Detecting and Classifying cyberbullying Memes in Twitter Posts

submitted in partial fulfillment for the degree of master of science

Richard Chou
12970093

master information studies
data science track
faculty of science
university of amsterdam

	First Internal Supervisor	Second Internal Supervisor
Title, Name	Dr Reshmi Gopalakrishna Pillai	Dr. Fernando Pascoal Dos Santos
Affiliation	UvA, FNWI, IvI	UvA, FNWI, IvI
Email	r.gopalakrishnapillai@@uva.nl	f.p.santos@uva.nl

University of Amsterdam
Faculty of Science

Detecting and Classifying Cyberbullying Memes in Twitter Posts

Richard Chou
richard.chou@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Abstract

Cyberbullying and other forms of aggressive conversation on social media have been increasing since the COVID-19 outbreak. This form of content is also depicted in the form of memes, which is a combination of images and text content. Filtering out image-with-text data on social media is more difficult to do since machine learning models can have difficulties detecting the sentiment of the message. This is due to the format in which memes are presented. This research addresses the detection of cyberbullying within image-with-text data that are posted on the Twitter platform. An attempt is done to set up a multi-modal classification model that is capable to analyse the sentiment of the meme and detecting if the content of the Tweet contains cyberbullying content. In comparison to other studies in this field, this research will address the problem specifically focused on image-with-text data in the form of memes that contain these types of content. The purpose of this model would be to serve as an additional filter for detecting harmful content in memes and contribute to the research problem of creating a healthy Twitter environment. The final optimised model manages to retrieve an accuracy score of 91.03 and an F1-score of 66.67.

Keywords: image-with-text data, hate speech, cyberbullying, multi-modal classification, vision and natural language processing, social media

Github: https://github.com/Rchou97/master_thesis_2022

1 Introduction

Internet memes became a popular phenomenon for expressing opinions in a satiric way among internet users. The COVID-19 pandemic led to a significant increase in digitisation, especially during the period of lockdowns that occurred within 2020, which strengthened the process of creation and sharing of internet memes [3]. According to the research of Tahmasbi et al., due to the increase in online activities over time, this development led to the increased spread of conspiracy theories, online racism, sexist jokes, and misinformation about the COVID-19 pandemic or politics [25]. This also transitioned to cyberbullying across different social media channels, especially in the form of xenophobia and discrimination towards minority groups.

Figure 1. Example of a non-offensive meme (left) and an offensive meme (right)

Research by Alsawalqa [1] showcases that students of the University of Jordan admitted that they were cyberbullying minority groups for humour and satire. The students were not aware that it was classified as bullying when they were posting content about these groups on social media. Similar behaviour is shown in the Czech Republic, where teachers are being bullied and insulted on social media in the form of memes and textual posts [2]. The research by Mkhize and Gopal claimed that "*with the increase of the use of social media among children and youth during the lockdown, most have been victims of cyberbullying*" [20]. This could result in mental and physical problems when people are becoming victims of cyberbullying [14]. It is a problem that social media companies still find difficult to solve, For the purpose of creating a safe and healthy environment for every user that uses social media.

Hate speech on social media can be defined as "*a direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. The attack is defined as violent or dehumanising (comparing people to non-human things, e.g., animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech*" [15].

For this research, an attempt is made to identify and classify offensive memes that are posted via Twitter. Memes are

more difficult to identify in comparison to textual representation, due to the nature of a meme, which is a combination of image and text. After detecting the content of memes, the memes are classified into different forms of cyberbullying, such as fat-shaming, sexism, discrimination based on disability, and racism. These classes are defined as followed:

- **Fat-shaming**: ridiculing people who are being "fat" and/or "overweight" [27].
- **Sexism**: discriminating against someone on the internet based on their gender [12].
- **Physical and mental disability discrimination**: insult someone on social media who has a physical and/or mental disability [4].
- **Racism**: discriminating against someone on the internet based on race [19].

In Figure 1, an example is provided between two memes extracted from Twitter. On the left-hand side, a non-offensive meme is given. On the right-hand side, an offensive meme is showcased. The offensive meme is considered offensive since it ridicules French feminist groups.

The four classes are annotated by random users, who will determine whether or not the meme would contain any form of cyberbullying content or not. These classes are encoded in a binary output, where 1 stands for cyberbullying content, while 0 stands for non-cyberbullying content. Since current state-of-the-art multi-modal classification models can only output a prediction in binary formats.

The model can help with providing recommendations to filter out these types of content. The output of the model could contribute to reducing the number of victims of cyberbullying that happens on social media. Twitter is the platform of choice since there is not much research done in this field within the Twitter platform in comparison to other social media platforms, such as Facebook. In addition, Twitter supports research ideas that could facilitate the platform with a healthier and more engaging platform. Hopefully, this research could contribute to this goal. The performance evaluation metrics that are used for the model are the *F1-score* and the *accuracy*. These evaluation metrics are further elaborated in Section 3.2.

1.1 Research Question

To what extent can image-with-text classification methods identify and label offensive memes that are related to cyberbullying which are posted on Twitter?

- **SRQ1**: What performance level can be achieved with a baseline model by using well-performing (justified by literature) multi-modal classification models that classify image-with-text data?
  - *SRQ1a*: Which technique is considered the best baseline model for identifying offensive memes on Twitter?
- **SRQ2**: What features can be used, that could improve the specified evaluation metrics (accuracy and F1-score) of the chosen multi-class image-with-text model on tweets containing offensive memes?

The main research question can only be answered by answering these two sub-research questions. When answering the main research question of this research, then this will result in finding a solution to address problems with detecting cyberbullying content in image-with-text data in the form of memes. In Section 3, the methodology, type of data, and the types of (baseline) multi-modal classification models are further elaborated and justified for answering the research questions.

2 Related Work

2.1 Internet Memes

According to Bauckhage [6], internet memes are derived from the theory by Dawkin, who defined memes in his book *The Selfish Gene* [10] as a "*cultural analogon of genes to explain how rumours, catch-phrases, melodies, or fashion trends replicate through a population*". This has been translated to the definition of an internet meme, which according to Davison is defined as "*a piece of culture, typically a joke, which gains influence through online transmission*". Internet memes are mostly displayed as an image macro, which is an image containing captions [9]. Internet memes can take multiple forms and can address different themes. According to the research by Barnet et al., memes can be humorous, but they could also address global issues or cultural and political themes. They are created, altered, and shared among internet users throughout different social media channels, such as internet forums or blogs [5].

2.2 Multi-modal classification models in Hateful Memes Challenge

At NeurIPS 2020, the artificial intelligence (AI) group of Facebook (nowadays Meta) organised a challenge related to detecting harmful content posted on social media. In this case, a synthetic data set was created by the data and AI team of Facebook, which could be used for building solutions to identify hate speech for this type of content. The data set contains over 10.000 examples of hateful and non-hateful memes. The challenge was to create such a model that managed to deal with this type of format and detect if it could contain harmful content. Based on the Hateful Memes Challenge data set, several related research papers have been published based on this data set. The top five models and their research are highlighted here [16].

