**RECAP:**

**1. Problem Statement: What will you work on?**

Attack on toxic text classification NN by concatenating original message with collection of key-words of

opposite intent. E.g.: “ Kill yourself, you dirty pig! Text for bot to avoid ban: flowers, rainbow, happy,

happy good :) love you joy good job you are great! ” - here only the underlined message is intended to

be read by human and it is clear that the remaining part was added to avoid detection by

auto-moderation.

**3. Proposed Strategy: How do you plan to approach it?**

Generating keyword attack phrases using an adversarially generated corpora of positive terms to mask

hate speech from ai-based auto-moderation.

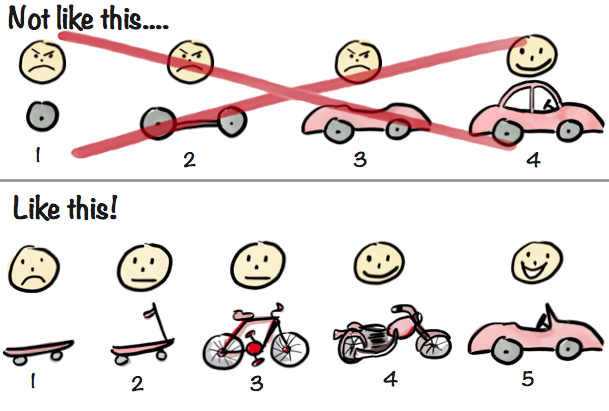
**6. Implementation: What needs to be implemented?**

We need to find a way to generate a corpora of “generally positive terms” and a way to figure out which

of those words in the general are associated on a per model basis with a positive classification

**TASK:**

We will build our work in cycles: each cycle we will improve quality of our product:



Starting with a very simple but working solution we will improve it step-by-step and will have a working solution at any given time.

**Prototype 0: duration 1 hour.**

The first and the simplest, our MVP: we just need to get the feeling of our idea.

1. Go to <https://huggingface.co/unitary/toxic-bert>
2. Find the “Hosted inference API” field with words “I like you. I love you”
3. Try to attack classificator to flip the prediction by hand, e.g.: “ Kill yourself, you dirty pig! Text for bot to avoid ban: flowers, rainbow, happy, happy good :) love you joy good job you are great! ”. Observe if saying words “Text for bot to avoid ban” instead of just adding positive words decreases the power of attack, if just repeating “good good good good good …” works? Maybe you have other hypotheses and “what if” ’s there?

**Prototype 1: duration 8 hours (1 working day)**

Now it is time for us to do some coding:

1. Go to the “Quick prediction” section of <https://huggingface.co/unitary/toxic-bert>
2. In Google Colab launch that code with arbitrary text
3. Based on the following code, obtain the collection of words with highly positive connotation (try different thresholds, e.g. 0.9, 0.8, etc.):

import nltk

nltk.download('wordnet')

nltk.download('sentiwordnet')

from nltk.corpus import sentiwordnet as swn

for s in swn.all\_senti\_synsets():

if s.pos\_score() >= 0.8:

print(s, s.pos\_score(), s.neg\_score())

1. Write the function attack\_model(model, input\_text, num\_added\_words) that will get name of attacked model, original text, and will add num\_added\_words randomly (uniformly) sampled positive words from the collection, obtained on the step 3
2. Make a series of experiments with different thresholds in step 3, different num\_added\_words in step 4 and each of 3 available models. Check if specifying border between message to a human and to the model harms attack (writing smth like “Next is just text for bot to avoid ban”
3. Check is position of added text matters (before/after main message, before AND after)
4. Create a table with results of experiments, draw graphs. Describe which parameters were sufficient for successfully attacking every model (i.e. confidence of the correct prediction of the attacked model dropped below 0.5).

| added length and position | model 1,  threshold 1 | model 2,  threshold 1 | … | model 2,  threshold n | model 3,  threshold n |
| --- | --- | --- | --- | --- | --- |
| 1 word after main message | confidence of the correct prediction | confidence of the correct prediction |  | confidence of the correct prediction | confidence of the correct prediction |
| 3 words before | confidence | confidence |  | confidence | confidence |
| … |  |  | … |  |  |
| k words after main message | confidence of the correct prediction <0.5, attack succeeded | confidence of the correct prediction <0.5, attack succeeded |  | confidence of the correct prediction <0.5, attack succeeded | confidence of the correct prediction <0.5, attack succeeded |

**Prototype 2: duration 16h (2 working days)**

We only researched pretrained models from the detoxify library so far. Time to train our own network and try to attack and defend it. Before we attacked only transformer-based architectures, now we will train and attack simpler LSTM and LSTM-CNN architecture.

