UNIVERSITY OF TARTU
1632

4N - Negative News Neural Nets Project

Classifying Media Articles with Machine Learning Models

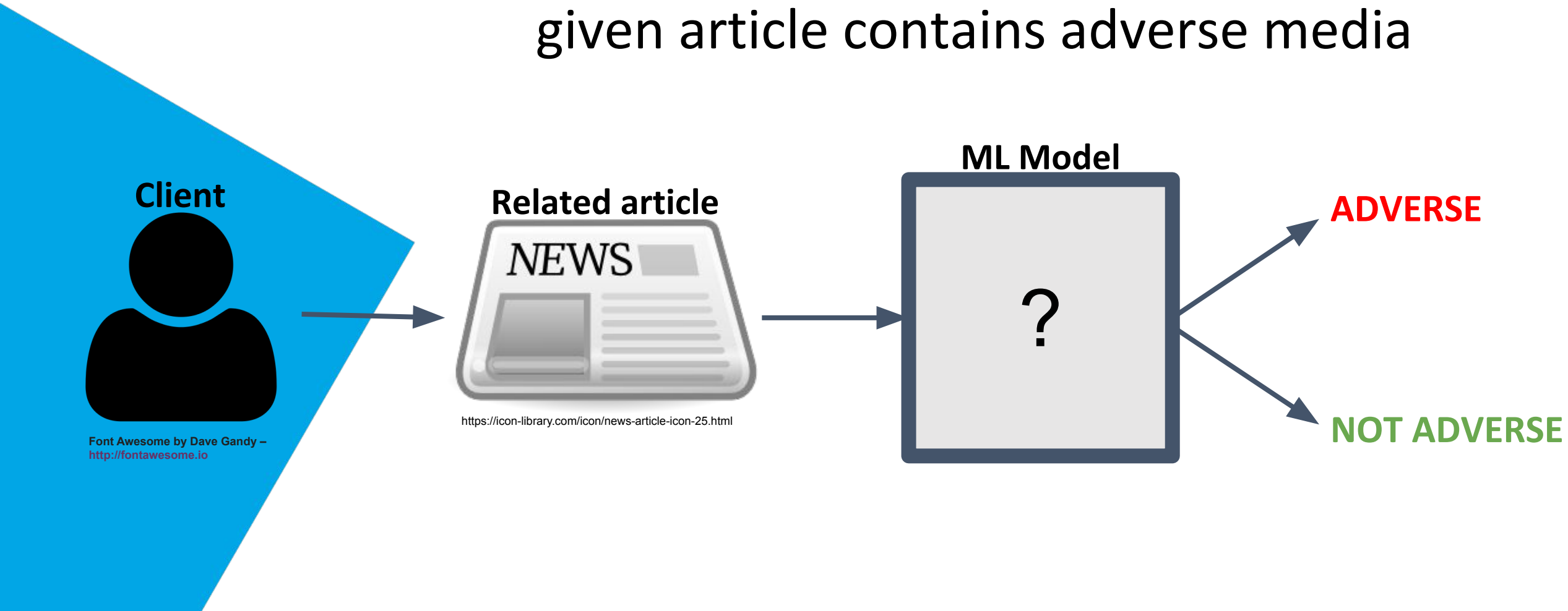
Owner: Kristjan Roosild

Team P18:
Canberk Ozen
Karl Hannes Veskus
Kristjan Roosild
Villem Oskar Ossip

UNIVERSITY OF TARTU
1632

Goal

A model that determines whether a given article contains adverse media



UNIVERSITY OF TARTU
1632

Preprocessing & Cleaning

Before pre-processing:

	article	label
8	Top 10 Crooked CEOs Bernie Madoff, who is sche...	1
10	Top fund manager forced to resign after BBC in...	1

After pre-processing:

	article	label
0	crooked ceos bernie madoff schedule sentence j...	1
1	fund manager force resign bbc investigation pu...	1