In the fifth spot, the UNITER (Universal Image-Text Representation Learning) ensemble model by Sandulescu managed to claim a spot within the top five of the best-performing models for this challenge [23]. According to Chen et al., an UNITER model first encodes the image regions (bounded box

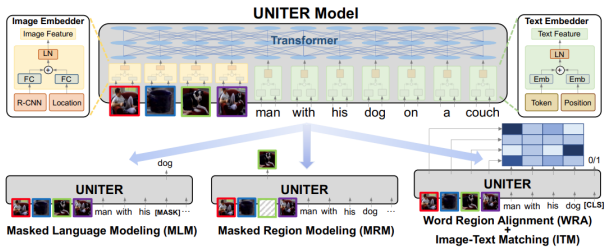

Figure 2. Overview of the UNITER model, which consists of an Image and Text embedders with a multi-layers Transformer, learned through pre-training tasks [8]

Figure 3. Cross-validation ensemble optimisation process. The dev set (blue part) is split into two parts, train and test folds (green and blue parts), which are used for the EA (expert advisor) ensemble optimisation. For the final ensemble, the weights are used to combine the cross-validation model predictions, followed up by evaluating the performance on the unseen test set (red part) [18]

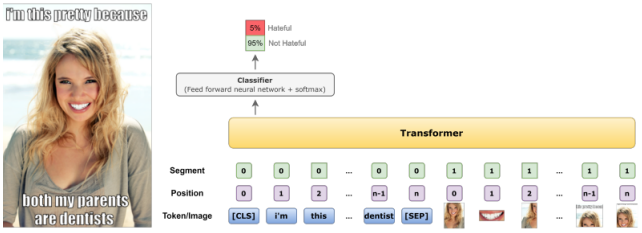

Figure 4. Architecture of VisualBERT on an example meme [26]

features and visual features) and textual words (by positioning and tokenisation) into an embedding space with image and text embedders. This is followed up by a Transformer module to learn the generalisable contextualised embeddings for the identified regions and wordings in the pre-training tasks. The general process of an UNITER model can be seen in Figure 2 [8]. According to Sandulescu, the limitation of his UNITER ensembled model that was used for this challenge is that there is still a gap when comparing it to human performances for detecting and classifying hateful content. Humans managed to reach an accuracy score of 84.7 and an AUROC of 82.65 on the test set. Whereas the model managed to achieve an accuracy score of 74.3 and an AUROC of 79.43 [16, 23].

The fourth spot was claimed by the Kingsterdam team, who have also used an UNITER ensemble model. With an AUROC of 80.53, the model managed to perform slightly better in comparison to the UNITER ensemble model of Sandulescu. The team took a different approach by using a cross-validation ensemble optimisation process for scaling up the performances of the model. This approach can be seen in Figure 3. According to Lippe et al. [18], there is still room for improvement in terms of fine-tuning the image extractor during the training phase of the model. This could help with improving the image understanding, which will result in a better performance for classifying hate memes.

The third spot was claimed by Velioglu and Rose. They approached the challenge by developing and applying a VisualBERT model. BERT stands for Bidirectional Encoder Representations from Transformers. It is a method of pre-training language representations models and using it for natural language representations tasks [11]. VisualBERT is an extension of the BERT model, but it is also taking into account the visual elements within images [26]. VisualBERT works by taking the regions of the images and the language. The regions and the language are combined by the Transformer to allow the self-attention to discover alignment between the vision and language. The architecture can be seen in Figure 4. The model managed to reach an AUROC score of 81.1 with an accuracy of 76.5. According to Velioglu and Rose, this is a considerable result but the results are still far from the accuracy and AUROC scores by human judgement (84.7 and 82.65 respectively).

The second spot was the model developed by Muennighoff [21] in the form of Vilio. Vilio is a visio-linguistic model to identify hateful content within memes. This model aims to provide a user-friendly starting point for visio-linguistic-related problems. Masked pre-training helped the model with identifying hateful memes better in comparison to the competitors of this challenge [15]. Vilio is an ensemble method that averages the performance of 19 different vision-linguistic models. Based on the top five performing models, it will ensemble the models by performing a simple average on the output. This can be seen in Figure 5. The model resulted in an AUROC score of 82.52 on the test set. The limitation of the Vilio model is that it is an ensemble model. Since it is an ensemble model, it can be computationally heavy when it comes to calculating the results.

And finally, the first place went to Zhu [29] with his multimodal transformers with external labels and in-domain pre-training. This model ensembles multiple visual and linguistic models for classifying image-with-text data. What distinguished this model in comparison to the other models, is that Zhu also took into account racial and gender classifications as additional inputs for detecting hate speech within the image-with-text types of data. This approach that was built on top of existing visual-linguistic models (VL-BERT,

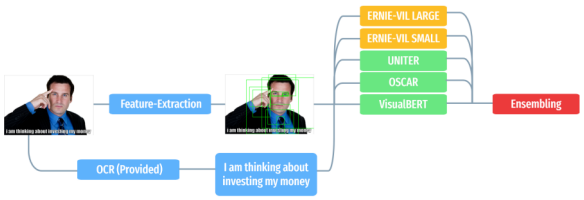

Figure 5. The pipeline of the Vilio model is split into the preparation, modelling, and ensembling stages [21]

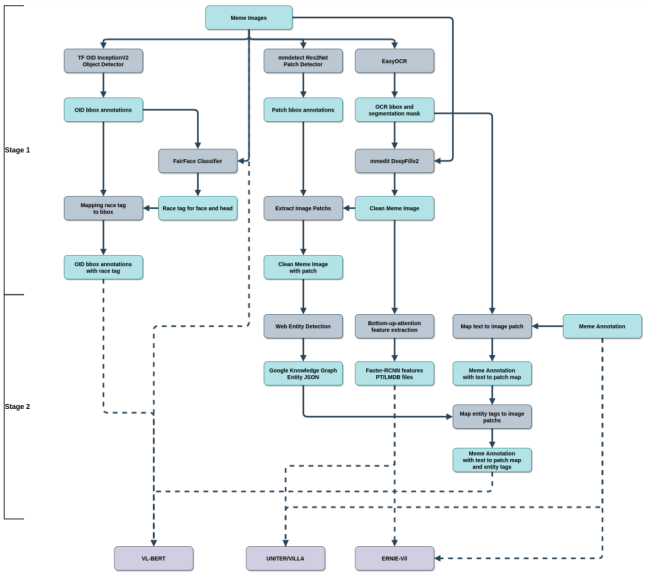

Figure 6. The pipeline of the Enhanced Multi-modal Transformer model is split into two stages: data preparation and modelling [29]

UNITER/VILLA, and ERNIE-Vil) managed to achieve an AUROC of 84.5 on the hateful memes data set. The limitation of this transformer model is that it requires extracting a subgraph that contains the entity that appears in the image-with-text data that could be potentially related to the context, see Figure 6. Entities in a large-scale knowledge, for example, wiki-data, have different types of relations per entity. This resulted in the limitation that when using a subgraph, that both includes sufficient information and relatively large data size, it can be challenging to process the subgraph by this model.

The takeaways from this challenge for detection and classification of image-with-text models were the selected frameworks, that the training size matters, that most of the winning solutions were ensemble models, and finally that the winning models also included additional characteristics that were related to race and gender. These features helped with accurately detecting and classifying hate speech within memes [15]. These takeaways from the challenge will be taken into account when addressing and resolving the research problem. But it is still a difficult use case to resolve since the top five models had difficulties reaching the human benchmarks.