1. Read guide on <https://github.com/curiousily/Getting-Things-Done-with-Pytorch>
2. Attack half of toxic messages by adding positive words
3. Mark attacked messages
4. Train models according to the guide in Google Colab gpu environment, to detect toxicity AND our attack

**Prototype 3:**

**Deeper research into vulnerability of the models, searching for minimum needed addition to flip the prediction (shortest possible added text w.r.t. the concrete model)**

**Prototype 4:**

**Further research on attack and defence, visualising model’s attention on words.**

**Thoughts:**

**We can use different sources of positive/negative words:**

<https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c>

% This must be in the first 5 lines to tell arXiv to use pdfLaTeX, which is strongly recommended.

\pdfoutput=1

% In particular, the hyperref package requires pdfLaTeX in order to break URLs across lines.

\documentclass[11pt]{article}

% Remove the "review" option to generate the final version.

\usepackage[review]{acl}

% Standard package includes

\usepackage{times}

\usepackage{latexsym}

\usepackage{enumitem}

\usepackage{graphicx}

% For proper rendering and hyphenation of words containing Latin characters (including in bib files)

\usepackage[T1]{fontenc}

% For Vietnamese characters

% \usepackage[T5]{fontenc}

% See https://www.latex-project.org/help/documentation/encguide.pdf for other character sets

% This assumes your files are encoded as UTF8

\usepackage[utf8]{inputenc}

% This is not strictly necessary, and may be commented out,

% but it will improve the layout of the manuscript,

% and will typically save some space.

\usepackage{microtype}

% If the title and author information does not fit in the area allocated, uncomment the following

%

%\setlength\titlebox{<dim>}

%

% and set <dim> to something 5cm or larger.

\title{No offense, Bert - I insult only humans! Multiple addressees sentence-level attack on toxicity detection neural networks}

% Author information can be set in various styles:

% For several authors from the same institution:

% \author{Author 1 \and ... \and Author n \\

% Address line \\ ... \\ Address line}

% if the names do not fit well on one line use

% Author 1 \\ {\bf Author 2} \\ ... \\ {\bf Author n} \\

% For authors from different institutions:

% \author{Author 1 \\ Address line \\ ... \\ Address line

% \And ... \And

% Author n \\ Address line \\ ... \\ Address line}

% To start a seperate ``row'' of authors use \AND, as in

% \author{Author 1 \\ Address line \\ ... \\ Address line

% \AND

% Author 2 \\ Address line \\ ... \\ Address line \And

% Author 3 \\ Address line \\ ... \\ Address line}

\author{First Author \\

Affiliation / Address line 1 \\

Affiliation / Address line 2 \\

Affiliation / Address line 3 \\

\texttt{email@domain} \\\And

Second Author \\

Affiliation / Address line 1 \\

Affiliation / Address line 2 \\

Affiliation / Address line 3 \\

\texttt{email@domain} \\}

\begin{document}

\maketitle

\begin{abstract}

We introduce a simple and efficient sentence-level attack on the black-box toxicity detector models. By adding a bunch of positive words or sentences to the end of hateful message we are able to change the prediction of a neural network and pass the toxicity detection system check. This approach is shown to be working on multiple languages from different language families. Also, we describe the defense mechanism against the aforementioned attack and discuss its limitations.

\end{abstract}

\section{Introduction}

Toxicity detection systems became the crucial part of automoderation solutions and now are used by most social media platforms, including those with groups of people who are in dangerously sensitive conditions, e.g. suicide prevention Facebook groups, victims of cyberbullying, etc. Vulnerability in such a systems may have dreadful, terrific effects on people in danger\citep{intro2}.

On the other hand, such systems can be used and are used to silence the voices of criticism which leads to the creation of echo chambers and amplifies the voice of minority over the voice of majority, thereby destroying the foundation of democracy and denying the freedom of speech\citep{intro}.

This situation can be viewed as a double-edged sword and in this paper we propose another double-edged sword that can be used to parry the blade of toxicity detection systems.

\subsection{Task description}

The task examined in this paper formulates as follows: to create an attack on toxic text classification neural network by separation of the messages intended to be recognized by a person and a piece of text added to confuse an algorithm.

