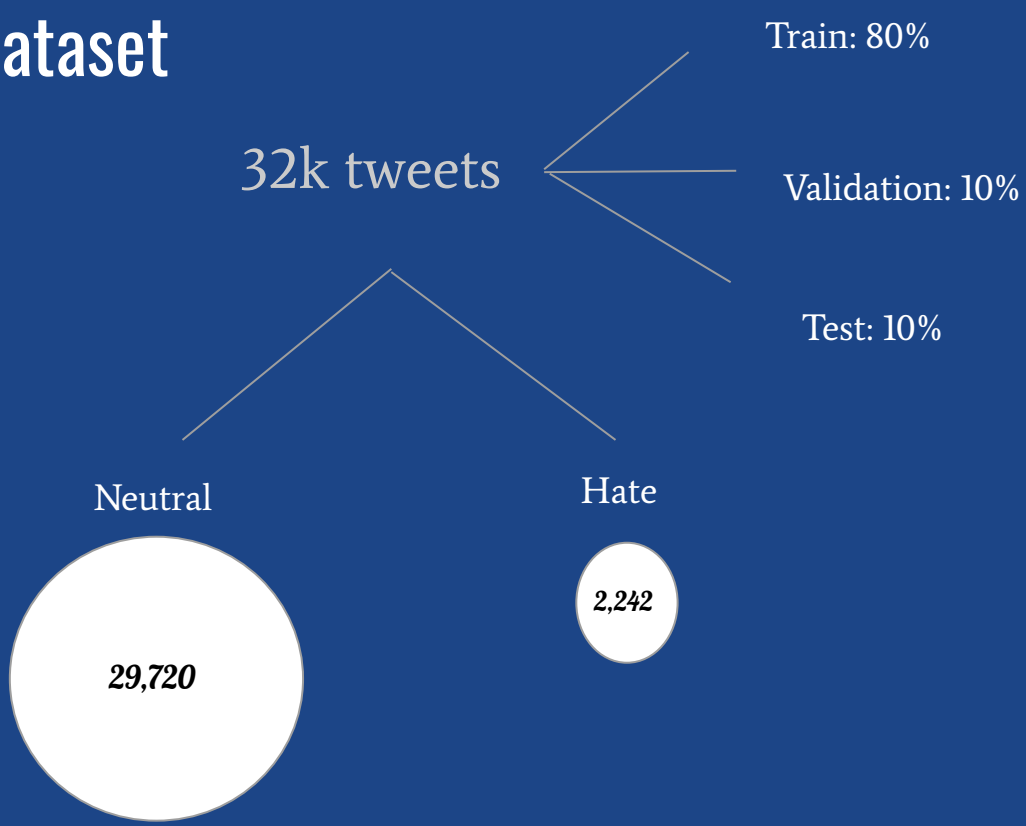
Architect

two most advanced deep learning models, BERT and Bi-LSTM, for hate detection using Twitter dataset.

Assess

The performance of models through error analysis.

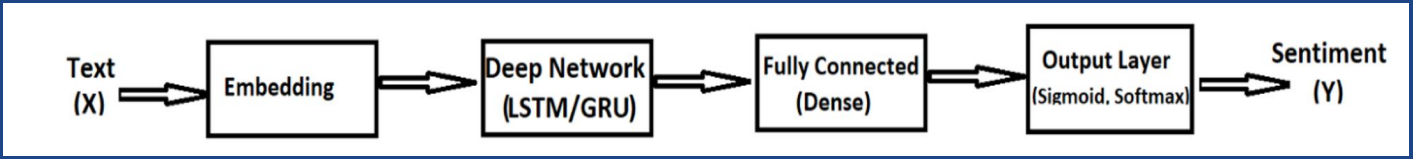
Approach: Twitter dataset



Davidson et al.: Automated Hate Speech Detection and the Problem of Offensive Language