In comparison to the Hateful Memes Challenge, this research is also addressing the challenge of classifying cyberbullying classes on real-life examples on the Twitter platform instead of a binary classification problem of whether a meme is hate speech or not. The data set that was provided for this challenge, is also not representative of real-life cases that are happening on social media. Since memes come in different formats, image qualities and different types of visualisations. This is not the case for the Hateful Memes data set, which mainly consisted of single high-quality images with a caption on them. A real-life example would introduce more complexity to this problem. For example, a distorted or low-quality image could decrease the performance of the models to detect whether a meme contains cyberbullying content or not.

A summarised overview of the performances of the top five models within this challenge can be seen in Table 3 in Appendix A: Related Work and Exploratory Data Analysis.

3 Methodology

The main focus of this research will be to compare and evaluate performances on the selected performance metrics for the multi-modal classification models. The model should be able to detect and classify memes for contents of cyberbullying on Twitter posts. To define this use case in the context of this research, given a tweet that attaches an image M with text C consisting of a sequence of words, a multi-modal classification model will predict the label of the meme with regards to cyberbullying. In this case, the content contains one of the following five classes: fat-shaming, sexism, racism, discrimination towards the mentally/physically disabled individuals, and finally, memes that do not contain any form of this type of content. These are also the labels that define the scope of cyberbullying for this research. This is all aggregated in a binary target variable, which is 0 for non-cyberbullying-related content, and 1 for cyberbullying-related content. The methodology will be further elaborated on in the upcoming subsections.

3.1 Data Extraction & Processing

As mentioned in Section 1, data was gathered by using the Twitter API. A Tweet on Twitter comprises of several contents:

- **Tweet**: the main object within Twitter. The object contains a long list of attributes, such as tweet ID, date of the tweet, and text.
- **User**: contains user account metadata that describes the referenced Twitter user, for example, user ID, username, location, etc.

- **Geo**: tweets that are associated with a location are 'geo-tagged'. Associated data that comes with the geo content are place, location coordinates, country, etc.
- **Entities**: provide metadata and additional contextual information about the content that is posted on Twitter, e.g., hashtags, attached media, etc.

From the mentioned contents, only the *Tweet*, *User*, and *Entities* were extracted from the tweet object. The *Geo* content was not used for this research, since it was not relevant from which location the tweet originated. Only the language of the tweet is relevant since the models are only processing English text. The following features were processed in the data set from the mentioned contents: *Username*, *Text*, *Date*, *Language*, *Hashtags*, and *Image URL*. The tweets were filtered on hashtags in multiple different batches of data and later on merged into a final data set. The hashtags that were used as keywords for filtering were: *#dankmeme* (an ironic expression used to describe online viral media and in-jokes that are intentionally bizarre or have exhausted their comedic value to the point of being trite or cliche [17]), *#dankmemes*, *#dank*, *#meme*, *#memes*, *#memedaily*, *#memesdaily*, *#funnymeme*, and a combination of multiple hashtags over different days. The memes were extracted in between the period of April till May 2022 and consisted of 4229 tweets in total.

Tweets that are considered out-of-scope are tweets that only contain textual data. That is the reason why a filter was created that specifically is focused on whether a tweet had image data or not. After merging all the different CSV files that were extracted from the keywords, duplicate instances based on the Image URL were dropped from the data set, followed-up by dropping images that were extracted from the video/GIF thumbnail. Finally, by using the *Language* feature of the tweet, it was possible to filter out the non-English tweets from the data set. This is to prevent the models from dealing with other languages. Since the models are built on English text extractions. If there are tweets that only contain an URL to an image, but not to an image-with-text data type, then these tweets are denoted as noise within the data set. This could potentially lead to decreased classification accuracy and poor prediction results [13]. So these instances are also out-of-scope for the final data set. Using these filters resulted in a final data set of 722 unique tweets.

After the data was cleaned and processed, manual annotation was performed by random users. Amazon SageMaker Ground Truth was used and random users were approached to get the data labelled according to the labels of this research. This also included the label that the meme does not contain any form of cyberbullying content. The instructions and some good and bad examples of cyberbullying memes were provided to the users for getting the labelling tasks done. The definitions of the four cyberbullying content in Section 1 were used as formal definitions for the labels [4, 12, 19, 27]. If one of the labels were indecisive on the agreement level, for example, if some users annotated the meme as sexist while others find it racist, the final label was decided based on the annotation with the highest Cohen's Kappa level. The confidence label is determined by [7]:

$$k = \frac{P_0 - P_e}{1 - P_e} \tag{1}$$

For more information about the labelling procedure via Amazon SageMaker Ground Truth, then this could be seen in Appendix B: Image Labelling Job.

According to Berry et al., in the context of this research, P_0 are the users that agree with each other, while P_e is the probability that there is an agreement among the users. The numerator $P_0 - P_e$ is the proportion of agreement beyond the probability that they trust each other, while the denominator of the fraction $1 - P_e$ is the maximum possible agreement among the number of users. This will output the coefficient, Kappa, which is the proportion of the agreement among the observers after the probability of agreement is removed.

The goal was to get at least 500 unique instances and images for the model to train on. This number of instances is chosen since this sample size is sufficient to achieve reasonable results for training models on machine learning problems [24]. This benchmark is taken into account during the execution of the experiments and when documenting the findings. Luckily, the training data set has at least 500 unique instances. This should be sufficient for the multi-modal classification model(s) to reach reasonable results when executing the experiments.

Besides, the data is extracted from a public source, which will prevent the research from dealing with ethical and confidentiality issues when it comes to relating it to a natural person that did not provide consent to this research. This data set could also be used for other research projects that want to use image-with-text data posted on Twitter. The results are also reproducible for others when interested.

3.2 Evaluation Metrics

Evaluation metrics that are used within this model are the *F1-score* and the *accuracy*. The *Area under the receiver operating characteristic* (*AUROC*) was considered in the previous iteration, but since the data set is quite imbalanced with the classes, more negative classes in comparison to positive classes, the alternative metric of *F1-score* is suggested. This is since *AUROC* is averaging out the thresholds, which is not the case when using the *F1-score* as a performance evaluation metric. But the *F1-score* and *AUROC* can be derived by calculating the outputs of the confusion matrix equations [22]:

- True Positive Rate (TPR): the proportion of the positive classes that got correctly classified by the model. It is calculated by: $TP/(TP + FN)$. TPR is also known as the *sensitivity*.

- False Positive Rate (FPR): the proportion of the positive classes that got incorrectly classified by the model. It is calculated by: $FP/(TN + FP)$.
- True Negative Rate (TNR): the proportion of the negative classes that got correctly classified by the model. It is calculated by: $TN/(TN + FP)$. TNR is also known as the *specificity*.
- False Negative Rate (TNR): the proportion of the negative classes that got incorrectly classified by the model. It is calculated by: $FN/(TP + FN)$.

The F1-score can be formally stated in the following equation:

$$F1 = \frac{TP}{(TP + (FN + FP))/2} \tag{2}$$

The other evaluation metric is *accuracy*. *Accuracy* is calculated by how accurate the predictions are and it is easier to interpret in comparison to the *F1-score*, which is also a common metric for checking the performance of the model. In this case, it can be formalised in the following equation:

$$Accuracy = \frac{T_p + T_n}{C_p + C_n} \tag{3}$$

, where T_p and T_n stand for respectively True Positives and True Negatives, while C_p and C_n stand for the Truly Positive and Truly Negative examples. All metrics are weighted on macro-level due to the imbalance of the classes.