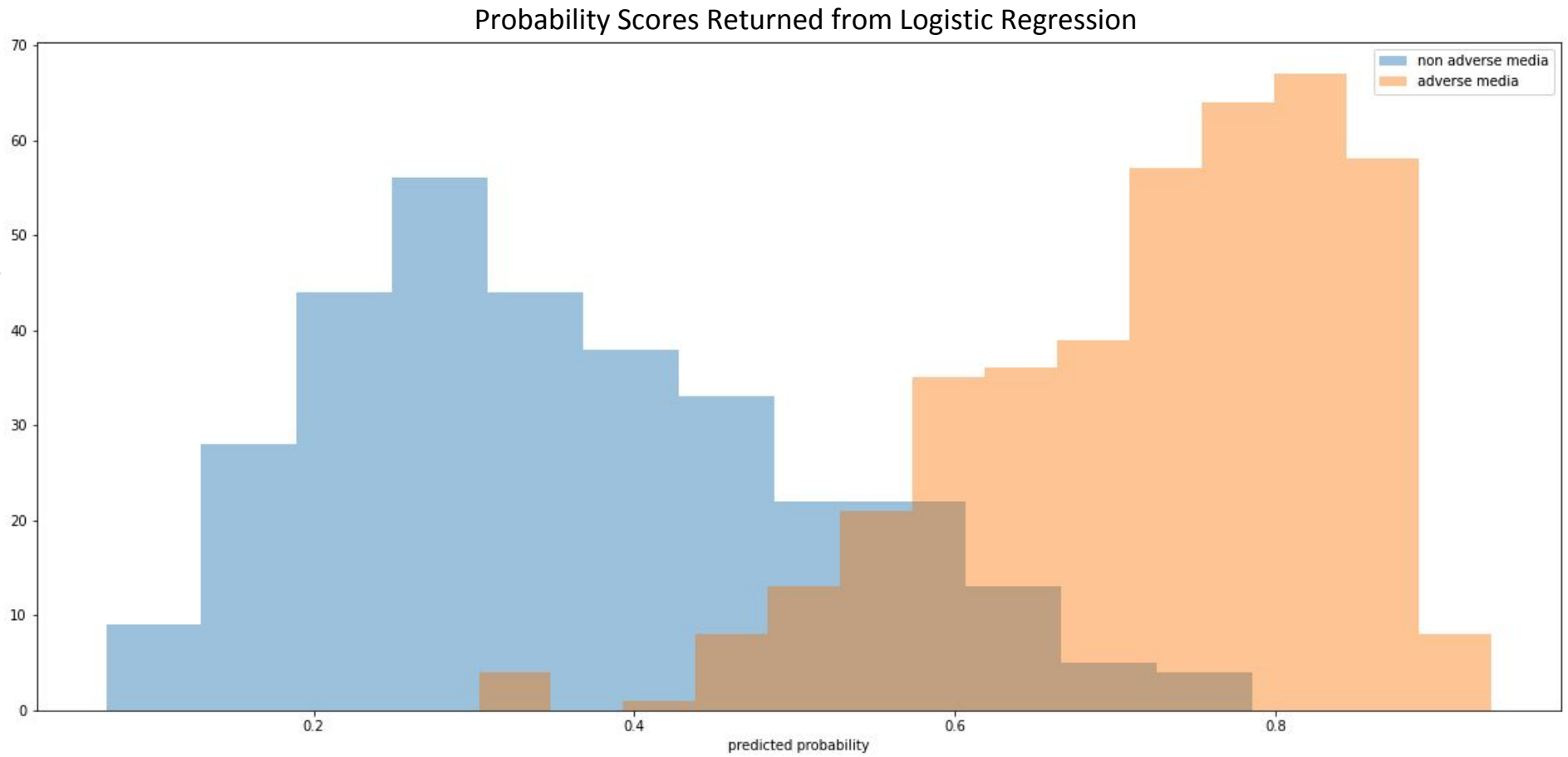
UNIVERSITY OF TARTU

Negative results (a.k.a. the path not to take)

- Additional automatically collected data
- Text augmentation
- LSTM
- Logistic Regression
- GloVe
- XLNet
- etc.

