* Have 5 red 5 green papers with you

For those of you who do not know me, my name is Michiel van Nederpelt

I hope you will, after the coming 15 minutes find offensive language interesting enough so you will want me to present the entire thesis when the most interesting experiments are carried out in a few weeks!

The topic of offensive language detection has seen major interest not only by the community of Natural Language Processing, but rather by society as a whole. At the same time the task of offensive language detection is not only a challenge for computational language understanding but also from a societal viewpoint; the task is hard to define. The reason behind this will be clear for most of you present here today

but for the good of clarity; what one might find offensive the other might not deem offensive at all.

Also living in a free-speech society brings its own challenges for classifying something as offensive and thereby allowing for detection and potential action by platforms.

The various forms of abusive and offensive content online pose risks to the users of social media platforms.

One such example is cyberbulling which has been linked to depression and suicide risk among teenagers, a topic very close to my heart as some might understand.

An example of the societal norms for offensive language, or rather *taking offense*, as a difficulty might be most prevalent in the terror enfolding after the Charlie Hebdo drawing of the prophet Mohammed

NEXT SLIDE

Without going into detail about the horrors, it shows that offensive language, drawings or cartoon interpretation is very person-, background-, race-, sexual orientation-, religion -dependent and *taking offense* might result in violence.

NEXT SLIDE

Direct abuse and insults containing profanity are fairly easy to identify, recognizing indirect insults for example, which often include metaphors and sarcasm, are a challenge to human annotators and, as a consequence, to most state-of-the-art systems.

One of the most common strategies to counter this problem is to firstly

train systems capable of identifying offensive language,

present the output to human moderators

which in turn can decide if the language is offensive or not and can then take action based on their decision.

For now let’s take a step back, no annotation guidelines, no rules, just your opinion. In front of you there are green and red pieces of paper. If you think the example is offensive, please show me the red paper, if not show the green.

2 examples

As we have seen, we do not agree/do agree because we have similar views/level of education or exposure to free speech.

MOVE TO TASK SLIDE

Typically, models are evaluated by measuring their performance on held-out test data using metrics such as accuracy, recall and F1 score.

However, this approach makes it difficult to identify specific model weak points. It also risks overestimating model performance due to increasingly well-evidenced systematic gaps and biases in offensive language datasets. To enable more targeted diagnostic insights, we introduce perturbations and a two-fold cross-domain study, making use of functional tests, for a transformer-based model showing near state-of-the-art performance for offensive language detection.

NEXT SLIDE

As offensive language detection models are improving performance, people have adapted accordingly. The use of special characters, numbers and small alterations of  
letters in words are a way of “cheating” the detection model. By monitoring BERT performance under increasing challenging conditions, the ability of the model to perform in the “real world” will be evaluated.

NEXT SLIDE

In the first three phases depicted here the model is evaluated  
under incresing challenging conditions.

Moving from in-domain training and testing

to training on one dataset and testing the system on another, known as cross-domain testing. Thereafter the first two phases are combined with the addition of perturbations in testing.

The fourth phase will be used to evaluate system performance while the system is trained on a larger dataset; the concatenation of the OLID and HASOC datasets

NEXT SLIDE

**Previous research**

Different abusive and offense language identification problems have been explored in  
literature ranging from aggression identification, cyber bullying, hate speech, toxic  
comments, and offensive language. Similarities between these subtasks have led scholars  
to group them together under the umbrella terms of “abusive language”, “harmful  
speech”, and “hate speech”. We make a distinction in the five broad categories in the thesis.

Each of these subtasks seeks to address a specific yet partially overlapping phenomenon but we will try to stay with the umbrella term for now, for the separate tasks please read the thesis ☺

NEXT SLIDE

I understand that in such a short time period the differences between the five tasks might be somewhat vague. In the review by Davidson et al. examples such as below are used to illustrate the difference between offensive language and hate speech:

NEXT SLIDE

NEXT SLIDE

I won’t read these aloud

NEXT SLIDE

Moreover, Waseem et al. adds to the critical reflection on the relationship and commonalities between different phenomena that have been grouped under the umbrella of “abusive language”, by introducing a two-fold conceptualization that considers

1. whether the abuse is directed at a specific target or towards a generalized group.

2. the degree to which it is explicit or implicit.

An insult targeted at an individual is commonly known as cyberbulling and insults targeted at a group are known as hate speech. The hierarchical annotation model proposed in the latter published works is also used in the datasets I use and they aim to capture this commonality.

The high number of resources and benchmark corpora for many different languages developed in a very narrow time span, from 2016 until now, confirms the growing interest of the community around abusive language in social media.

**Data**

The three-level hierarchical model used in the datasets used is.

**1. Offensive language detection**

2. Categorization of offensive language

3. Offensive language target identification

NEXT SLIDE

The last two categories are beyond the scope of this research, focus will be on the first category (A) thus only hate speech and offensive language identification.

**Difference datasets**

Both are publicly available although you need permission for the HASOC dataset which I got yesterday evening YESSSS!!!

The main difference however is that in the HASOC dataset Offensive language and hate speech are detected whilst in OLID just the umbrella term of offensive language is used.

Also see the much larger corpus for OLID and the skewed distribution between train and test instances

**Methods**

As said, I make use of the BERT model which shows state of the art results for offensive language detection. I will not discuss the model architecture now as it will probably have been discussed by other students before me and I assume you have all been in contact with the exciting possibilities of the Neural Network.

However, I want to give some attention to transfer learning.

The idea behind Transfer Learning is to make systems less task specific, or *isolated*, to counter high costs and thereby decreasing expenses for new system applications. The main goal of Transfer Learning is in essence in its name; to transfer learned knowledge.

There have been different designs for transfer learning, but we will only briefly touch upon sequential transfer learning:

In Sequential Transfer Learning, tasks are learned in two stages, or a twofold sequence. The first phase, the pre-training phase, is the foundation of the model in which general representations are learned on a source task or domain. This learned ”foundation” is where the consecutive learning is based upon and is followed by an adoption phase during which the learned knowledge is applied to a target task or domain.

**Behavioral testing (needs work)**

When a high F1-score is observed for a certain model, with publishers even stating the model outperforming humans in a similar task, this might not at all be a good indicator how the systems really perform, or how systems have learned. In real-world scenario’s humans, or most of them, can relatively easily adapt to a changing sentence, word order, spelling mistake, extra information etc. whilst a system may not.

So, while a system may perform very well on a certain domain or a subset of real-world data it is very important to understand why and how predictions are made.

Behavioural testing is concerned with testing different capabilities of a system by validating the input-output behaviour, presuming to have no knowledge of the internal structure and treating the internal architecture as a black box.

**Checklists will not be discussed here, if you want to know more read thesis ☺**

**Results so far**

As the process of training, testing, validation, scoring etc. is now ongoing I can only present some preliminary results. These are not the most interesting so if you would have me, I would very much like to present the finished work in about a month!

For now, let’s see where we are

For the in-domain training and testing we find these scores. As it will take a while to look at let’s skip to the cross-domain and later revisit this.

MAKE IN-DOMAIN CROSS-DOMAIN COMPARISON ADD ABOUT THREE MINUTES OF RESULTS DISCUSSION !!

Take home message: We can increase the detection of offensive language, we can have platforms ban people ventilating offensively, remove tweets, posts and comments, but we should have our common goal in mind: keep free, respectful, speech above else. The difficulty will be to not stray away from this goal and land an era of new speak as George Orwell described in 1984