First Approach

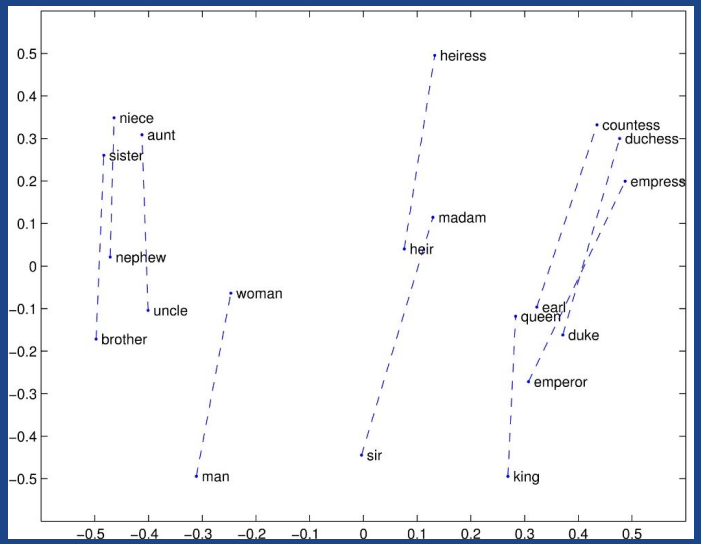
Bi-LSTM

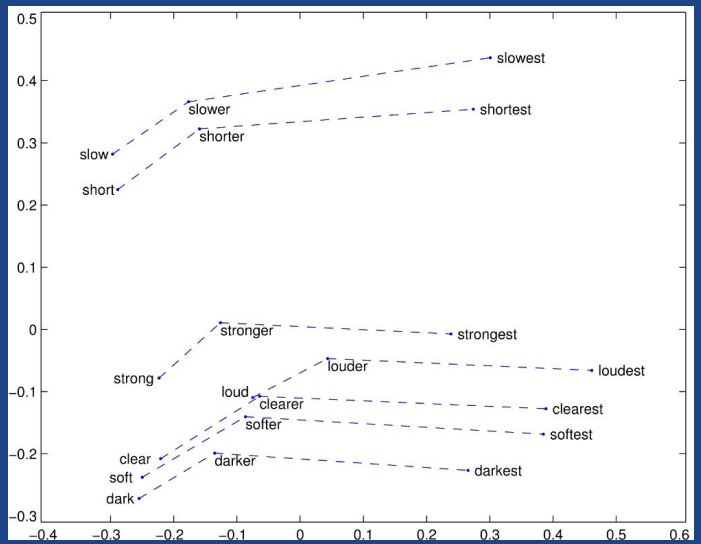


Bi-LSTM

Word Embedding by GloVe for vector space representations

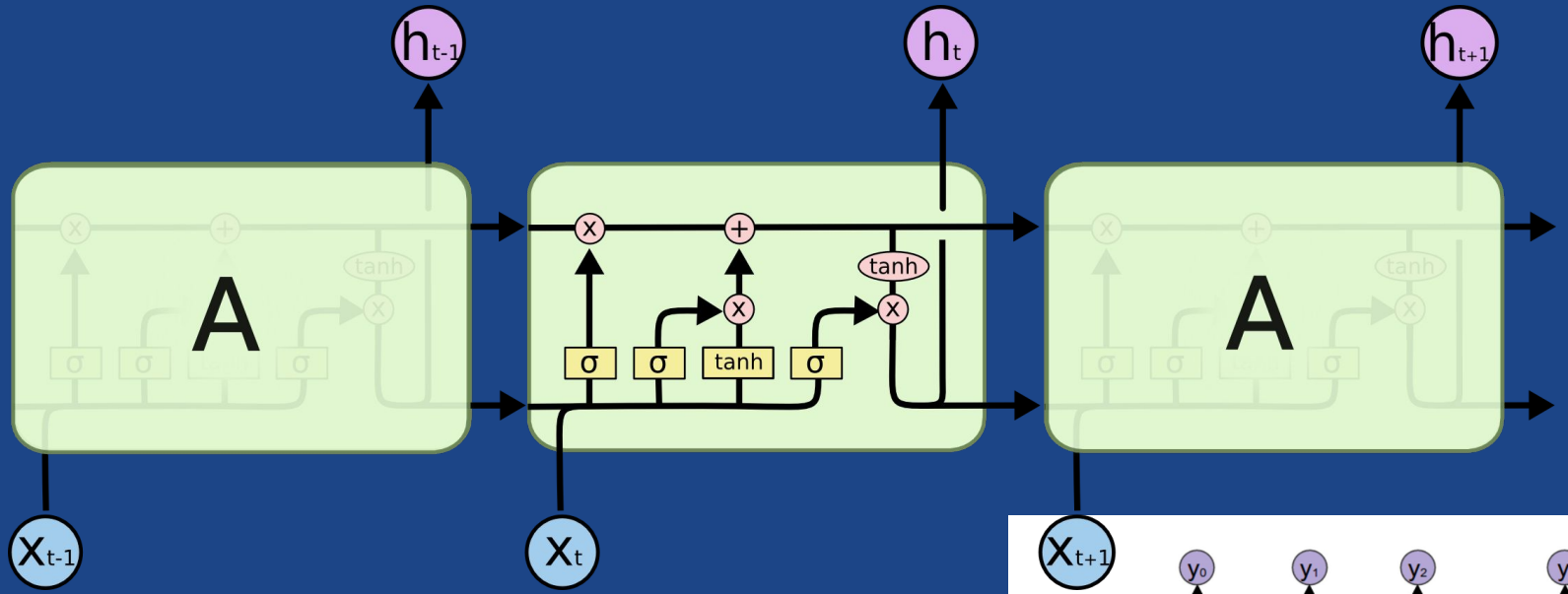
glove.6B.100d 6 B tokens of 400 k vocab forming a 100 dimension form Wikipedia 2014 and Gigaword 5 corpora