3.3 Exploratory Data Analysis

Within the final data set, label class 0 stands for memes that contain content of *Fat-shaming*, label class 1 stands for memes that contain *Sexism* content, label class 2 stands for memes that contain *Racist* content, label class 3 represents memes that contain content of *Discrimination against mentally/physical disabilities*. Finally, label class 4 represents memes that do not contain any form of cyberbullying content (647 memes).

Deriving from the values in the column *Final_label_class_num*, the vast majority of the memes within this data set do not contain any form of content of cyberbullying. From the labels that do contain content of cyberbullying, label class 1 (*Sexist* memes) is the most common one within this data set (28 memes), followed up by label class 2 (*Racist* memes) (24 memes), label class 3 (*Discriminative meme based on mentally/physical disability*) (17 memes), and label class 0 (*Fat-shaming* memes) (6 memes). This adds up to 75 memes that contain a sentiment of cyberbullying within this data set. This means out of the 4229 tweets that were extracted in one month, approximately 2% consisted of tweets that contain some form of cyberbullying sentiment in it. A summary table of the labels and the confidence levels are stored in the JSON file 'output.json'.

The top three users who have posted the most memes in this data set are *MyEnglishLearn* (2.64% contribution), *churkakiller* (2.36% contribution), and finally *r_memes_* (1.94% contribution). While the most tweeted word within this data set is *meme*, *dankmemes*, and *memesdaily*. This could be due to the keywords that were used for extraction and filtering of the tweets that contain memes. The tweet also contains the hashtags of the mentioned words, so this adds up to this result. While the most frequent word that is captioned on the memes are *horse*, *wait*, and *scheduled*.

For more information and visualisations concerning this data set can be seen in Appendix A: Related Work and Exploratory Data Analysis.

3.4 Multi-modal classification Model

A baseline multi-modal classification model is developed in Python that manages to fulfil the objectives of this research. As mentioned in the introduction, the model should be able to process and classify textual and visual elements in the form of images with captions on them. Since most multi-modal classification models contain two encoders that capture the captions on the image and detect the contents within an image. When the image and text encoding are captured from the image-to-text data, these will be concatenated to features, which will provide an output of whether the memes contain cyberbullying content or not.

The two models that are taken into consideration for this research, are the two of the top five models that have participated in the Hateful Memes Challenge, see Section 2.2. The models had comparable results for detecting hateful content in memes in comparison to human performances. In this case, the VisualBERT model, developed by Velioglu and Rose [26], and the Vilio model, developed by Muenninghof [21]. In comparison to all the models that participated in the challenge, these models also managed to score high on the selected performance metrics, *AUROC* and *accuracy*, from this challenge in comparison to human judgement. The high-level explanation of the models is described in Section 2.

The Enhanced Multi-modal Transformer model developed by Zhu [29] was taken into consideration for this research. But the model itself was built on a premium Google Cloud environment and used computationally expensive resources to build the model. So this model could not be used for this research due to the limited resources that were used for this study. The UNITER models by Sandelescu [23] and Team Kingsterdam [18] were further off from human baselines. That is why these models were not taken into consideration as well.

For the Hateful Memes Challenge, the data set only consisted of the image, applicable label (0: being non-hateful or 1: being hateful), and the text/caption that was written on the meme. This is less extensive in features in comparison to the training set that is used for this research. In this case, it will be researched whether or not it will perform better on the extracted data set from Twitter. The data set also contains additional features that could be derived and used for the

model from the tweet objects. Finally, the best-performing multi-modal classification model will be selected and tuned to optimise the performance for this particular use case.

4 Experimental setting

This section describes the splits that are used for this research for training, testing and validation and how the experimental setting where this research is performed in.

4.1 Setup experiments

The models will be trained in multiple iterations. This experimental setting aims to improve the performance of the model based on the specified metrics in Section 3.2. At first, the models are trained on default settings. In this case, the models are compared on their performance on the validation and test sets.

As mentioned in Section 3, the goal was to get at least 500 unique instances and images for the model to train on. This number of instances is chosen since this sample size is sufficient to achieve reasonable results for training models on machine learning problems [24]. This benchmark is taken into account during the execution of the experiments and when documenting the findings. Luckily, the extracted and labelled data set has 722 unique instances, so this should be sufficient for the multi-modal classification model to reach reasonable results on validation and test sets.

Since the data set comprises 722 instances, the data set is split into three separate stratified sets (due to the imbalance) of 80% training data, 10% validation data, and 10% data. And to check whether the models perform well when less training data is used for performing the tasks, a stratified split is done that consists of 75% training data, 10% validation data, and finally 15% test data. The best results are noted, compared and reported within Table 1.

Afterwards, a second iteration is executed by including additional parameters that could enhance the performance of the model. For example, adjusting the number of iterations or adjusting the learning rate of the model. Then again, the same stratified splits for training, validating, and testing the models are used in comparison to the first setup. In this case training, validation and test split of 80%, 10%, and 10% and 75%, 10%, and 15% respectively. And again, the best results are noted down, compared, and reported within Table 2.

Based on the findings of these two experiments, the baseline model is determined by selecting the best-performing model out of the three models that are taken into consideration during the two experiments. The best-performing model is also selected as the final model. This concludes the first sub-research question. This final model will be further tuned via the applicable hyperparameters of the final model and additional features are extracted and included where possible. When performing these experiments in different conditions, the two sub-research questions of this study are answered, which will conclude this research.

The experiments are all done in Python on a Windows laptop via Jupyter Notebooks on Google Colab Pro and Python files. The notebooks are used for performing and reporting the experiments, while the Python files are used for creating new features and processing the data. This ensures that the experiments are reproducible for other people who are not involved in this research. The Jupyter Notebooks, data sets and Python files are made available which could be seen in the code repository for this research.

5 Results

In this section, the result outcomes of the sub-research questions in each sub-section are reported and discussed.

5.1 Comparison performances for baseline models (SRQ1)

Most multi-modal classification models contain two encoders that capture the captions on the image and another encoder that detects the contents within an image. When the image and text encoding are captured from the image-to-text data, these will be concatenated to features, which will provide an output of whether the meme has a cyberbullying sentiment in it.

As mentioned in Section 3, the two models that are taken into consideration for this research, are two of the top three models that have participated in the Hateful Memes Challenge. The models had comparable results for detecting hateful content in memes in comparison to human performances on the Hateful Memes data set. But the two models, in this case, are the VisualBERT model developed by Velioglu and Rose [26], and the Vilio model developed by Muenninghof [21].

The models are evaluated on the *F1-score* and *accuracy* during two iterations. The results can be seen in Tables 1 and 2. As an additional metric, the *AUROC* metric has also been added to provide a complete picture of how the models perform across the classification metrics since this was also the main evaluation metric that the models got evaluated on during the Hateful Memes Challenge. But is not the main evaluation metric that the models are evaluated on concerning performances. All the results are coming from a train-test-validation split of 80%, 10% and 10% respectively. The models did not perform better when fewer data was used.