UNIVERSITY OF TARTU
1632

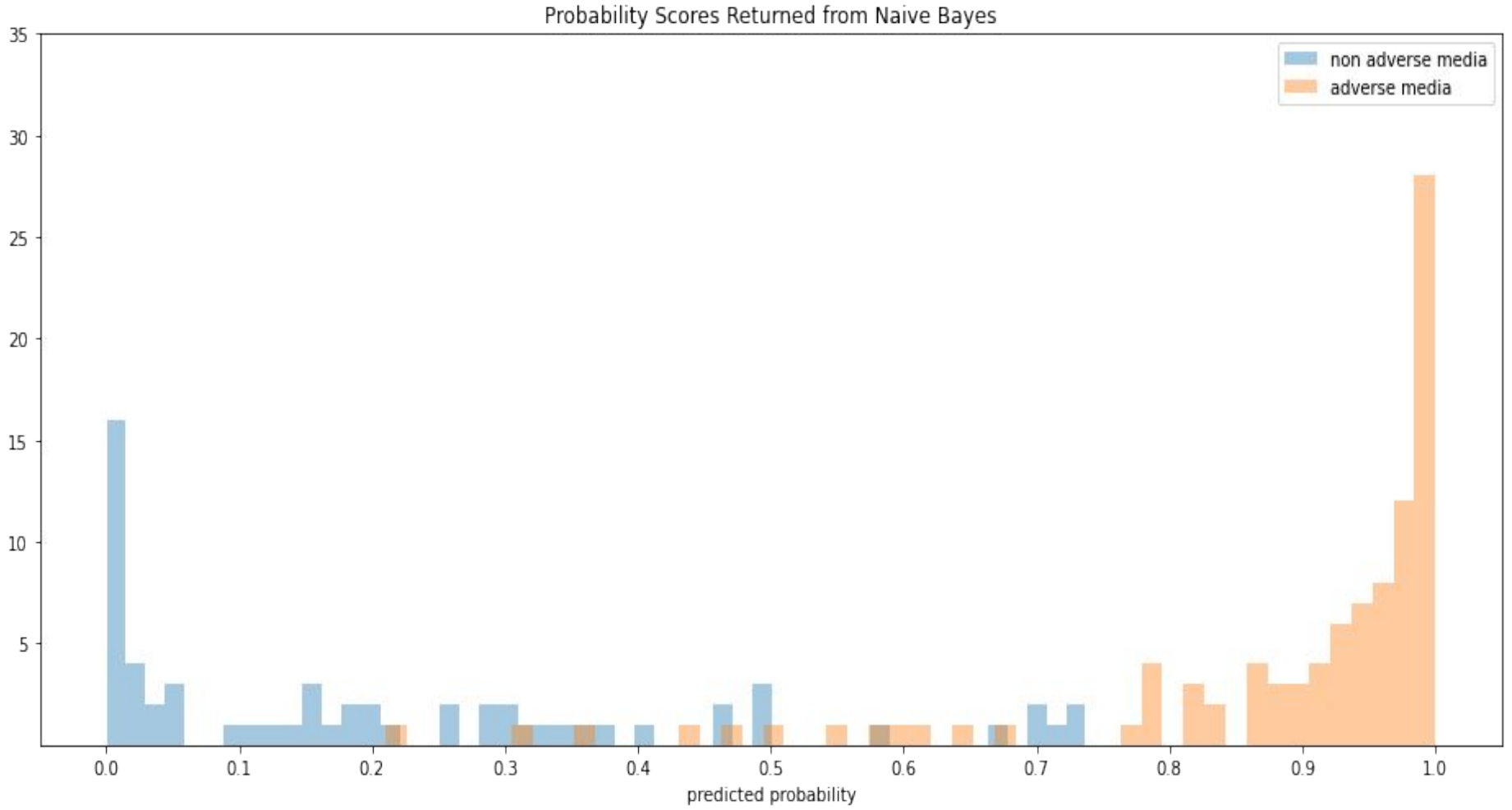
Negative results



UNIVERSITY OF TARTU

Naive Bayes

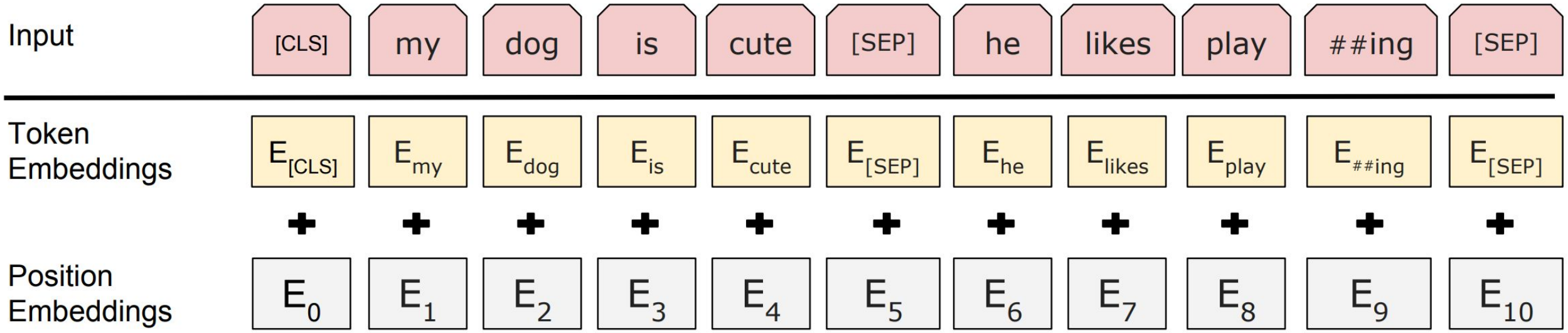
- **Public test F1-score: 93.9%**



UNIVERSITY OF TARTU

BERT

Bidirectional Encoder Representations from Transformers



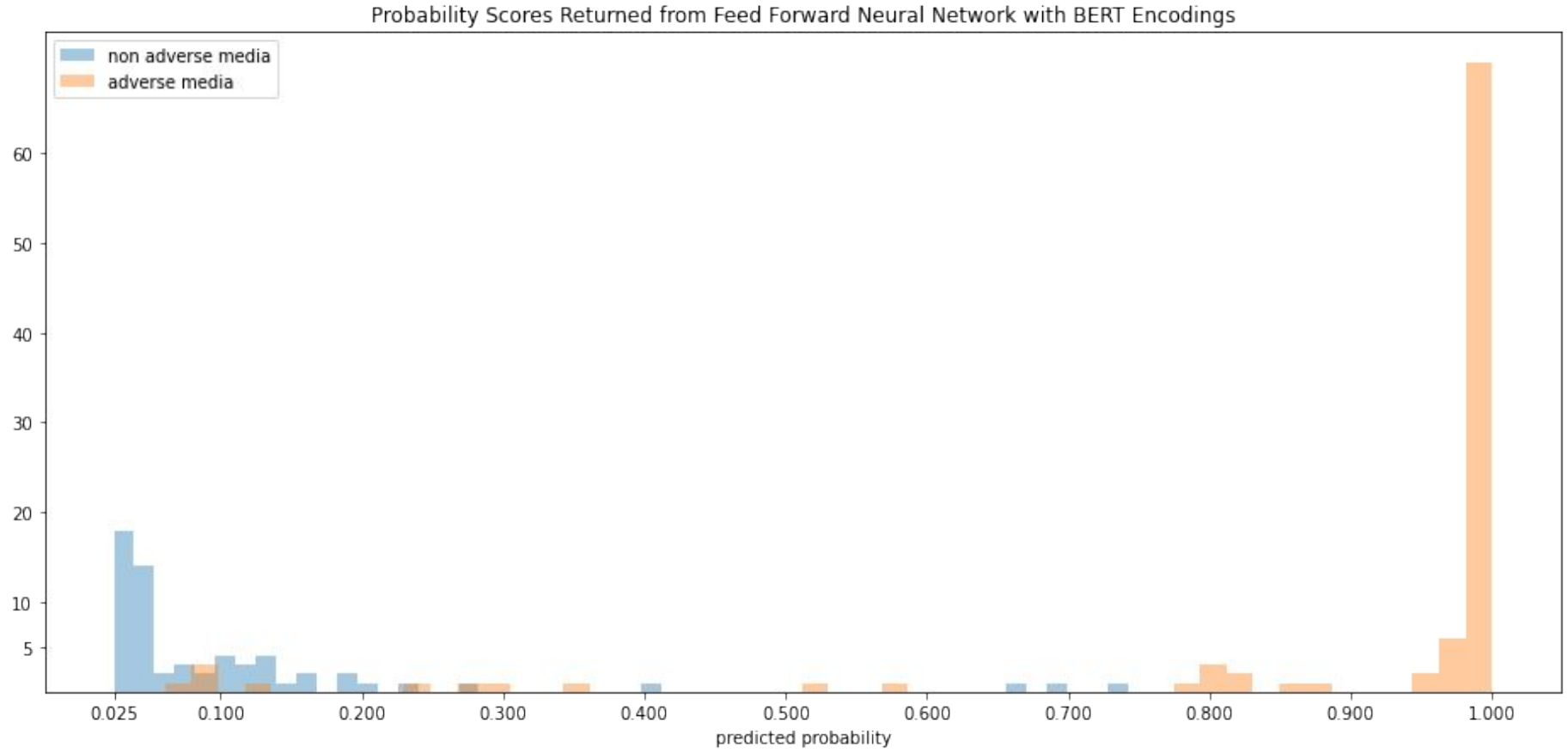
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (https://arxiv.org/pdf/1810.04805.pdf)

- Max length of ~500 words
- Full articles result in OoM errors

UNIVERSITY OF TARTU

BERT Results

- **Public test F1-score: 94.3%**



UNIVERSITY OF TARTU

Learnings

- Quality over quantity
- Keep it simple

UNIVERSITY OF TARTU
1632

Links:

https://github.com/kristjanr/ut-ml-adverse-media