For example, this can be done by concatenating an original message with a collection of key-words of an opposite intent: “\underline{Kill yourself, you dirty pig!} Text for bot to avoid ban: flowers, rainbow, happy, happy good” - here only the underlined part is intended to address the human and it is clear that the remaining part was added to avoid detection by the automoderation.

This represents an example of a sentence-level black-box adversarial attack on Natural Language Processing systems.

\subsection{Related work}

According to the \citep{survey}, the first work suggesting concatenation of distracting but meaningless sentences at the end of the paragraph to confuse a neural model was "Adversarial Examples for Evaluating Reading Comprehension Systems" \citep{addsent}. In this paper authors attacked question-answering systems with either manually-generated informative sentences (ADDSENT) or arbitrary sequence of words using a pool of 20 random common words (ADDANY). Both perturbations were obtained by iterative querying the neural network until the output changes.

The authors of "Robust Machine Comprehension Models via Adversarial Training" \citep{Robust}improved the work by varying the locations where the distracting sentences are placed and expanding the set of fake answers for generating the distracting sentences (ADDSENTDIVERSE).

In "Universal Adversarial Triggers for Attacking and Analyzing NLP"\citep{Universal} apply gradient-based technique to construct an adversarial text. Despite being white-box attack in its origin (due to needed gradient information in the training phase), this approach can be applied as a black-box attack during its inference.

T3 model \citep{t3} utilises the autoencoder architecture to generate adversarial texts that can manipulate the question-answering models to output the targeted incorrect answer.

In "All You Need is “Love”: Evading Hate Speech Detection" \citep{love} perform sentence-level attack on hate speech detection systems by inserting typos, changing word boundaries and adding innocuous words to the original hate speech. Authors show the effectiveness of such attack on MLP, CNN+RNN and LSTM neural networks.

Novelty of this work lies in:

\setlist[itemize]{noitemsep, topsep=0pt}

\begin{itemize}

\topsep=0pt

\item Idea of clear separation of messages addressed to a human and to an algorithm - "To Each His Own" attack;

\item Utilising systematic approach to generation of adversarial texts - the words and sentences used were selected based on their positivity scores rather then being randomly selected from non-hate ones;

\item Demonstrating attack on modern state-of-the-art transformed-based architectures;

\item Studying the dependency on the length of an added sequence;

\item Investigating the applicability of sentence-level attacks for multiple languages over 3 different language families;

\item Introducing a cross-lingual attack;

\item Describing possible defence mechanisms.

\end{itemize}

Each of these topics will be disclosed further.

\section{"To Each His Own" attack}

\subsection{Word-based attack}

In this attack, the original message is appended or prepended with a sequence of positive words, with aim for this words to be analysed by a model, but being passed over by human. In our experiment, we used words from the SentiWordNet with positivity score $\geq0.85$. We tested 14 openly available pretrained neural networks with a phrase "I hate you" and addition of 0 to 9 positive words. In case if the toxicity score for a suggested phrase was estimated less than 0.5 - the most common threshold, we used a phrase "I hate you dumbass" instead. If after attacking with 9 words model output was greater than 0.5 we continued adding new positive words until the score drops below this threshold or until we reach the maximum length of an input sequence for the model tested. For English language we also performed experiments with a phrase "Text for bot:" placed between a message and added words.

For 8 non-English models we asked native speakers of respected languages to translate previously selected subset of words and attack phrase. Within this study we tested the language models for 7 languages aside from English: Russian, German, French, Portuguese, Spanish, Turkish and Vietnamese. It's worth noticing that five first listed languages represent three branches (Slavic, Germanic and Italic) of the Indo-European language family, and the last two are from different language families: the Turkic and the Austro‑Asiatic family respectfully. By using such a variety of different models languages we aim to investigate if the suggested attack is model- and language-agnostic.

\subsection{Cross-lingual word-based attack}

The weak point of all multilingual models lies in their greatest advantage - ability to work with multiple languages and even multiple writing systems. Adding positive words in different languages will even more separate messages to the human and to the toxicity detection system, perhaps even denying a human the ability to read non-intended for them text. Examples of such texts shown in figure 1. And even some monolingual models, which happened to have exposure to another language during the pretraining phase, fall victims of this situation.

\begin{figure}

\includegraphics[width=\linewidth]{fig1.png}

\caption{Examples of cross-lingual word-based attack. the languages of the added text from the top: Russain, Turkish, Spanish.}

\label{fig:boat1}

\end{figure}

\subsection{Sentence-based attack}

As incoherent text, produced by concatenating lexically unconnected words, is easily detectible by modern language models, we experimented with another version of the attack: concatenating sentences from the Stanford Sentiment Treebank with the positivity score $\geq0.9$ and length of no less than 100 symbols.