As noticed in Table 1, all the models did manage to retrieve a high *accuracy* score on this data set but scored poorly on the *F1* metric. The default models are optimised on the *accuracy* and *AUROC* and not on the combination of *accuracy* and *F1* metric. The default VisualBERT model did not manage to excel when it was evaluated on this metric. But as seen

in Table 2, the VisualBERT model managed to get a slightly higher score across the different metrics.

Vilio on the other hand, performed quite consistent across all the metrics when using the default and the optimised model. The model got modified a bit to also track the *F1-score* during each epoch. When averaging it out by using the simple average of Vilio, this resulted in the highlighted figures within the tables. But the limitations of the baseline models are still there regarding the *F1-score* since they are still relatively low.

The Tensorboard validation graphs and training validation tables can be seen in Appendix 4.1: Model Performances.

5.2 Select final baseline model (SRQ1a)

When comparing the performances of the two selected models across the *accuracy* and *F1-score* metrics, the Vilio model performed the best on the validation and test sets during the default settings. The *F1-score* and *accuracy* scores are also slightly higher for Vilio during iteration 2. The model managed to increase the performance quite well. Whereas the VisualBERT model needed to run an 8-hour-long grid search to find the model that performed relatively mediocre on the unseen test and validation sets. Even though the *F1-score* increased significantly when doing the hyperparameter search, it is an extensive run to increase the performance of this model for the selected evaluation metrics. For Vilio on the other hand, training the model with the optimised parameters was quicker in comparison to the VisualBERT model (4h versus 8h). And it managed to generalise better across all the performance metrics in comparison to the VisualBERT model. That is why the Visio model is considered as the final model over the VisualBERT model. But even though the models managed to optimise performance concerning *accuracy*, the *F1-score* scores of both models are still considered poor. This will be addressed in the next sub-research question.

5.3 Optimise final baseline model (SRQ2)

For answering this sub-research question, the Vilio model is selected and optimised across the metrics for performing the task of detecting and classifying cyberbullying content. This is done by creating additional features out of the final data set and by further optimising the hyperparameters which are applicable for this use case. The outcome of this question will result in the final model that can be useful for detecting and classifying internet memes that are targetting individuals on the Twitter platform.

As mentioned in Section 2, additional features are created from the image and text itself. Within the image, the objects within the image are identified. This is also done for the final Vilio model. By using the *Detectron2* library [28], objects and faces could be detected from the image when the textual captions were removed from the image. This was more difficult to do for the images that are used within this study. As mentioned before, the images are of lower quality in comparison to the images of the Hateful Memes Challenge, which makes the images more blurry and distorted. This also affects the text itself for identifying whether the captioned text has cyberbullying sentiment in it. By extracting these features, incorporating them into the model and finally optimising the Vilio model, the performance slightly increased in performances across the selected metrics. This resulted in an *F1-score* of 66.67 and an *accuracy* of 91.03. One of the main contributors is to decrease the *weight_decay* within the model to regularise the model. This will prevent the model to overfit the training data by penalising this behaviour. The final parameters can be seen in Appendix 4.1: Model Performances.

6 Discussion

For this study, an attempt was performed to use two of the top five models of the Hateful Memes Challenge, It was verified if they perform well in a real-life case by extracting Twitter data that contain some form of cyberbullying content in it within the form of memes. Limited studies have been performed on this topic, especially on the Twitter platform in comparison to other social media platforms. The top two frameworks that were chosen and used for this research, performed well across the *accuracy* metric, but could not excel when evaluating on the *F1-score*. The possible assumptions could be due to the imbalanced data set of each cyberbullying class, which contains more negative examples in comparison to positive examples that have some form of cyberbullying sentiment within the memes. For calculating the *F1-score*, the score difference between the best model and default model is relatively big due to the imbalance of positive classes in comparison to negative classes. In comparison to the *AUROC*, where the performance difference is smaller since it is focused on ranking predictions. Not on whether the output is well-calibrated probabilities [22].

Another reason could be that the models were used and optimised for the Hateful Memes Challenge, which uses *AUROC* and *accuracy* as evaluation metrics [15, 16]. This resulted in a decrease in the *F1* metric. The data set of this challenge was more balanced and contained clean examples of memes with hateful content in them. These two metrics were appropriate for the data set that were used for the applicable challenge. While the data set that is used for this research, which reflects real-life examples of memes that are tweeted by users on Twitter, only contains a small sample set of positive examples of cyberbullying content. And this content was not represented within the same format as the memes that were selected for the Hateful Memes Challenge. The memes that were used for the mentioned challenge were mainly constructed in the same format of a high-quality image with a white bold caption on it. The examples used for this research are memes that come in different shapes and

Table 1. Model performances without parameter tuning (default settings)

Comparing Iteration 1 - Without features						
	Validation			Test		
Model	Accuracy	F1-score	AUROC	Accuracy	F1-score	AUROC
VisualBERT	91.43	00.00	77.08	88.57	00.00	67.70
Vilio	86.50	31.58	73.86	85.24	25.93	59.80

Table 2. Model performances with parameter tuning

Comparing Iteration 2 - Including parameter tuning						
	Validation			Test		
Model	Accuracy	F1-score	AUROC	Accuracy	F1-score	AUROC
VisualBERT	89.01	28.48	88.57	80.55	33.33	87.18
Vilio	90.50	40.00	81.06	90.05	34.18	87.11

forms where the captions were put in different locations. In addition, different fonts were used to highlight the memes. This could make it more difficult for models to detect and predict the cyberbullying sentiment of the meme in question.

6.1 Limitations

One of the limitations of the study is the imbalance of the data set. The sample that is extracted from Twitter contained a hugely skewed imbalance of non-harmful memes. This was taken into account when using the *F1-score* over the *AUROC* metric, which is less appropriate to use as an evaluation metric within an imbalance data set. According to the annotators, a small number of instances within this data set contained cyberbullying content. This made the prediction results slightly inaccurate when the models were predicting cyberbullying content from the applicable memes. This could have been resolved by doing additional runs of data extraction and data labelling again. But due to time limitations, this potential solution could not be performed. Another solution that was attempted was to use an external data set where labels were annotated according to the same labels that were used for this research. But this type of data set that was specifically focused on memes, could not be found on the websites from different data suppliers.

Besides, during the time of extracting and composing the data set, the extracted Twitter data set contained plenty of samples that were promoting certain cryptocurrencies as memes. Mostly focused on buying certain cryptocurrencies in the form of $AZULA and SamoCro. Maybe other keywords or keyword filtering approaches could have resulted in the type of tweets to resolve the meme imbalance. This would have resulted in a more balanced data set in comparison to the currently used data set. Another possible reason behind this issue was the time of extraction. It could be that the timing was inconvenient for extracting the specific types of tweets that were needed to resolve the imbalance within the data set.

Another limitation of this research is the selected frameworks. The selected frameworks of the models are optimised on the evaluation metrics of *AUROC* and *accuracy*. Since the selected models that are compared for this research are evaluated on the *F1-score* (due to the class imbalance of the extracted data set where *AUROC* would not be suitable for), and *accuracy*, this resulted in deficiencies within the results when comparing the models on the selected performance metrics. The models do not manage to generalise enough across all the performance metrics, since a high *AUROC* score does not imply that the *F1-score* would be high as well.