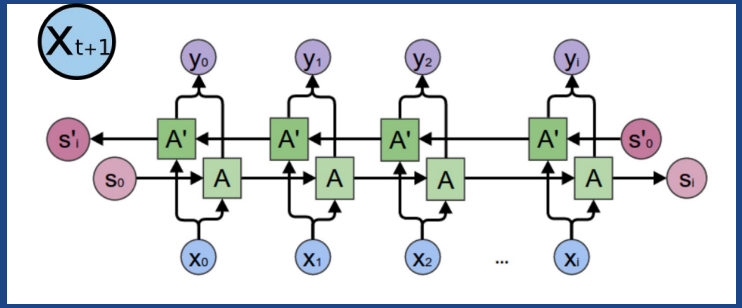




Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Bi-LSTM

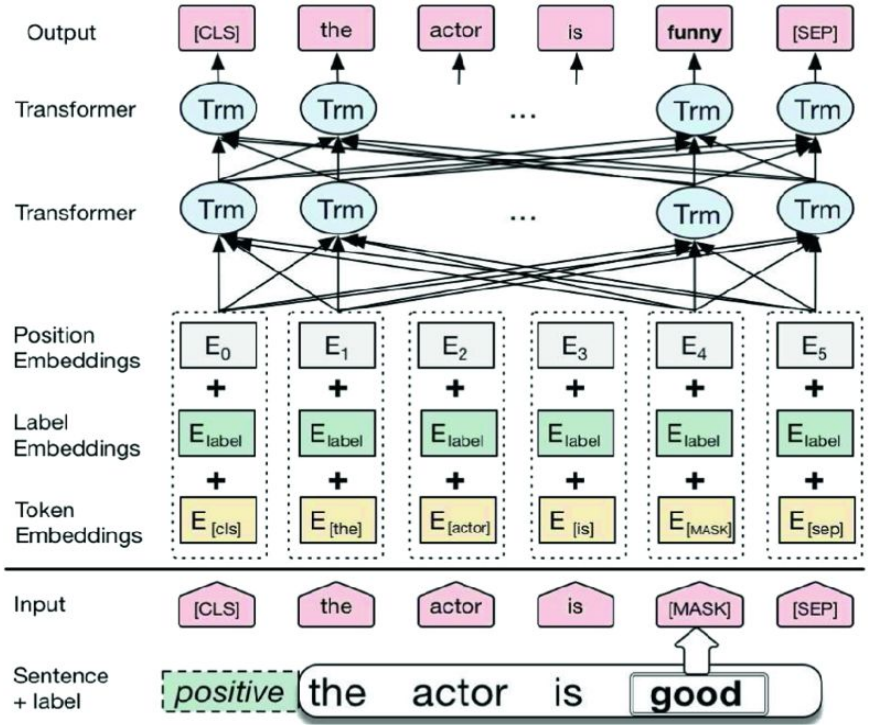

h t-1
h t
h t+1
A
A
tanh
σ
σ
tanh
σ
tanh
X t-1
X t
X t+1


y0
y1
y2
yi
s'i
A'
s'0
s0
A
si
x0
x1
x2
...
xi

Second Approach

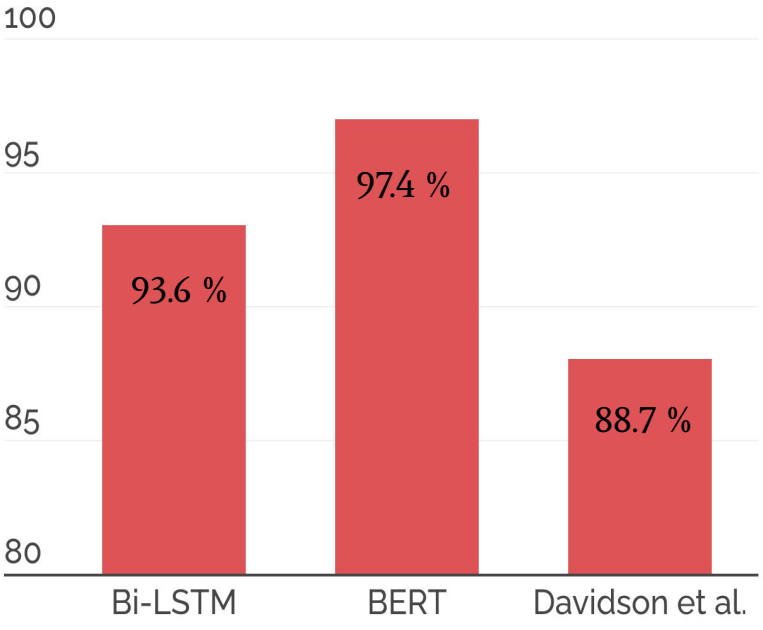
BERT

BERT



Wu X., Lv S., Zang L., Han J., Hu S. (2019) Conditional BERT Contextual Augmentation. In: Rodrigues J. et al. (eds) Computational Science – ICCS 2019. ICCS 2019. Lecture Notes in Computer Science, vol 11539. Springer, Cham. https://doi.org/10.1007/978-3-030-22747-0_7

Evaluation

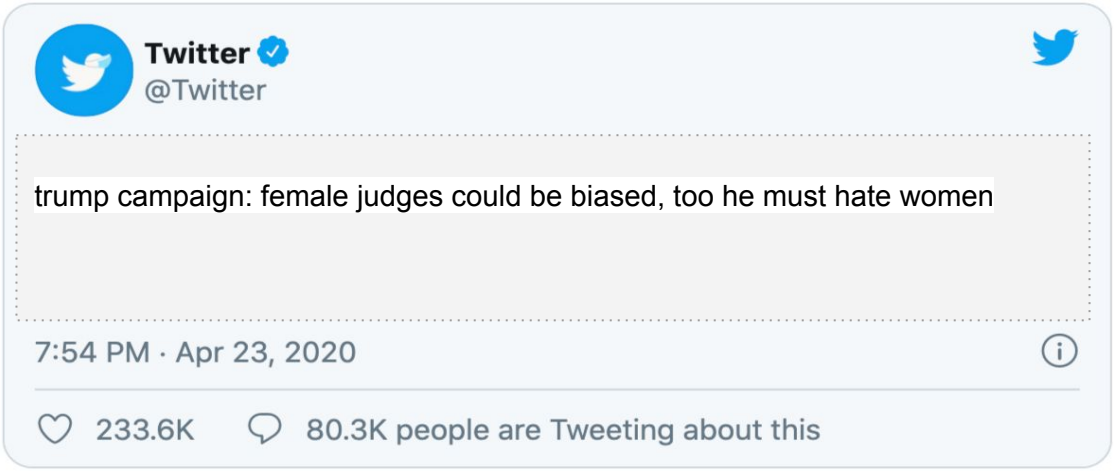