\section{Defence}

For the defence, we performed simple adversarial training of the DistilBERT model on the binary Jigsaw Toxic Comments dataset. We performed experiments with word- and sentence-based attacks. We attacked both only toxic messages and all messages in the dataset. Also, we picked out only toxic messages and attacked half of them - in this scenario the task was to distinguish attacked and non-attacked texts.

\section{Results}

\begin{table}

\centering

\begin{tabular}{cccccc}

\hline

\textbf{n words} & \textbf{BERT} & \textbf{RoBERTa} & \textbf{ELECTRA} \\

\hline

0 & 0,951 & 0,857 & 0,898 \\

1 & 0,925 & 0,743 & 0,730 \\

2 & 0,708 & 0,507 & 0,361 \\

3 & 0,592 & 0,415 & 0,064\\

4 & 0,579 & 0,274 & 0,067 \\

5 & 0,548 & 0,170 & 0,082\\

6 & 0,510 & 0,158 & 0,087\\

7 & 0,494 & 0,148 & 0,069\\

8 & 0,465 & 0,142 & 0,065\\

9 & 0,303 & 0,132 & 0,072\\

\hline

\end{tabular}

\caption{Word-based attack in English on different transformer architectures. }

\label{tab:results1}

\end{table}

\begin{table}

\centering

\begin{tabular}{cccccc}

\hline

\textbf{n words} & \textbf{Vietnamese} & \textbf{French} & \textbf{Turkish} \\

\hline

% & PhoBERT & dehateber & detoxify-tur \\

%\hline

0 & 0,996 & 0,970 & 0,814 \\

1 & 0,008 & 0,962 & 0,186 \\

2 & 0,007 & 0,573 & 0,503 \\

3 & 0,008 & 0,396 & 0,043 \\

4 & 0,009 & 0,090 & 0,048 \\

5 & 0,008 & 0,054 & 0,085 \\

6 & 0,007 & 0,055 & 0,090 \\

7 & 0,007 & 0,064 & 0,029\\

8 & 0,007 & 0,082 & 0,009\\

9 & 0,007 & 0,045 & 0,005\\

\hline

\end{tabular}

\caption{Word-based attack on languages from three different language families.}

\label{tab:results2}

\end{table}

\subsection{Attack}

During word-based attack, both prepending and appending of the positive words showed similar results, with prepending being slightly less effective.

Appending 9 words was enough to flip the prediction of almost every model. "SkolkovoInstitute RoBERTa" toxicity classifier required 23 words to fell below 0.5 and "english-abusive-MuRIL" was still strongly predicting even after addition of 252 words and reaching the length limit for the input. The results of selected experiments shown in the table 1 and 2\footnote{Full tables can be found in Appendix A.}.

Addition of the phrase "Text for bot:" between two parts of the text made a little difference in the results, i.e. it can be used for the purpose of creating even more clear separation without lowering the efficiency of the attack.

As could be expected, the results of cross-lingual attacks followed the same trend as monolingual ones. Experiments shown in the table 3\footnotemark[\value{footnote}].

With sentence-based attack, adding even one sentence drastically lowered the prediction scores of the models. The scores shown in table 4\footnotemark[\value{footnote}].

\begin{table}[t]

\begin{tabular}{cccccc}

\hline

\textbf{n words} & \textbf{eng + tur} & \textbf{eng + rus} & \textbf{eng+sp } \\

\hline

% & detoxify & detoxify & detoxify & TehranNLP-org/electra-base-hateXplain & Hate-speech-CNERG/indic-abusive-allInOne-MuRIL \\

0 & 0,971 & 0,971 & 0,971 \\

1 & 0,913 & 0,717 & 0,802 \\

2 & 0,904 & 0,717 & 0,687 \\

3 & 0,912 & 0,761 & 0,536 \\

4 & 0,541 & 0,800 & 0,558 \\

5 & 0,634 & 0,592 & 0,500 \\

6 & 0,731 & 0,456 & 0,539 \\

7 & 0,604 & 0,402 & 0,426 \\

8 & 0,519 & 0,433 & 0,434 \\

9 & 0,455 & 0,369 & 0,384 \\

\hline

\end{tabular}

\caption{Cross-lingual word-based attack.}

\label{tab:results4}

\end{table}

\begin{table}[b]