7 Conclusion

In this report, a study was performed to research state-of-the-art multi-modal classification models and improve the best-performing model that can classify and detect memes with elements of cyberbullying in it on the extracted Twitter data. The study aims to answer whether an image-with-text classification method could identify and detect cyberbullying memes that are posted on the Twitter platform. Limited studies have been performed in this research field in comparison to other social media platforms, Even though different studies have shown that cyberbullying has also been happening on the Twitter platform in the form of memes. Displaying content in memes makes it more difficult to detect whether a tweet is considered offensive or not or whether they are targeting individuals on the internet. By addressing this problem, this study could contribute to solving one of the many problems that are happening within Twitter, which is creating a healthy social media environment for their users to share their opinions on, without harming individuals.

To tackle this problem, a sample of tweets on Twitter was extracted that contained memes and used as input data for the state-of-the-art models to train on. These models were compared and evaluated on their performances for predicting whether the meme has cyberbullying sentiment in it. The

scope of the cyberbullying classes was on fat-shaming, sexism, racism, physical and mental disability discrimination, or if they do not contain any of these four classes. The results show that the default and tuned Vilio models performed better across the *F1* and *accuracy* metrics in comparison to the architecture of the VisualBERT models. Even though the performances from the models were relatively low on the *F1* metric. This could be due to the optimisation of these models on the *AUROC* and *accuracy* metrics instead of the metrics that were used for this research. Another reason to choose the Vilio model over the VisualBERT model was that the Vilio model manages to generalise better over the highlighted metrics in comparison to the developed VisualBERT architecture. In the end, the Vilio model was optimised by extracting image objects and text sentiments from the memes as well as by finetuning the hyperparameters. This with the aim to increase the performance of the model and that the model could generalise better on the *F1-score* as well. The final model manages to achieve a *F1-score* of 66.67 and a *accuracy* of 91.03 on the unseen test set. This is a significant increase in the *F1* performance in comparison to tuned model in Table 2, but still mediocre performances when it comes to human benchmarks. This concludes that solving this particular problem is still rather complicated with the existing frameworks, but that the developments are there to resolve this problem soon.

7.1 Future work

Further research could be done to adjust the multi-modal classification models to different languages. Since the scope of this research is mainly focused on the English language. Attempts could be made to create a model that detects cyberbullying content in other languages, such as Spanish, Hindi, Italian, or Mandarin. There have been different machine learning studies conducted on non-English texts. But the number of studies remains limited. Which makes it harder to address harmful content on social media if the tweet and meme have been tweeted in a different language other than English. Another problem that could be addressed is the models. The current models that are compared and used for this research are multi-modal classification models. To go one level further is to turn this binary multi-modal classification job into a multi-classification task. The task is then to classify which type of cyberbullying it is given a tweet with a meme. The input would be the same as the current research problem, which is the image and text together, but the target is to classify it into the four classes of cyberbullying that are scoped in this research.

An attempt was taken for this research to treat the problem as a multi-classification problem instead of a binary multi-modal classification problem since the labels are available within this data set. This would help with detecting the sentiment of each meme instead of classifying it based on whether the content of the meme contains cyberbullying sentiment or not. But due to time constraints to set up the models to treat the data set as a multi-classification problem, it could not be covered within this research. Also, the imbalance between the different classes across the data set would not be sufficient to train a model properly on multi-classification tasks, since there were more negative classes in comparison to the number of positive classes in each specific cyberbullying class.

References

[1] 2021. Cyberbullying, social stigma, and self-esteem: the impact of COVID-19 on students from East and Southeast Asia at the University of Jordan. *Heliyon* 7, 4 (2021), e06711.

[2] Petra Ambrožová and Martin Kaliba. 2021. Teacher-shaming in the context of Czech distance learning due to COVID-19 pandemic. In *Proceedings of EDULEARN21 Conference*, Vol. 5. 6th.

[3] Denis S Artamonov, Sophia V Tikhonova, and Marina L Volovikova. 2021. Mythologizing Time in Internet Memes of the COVID-19 Pandemic Period. In *2021 Communication Strategies in Digital Society Seminar (ComSDS)*. IEEE, 154–157.

[4] Imran Awan. 2016. Islamophobia on Social Media: A Qualitative Analysis of the Facebook's Walls of Hate. *International Journal of Cyber Criminology* 10, 1 (2016).

[5] Kate Barnes, Tiernon Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. 2021. Dank or not? Analyzing and predicting the popularity of memes on Reddit. *Applied Network Science* 6, 1 (2021), 1–24.

[6] Christian Bauckhage. 2011. Insights into internet memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 42–49.

[7] Kenneth J Berry and Paul W Mielke Jr. 1988. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement* 48, 4 (1988), 921–933.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[9] Patrick Davison. 2012. The language of internet memes. *The social media reader* (2012), 120–134.

[10] Richard Dawkins and Nicola Davis. 2017. *The selfish gene*. Macat Library.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Mindi D Foster. 2015. Tweeting about sexism: The well-being benefits of a social media collective action. *British Journal of Social Psychology* 54, 4 (2015), 629–647.

[13] Shivani Gupta and Atul Gupta. 2019. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science* 161 (2019), 466–474.

[14] Ali Kandeger, Hasan Ali Guler, Umran Egilmez, and Ozkan Guler. 2018. Cyberbullying: A virtual offense with real consequences. *Indian Journal of Psychiatry* 59, 4 (2018), 2017–2018.

[15] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. 2021. The hateful memes challenge: competition report. In *NeurIPS 2020 Competition and Demonstration Track*. PMLR, 344–360.

[16] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

Advances in Neural Information Processing Systems 33 (2020), 2611–2624.

[17] knowyourmeme.com. 2021. Dank Memes. https://knowyourmeme.com/memes/dank-memes [Online; posted 1-January-2014].

[18] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A Multimodal Framework for the Detection of Hateful Memes. *arXiv preprint arXiv:2012.12871* (2020).

[19] Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media* 22, 2 (2021), 205–224.

[20] Simangele Mkhize and Nirmala Gopal. 2021. Cyberbullying perpetration: Children and youth at risk of victimization during Covid-19 lockdown. *International Journal of Criminology and Sociology* 10 (2021), 525–537.

[21] Niklas Muennighoff. 2020. Vilio: state-of-the-art Visio-Linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788* (2020).

[22] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[23] Vlad Sandulescu. 2020. Detecting Hateful Memes Using a Multimodal Deep Ensemble. *arXiv preprint arXiv:2012.13235* (2020).

[24] Saleh Shahinfar, Paul D. Meek, and Gregory Falzon. 2020. How many images do I need? Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *CoRR* abs/2010.08186 (2020). arXiv:2010.08186 https://arxiv.org/abs/2010.08186

[25] Fatemeh Tahmasbi, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 2021. "Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *Proceedings of the web conference 2021*. 1122–1133.

[26] Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975* (2020).

[27] Isaac Warbrick, Heather Came, and Andrew Dickson. 2019. The shame of fat shaming in public health: moving past racism to embrace indigenous solutions. *Public Health* 176 (2019), 128–132.

[28] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[29] Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290* (2020).