FI-Score

BERT	Bi-LSTM
0.7	0.6

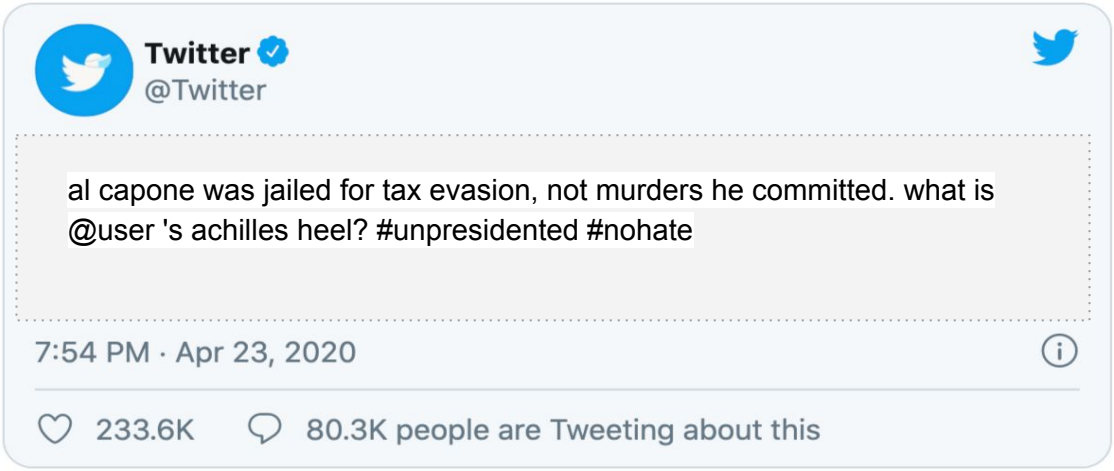
Error Analysis

False positive by LSTM, but True negative by BERT



Error Analysis

False negatives by BERT but True positive by LSTM



Conclusions

- BERT achieve the best performance in classification.
- Both Models get higher accuracy than original davidson et al. results.
- Both Models were able to handle unbalanced classes better.

Future Work

Big dataset for training.

Develop better word Embedding for slang words and abbreviations

龍 odio

Cross lingual unified model for non english languages

Code repository on GitHub

https://github.com/ramiabulfadl/Hate-detection