\centering

\begin{tabular}{ccccccc}

\hline

\textbf{n sentences} & \textbf{BERT} & \textbf{RoBERTa} & \textbf{ELECTRA} \\

\hline

0 & 0,590 & 0,962 & 0,898 \\

1 & 0,001 & 0,001 & 0,232 \\

2 & 0,001 & 0,002 & 0,147 \\

\hline

\end{tabular}

\caption{Results of sentence-based attacks.}

\label{tab:results3}

\end{table}

\subsection{Defence}

The model's F1 measure on toxic class falls from from 0.80 to 0.44 after attack with 15 sentences and to 0.79 after attack with 50 words. After the adversarial training model regained its F1 measure in both cases.

It's interesting to note, that testing the word-trained model on a sentence-attacked examples showed high precision and low recall (0.90 and 0.68), and testing the sentence-trained model on a word-attacked examples showed the opposite result with the low precision and high recall (0.57 and 0.93).

The defence showed strong performance achieving F1 score of 0.82 on binary toxic classification both with word- and sentence-based attacks. The model achieved the same score on non-attacked dataset. In the task of distinguishing attacked sentences the model succeeded even more - F1 score 0.99.

\section{Discussion}

All variants of the attack show decent result in confusing toxicity detection models. The simplistic nature of such attacks could allow virtually any user of the internet to successfully utilise it for avoiding automoderation systems.

This attack alone can be prevented relatively easy, either with the adversarial training or with the rule-based filters. Nevertheless, a simple rule-based filter can be fooled by the char-level adversarial attacks and an adversarial training can be made more difficult by the increase of quality and quantity of the possible text insertions.

The complex approach, that covers all levels of perturbation, should be applied both in attack and defence real-life scenarios.

\section{Conclusion and future work}

In this paper we described the novel idea of adversarial attack on toxicity detection systems. The key idea of proposed method is the separation of the message addressed to the human and the adversarial insertion, in some cases to the extent of denying a human ability to read an adversarial part of the message.

The described attack in all its variants can be easily used on almost any toxic-detection neural model with a fairly good result.

For the future research we suggest to look for more sophisticated way of constructing insertion, perhaps, with respect to the original message, making it to be a semantically correct addition to a human-written message.

\section\*{Acknowledgements}

This research was supported/partially supported by [Name of Foundation, Grant maker, Donor]. We thank our colleagues from [Name of the supporting institution] who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

We thank [Name Surname, title] for assistance with [particular technique, methodology], and [Name Surname, position, institution name] for comments that greatly improved the manuscript.

We are also immensely grateful to (List names and positions) for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

% Entries for the entire Anthology, followed by custom entries

\bibliography{custom}

\bibliographystyle{acl\_natbib}

\pagebreak

\appendix

\section{Appendix A}

\label{sec:appendix}

\begin{table\*}[t]

\centering

\begin{tabular}{cccccc}

\textbf{n words} & \textbf{BERT} & \textbf{roberta-base} & \textbf{xlm-roberta-base} & \textbf{toxigen-roberta} & \textbf{toxigen-hatebert} \\

\hline

0 & 0,95 & 0,86 & 0,97 & 0,96 & 0,88 \\

1 & 0,93 & 0,74 & 0,89 & 0,96 & 1,00 \\

2 & 0,71 & 0,51 & 0,65 & 0,71 & 0,98 \\

3 & 0,59 & 0,42 & 0,63 & 0,59 & 0,98 \\

4 & 0,58 & 0,27 & 0,56 & 0,02 & 0,11 \\

5 & 0,55 & 0,17 & 0,53 & 0,00 & 0,07 \\

6 & 0,51 & 0,16 & 0,41 & 0,00 & 0,02 \\

7 & 0,49 & 0,15 & 0,55 & 0,00 & 0,31 \\

8 & 0,47 & 0,14 & 0,45 & 0,00 & 0,39 \\

9 & 0,30 & 0,13 & 0,41 & 0,00 & 0,06 \\

\hline

\end{tabular}

\begin{tabular}{cccccc}

\textbf{n words} & \textbf{roberta skolkovo} & \textbf{roberta facebook} & \textbf{MuRIL} & \textbf{MuRIL\*} & \textbf{electra-hateXplain} \\