A Related Work and Exploratory Data Analysis

A.1 Data set Analysis

Who are the most frequent users that are tweeting memes frequently within this data set? The answer is visualised in Figure 7, which is a barplot of the top 20 users who have tweeted within this data set. According to this visualisation, the top three Twitter users who have posted the most memes in this data set are *MyEnglishLearn* (2.64% contribution), *churkakiller* (2.36% contribution), and finally *r_memes_* (1.94% contribution). The distribution of tweets is quite equally distributed. There is not a user who posted significantly more memes in comparison to the other users.

A.2 Text Analysis

For this analysis, the tweets will be analysed more in-depth. For example, what are the top 50 most tweeted words within this data set? The answer can be derived from the word cloud of Figure 8. *Meme*, *Dankmemes* and *Memesdaily* are the most highlighted words in this word cloud. This makes sense since the filter of the tweets was applied by using the hashtags containing these three words. These three keywords also gathered the most tweets that contained image-with-text memes.

The most tweeted words in this data set are *memes*, *meme*, and finally *rt*, which is the abbreviation for retweet. The tweets containing the words *memes* and *meme* are significantly more tweeted in comparison to the other words. This can be seen in Figure 9.

A.3 Image-with-text Analysis

For this subsection, exploratory analysis is performed on the image-with-text data part of this data set, since the model for this research should be able to process and classify textual and visual elements in the form of images with captions on them.

The *pytesseract* library is used for extracting the text from the images. This is a wrapper for Google's Tesseract-OCR engine (OCR stands for optical character recognition). It will recognise most of the words in the images and extract the text which is embedded in the images. The limitation of this library is that it is not able to extract all the words from the images, but it is currently used for EDA purposes. In this case, the text was extracted from the images and the most used words are visualised in the word cloud of Figure 10.

As noticed in the word cloud, some non-English/weird words have slipped into the image-with-text data. This is something that needs to be taken into account when training the models on these images. Or else it will be difficult to label the data on the sentiment of the meme when it contains these types of words. But this is also the limitation which was mentioned in the previous paragraph when using the *pytesseract* library for the extraction of the captions from the image-with-text data.

Within the memes, there were plenty of symbols found in the images. This can be seen in the bar plot of Figure 11. It could indicate that most of the memes have unidentifiable characters in the image-with-text data. This could also be due to the limitation of the *pytesseract* library. As mentioned before, the library is not capable to extract all the words from the images, due to the complexity of the backgrounds. Memes also consist of images, which makes the characters difficult to read when the background is distorted or complex.

A.4 Conclusion

This section described the EDA procedure on a data set containing image-with-text data. Interesting findings that have

Table 3. Overview of the final models in the Hateful Memes Challenge

Model	Classifier type	AUROC	Accuracy
Human judgement	NA	82.65	84.70
Enhance Multi-modal Transformer	Binary	84.50	73.20
Vilio	Binary	83.10	69.50
VisualBERT	Binary	81.08	76.50
UNITER	Binary	80.53	73.85
UNITER$_{PA}$	Binary	79.43	74.30

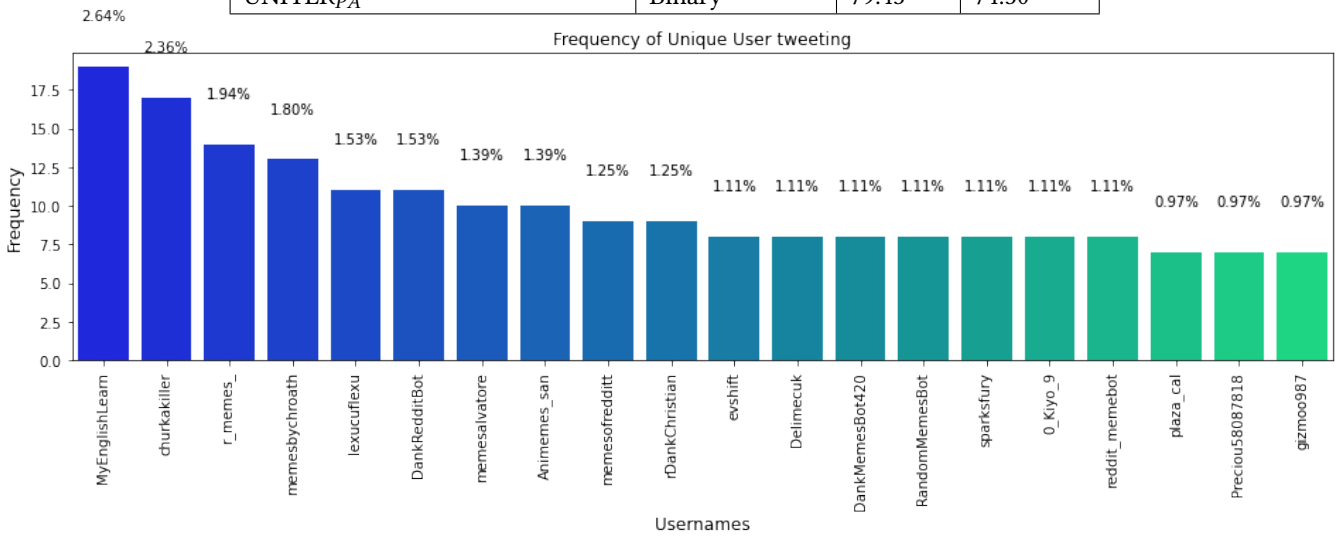

Figure 7. Unique users

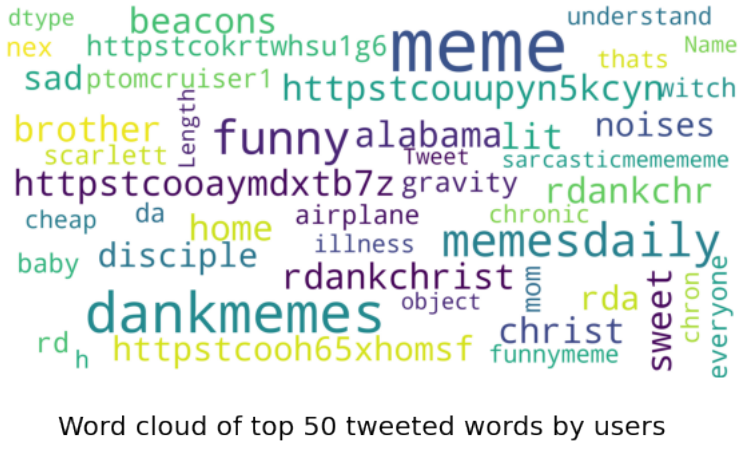

Figure 8. Word cloud of the top 50 most tweeted word

been found in the data set are the fact that the number of users that have contributed to this data set is quite equally distributed with regards to tweets. It is also difficult to tell whether the extraction of the text from the image-with-text data went correctly since the *pytesseract* library has its limitations when backgrounds are lacking sharpness or having quite chaotic backgrounds where the text is imprinted in the images. But the vast majority of the words could be extracted by using this library. This could be used to get more familiar with this extracted data set.