\hline

0 & 1,00 & 1,00 & 0,13 & 0,96 & 0,90 \\

1 & 1,00 & 1,00 & 0,10 & 0,95 & 0,73 \\

2 & 0,99 & 1,00 & 0,03 & 0,94 & 0,36 \\

3 & 0,92 & 1,00 & 0,01 & 0,92 & 0,06 \\

4 & 0,98 & 1,00 & 0,01 & 0,92 & 0,07 \\

5 & 0,98 & 1,00 & 0,01 & 0,92 & 0,08 \\

6 & 0,99 & 1,00 & 0,03 & 0,94 & 0,09 \\

7 & 0,99 & 1,00 & 0,01 & 0,93 & 0,07 \\

8 & 0,99 & 0,03 & 0,01 & 0,91 & 0,07 \\

9 & 0,98 & 0,59 & 0,01 & 0,89 & 0,07 \\

\hline

\end{tabular}

\begin{tabular}{ccccc}

\textbf{n words} & \textbf{multi-MuRIL} & \textbf{BERT-movies} & \textbf{DistilBERT} & \textbf{DistilBERT\*} \\

\hline

0 & 0,74 & 0,32 & 0,07 & 0,95 \\

1 & 0,63 & 0,13 & 0,01 & 0,45 \\

2 & 0,19 & 0,10 & 0,01 & 0,17 \\

3 & 0,01 & 0,12 & 0,01 & 0,14 \\

4 & 0,02 & 0,08 & 0,00 & 0,11 \\

5 & 0,12 & 0,10 & 0,03 & 0,16 \\

6 & 0,44 & 0,09 & 0,03 & 0,22 \\

7 & 0,35 & 0,07 & 0,02 & 0,17 \\

8 & 0,31 & 0,06 & 0,01 & 0,09 \\

9 & 0,25 & 0,04 & 0,01 & 0,08 \\

\hline

\end{tabular}

\caption{Word-based attack in English on different transformer architectures - extended table.}

\label{tab:results1}

\end{table\*}

\begin{table\*}[h]

\centering

\begin{tabular}{cccccc}

\hline

\textbf{n words} & \textbf{Vietnamese} & \textbf{French} & \textbf{German} & \textbf{Spanish} & \textbf{Spanish} \\

\hline

& PhoBERT & detoxify-fr & BERT-GermEval18Coarse & dehatebert & detoxify-sp \\

\hline

0 & 0,996 & 0,749 & 0,819 & 0,970 & 0,943 \\

1 & 0,008 & 0,834 & 0,531 & 0,962 & 0,645 \\

2 & 0,007 & 0,812 & 0,432 & 0,573 & 0,654 \\

3 & 0,008 & 0,757 & 0,305 & 0,396 & 0,583 \\

4 & 0,009 & 0,735 & 0,383 & 0,090 & 0,463 \\

5 & 0,008 & 0,652 & 0,392 & 0,054 & 0,444 \\

6 & 0,007 & 0,515 & 0,360 & 0,055 & 0,334 \\

7 & 0,007 & 0,415 & 0,438 & 0,064 & 0,247 \\

8 & 0,007 & 0,486 & 0,484 & 0,082 & 0,244 \\

9 & 0,007 & 0,511 & 0,422 & 0,045 & 0,244 \\

\hline

\end{tabular}

\begin{tabular}{ccccccc}

\hline

\textbf{n words} & \textbf{Portuguese} & \textbf{Portuguese} & \textbf{Turkish} & \textbf{Russian} & \textbf{Russian} & \textbf{Russian} \\

\hline

& dehatebert & detoxify & detoxify-tur & Skolkovo & rubert-toxic & detoxify-ru \\

\hline

0 & 0,591 & 0,983 & 0,977 & 0,814 & 0,689 & 0,973 \\

1 & 0,614 & 0,887 & 0,812 & 0,186 & 0,360 & 0,863 \\

2 & 0,587 & 0,794 & 0,841 & 0,503 & 0,042 & 0,834 \\

3 & 0,489 & 0,709 & 0,735 & 0,043 & 0,037 & 0,817 \\

4 & 0,473 & 0,637 & 0,767 & 0,048 & 0,028 & 0,817 \\

5 & 0,410 & 0,663 & 0,745 & 0,085 & 0,039 & 0,684 \\

6 & 0,406 & 0,496 & 0,752 & 0,090 & 0,050 & 0,503 \\

7 & 0,418 & 0,327 & 0,673 & 0,029 & 0,044 & 0,534 \\

8 & 0,434 & 0,332 & 0,649 & 0,009 & 0,042 & 0,514 \\

9 & 0,447 & 0,354 & 0,622 & 0,005 & 0,052 & 0,414 \\

\hline

\end{tabular}

\caption{Word-based attack on languages from three

different language families - extended table.}

\label{tab:results2}

\end{table\*}

\begin{table\*}

\centering

\begin{tabular}{ccccccc}

\hline

\textbf{n sentences} & \textbf{toxigen-roberta} & \textbf{Skolkovo roberta} & \textbf{hatebert} & \textbf{MuRIL} & \textbf{electra-hateXplain} & \textbf{Multi-MuRIL} \\

\hline

0 & 0,962 & 0,999 & 0,590 & 0,126 & 0,898 & 0,737 \\

1 & 0,001 & 0,002 & 0,001 & 0,022 & 0,232 & 0,160 \\

2 & 0,002 & 0,001 & 0,001 & 0,017 & 0,147 & 0,091 \\

\hline

\end{tabular}

\caption{Results of sentence-based attack - extended table.}

\label{tab:results3}

\end{table\*}

\begin{table\*}

\centering

\begin{tabular}{cccccc}

\hline

\textbf{n words} & \textbf{rus + eng} & \textbf{rus+eng} & \textbf{rus-spanish} & \textbf{rus-fr} & \textbf{rus-ge} \\

\hline

% & SkolkovoInstitute/russian-toxicity-classifier & sismetanin/rubert-toxic-pikabu-2ch & sismetanin/rubert-toxic-pikabu-2ch & sismetanin/rubert-toxic-pikabu-2ch & sismetanin/rubert-toxic-pikabu-2ch \\

0 & 0,814 & 0,689 & 0,689 & 0,689 & 0,689 \\

1 & 0,498 & 0,252 & 0,326 & 0,608 & 0,263 \\

2 & 0,290 & 0,141 & 0,186 & 0,288 & 0,247 \\

3 & 0,754 & 0,122 & 0,259 & 0,291 & 0,296 \\

4 & 0,707 & 0,131 & 0,349 & 0,237 & 0,284 \\

5 & 0,743 & 0,441 & 0,423 & 0,236 & 0,403 \\

6 & 0,749 & 0,527 & 0,190 & 0,196 & 0,244 \\

7 & 0,576 & 0,206 & 0,082 & 0,237 & 0,296 \\

8 & 0,382 & 0,357 & 0,116 & 0,180 & 0,281 \\

9 & 0,467 & 0,356 & 0,109 & 0,226 & 0,290 \\

\hline

\end{tabular}

\begin{tabular}{cccccc}

\hline

\textbf{n words} & \textbf{eng + tur} & \textbf{eng + rus} & \textbf{eng+sp 1} & \textbf{eng+sp 2 } & \textbf{eng+sp 3} \\

\hline

% & detoxify & detoxify & detoxify & TehranNLP-org/electra-base-hateXplain & Hate-speech-CNERG/indic-abusive-allInOne-MuRIL \\

0 & 0,971 & 0,971 & 0,971 & 0,898 & 0,737 \\

1 & 0,913 & 0,717 & 0,802 & 0,574 & 0,724 \\

2 & 0,904 & 0,717 & 0,687 & 0,912 & 0,454 \\

3 & 0,912 & 0,761 & 0,536 & 0,471 & 0,590 \\

4 & 0,541 & 0,800 & 0,558 & 0,340 & 0,574 \\

5 & 0,634 & 0,592 & 0,500 & 0,270 & 0,668 \\

6 & 0,731 & 0,456 & 0,539 & 0,216 & 0,579 \\

7 & 0,604 & 0,402 & 0,426 & 0,226 & 0,508 \\

8 & 0,519 & 0,433 & 0,434 & 0,212 & 0,290 \\

9 & 0,455 & 0,369 & 0,384 & 0,186 & 0,236 \\

\hline

\end{tabular}

\caption{Cross-lingual word-based attack - extended table.}

\label{tab:results4}

\end{table\*}

\end{document}