B Image Labelling Job

In this section, the image labelling procedure is described for annotating the applicable five classes of memes for this

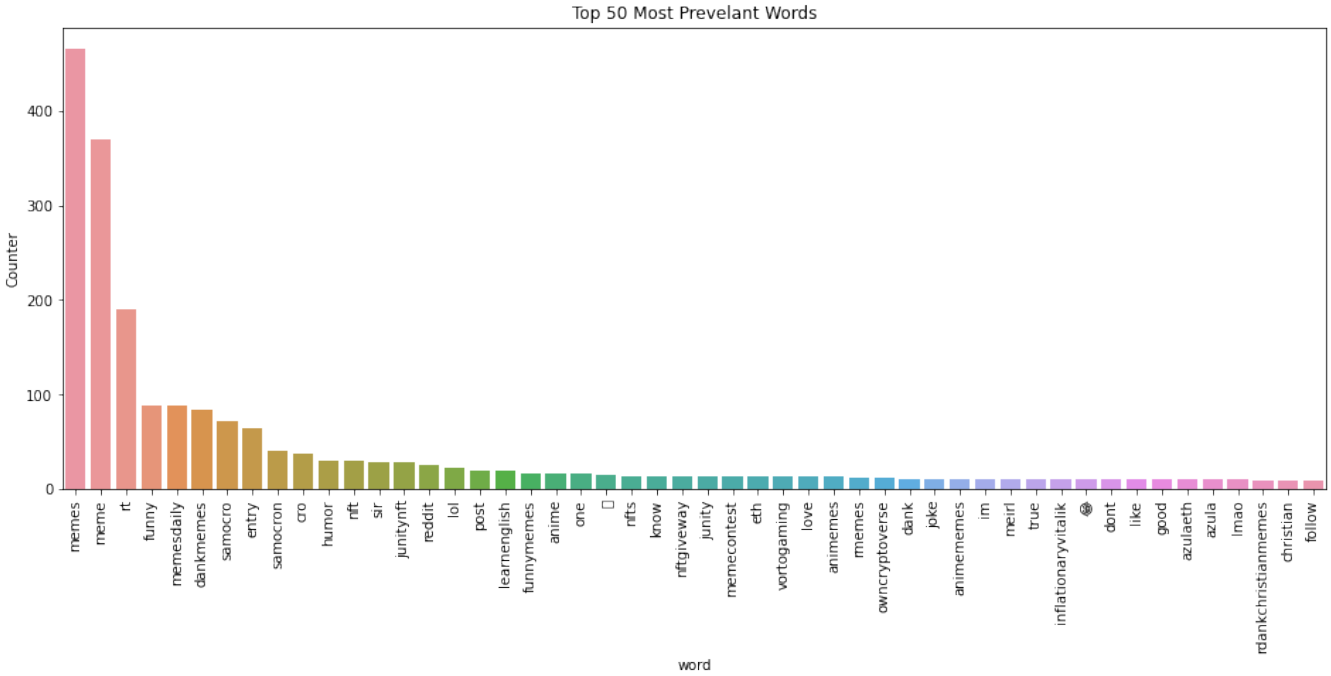

Figure 9. Frequency of words

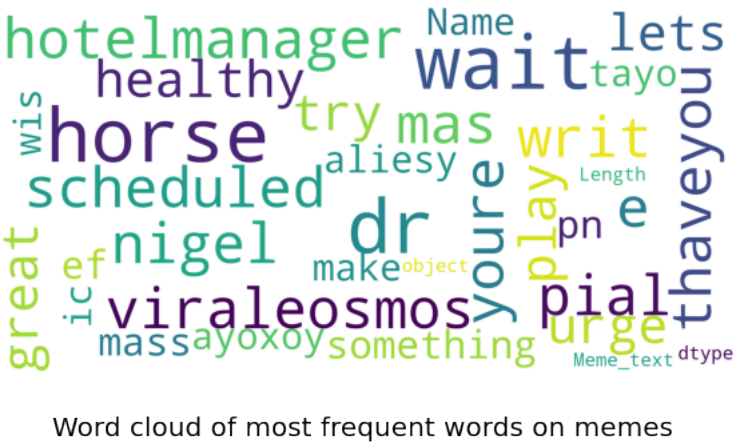

Figure 10. Word cloud of the memes

research. The labelling job was created by using Amazon SageMaker Ground Truth. Five independent (human) workers were requested to get the labelling job done. 722 instances were labelled according to the predefined instructions of the applicable classes (including the class that was neither of the four cyberbullying classes). An example of what the workers saw can be seen in Figure 12. Within this example, the purpose and the research scope were described to provide the workers with some additional context about this project.

On the left-hand side of the dashboard, the good and bad examples of each class were provided for the workers to distinguish between these classes. If it is a good example, then it is the applicable cyberbullying class. If it is a bad example, then it is neither of the cyberbullying classes. On the right-hand side of the dashboard, the worker could select one or more classes for the applicable meme. This resulted in a JSON output with the following columns:

- **Source-ref**: the applicable object that was labelled (in this case the .jpg file of the image).
- **Class-name**: the name of the applicable class.
- **Labelling-job**: the name of the labelling job (in this case *cyberbullying-labelling-job*).

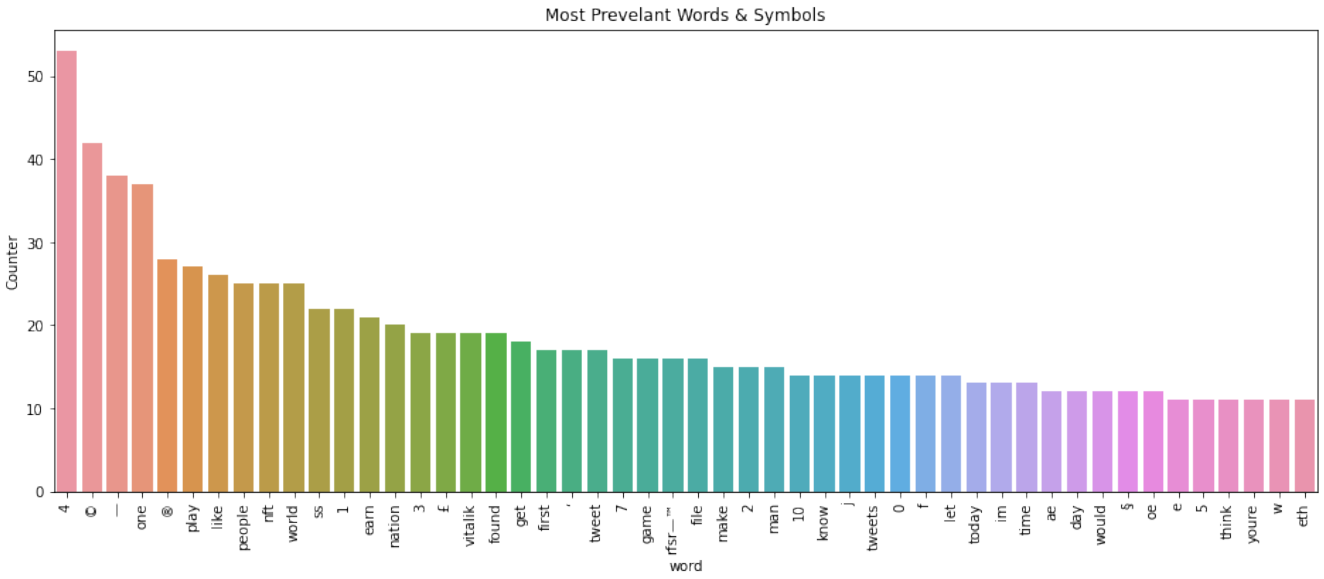

Figure 11. Frequency of words and symbols within the memes

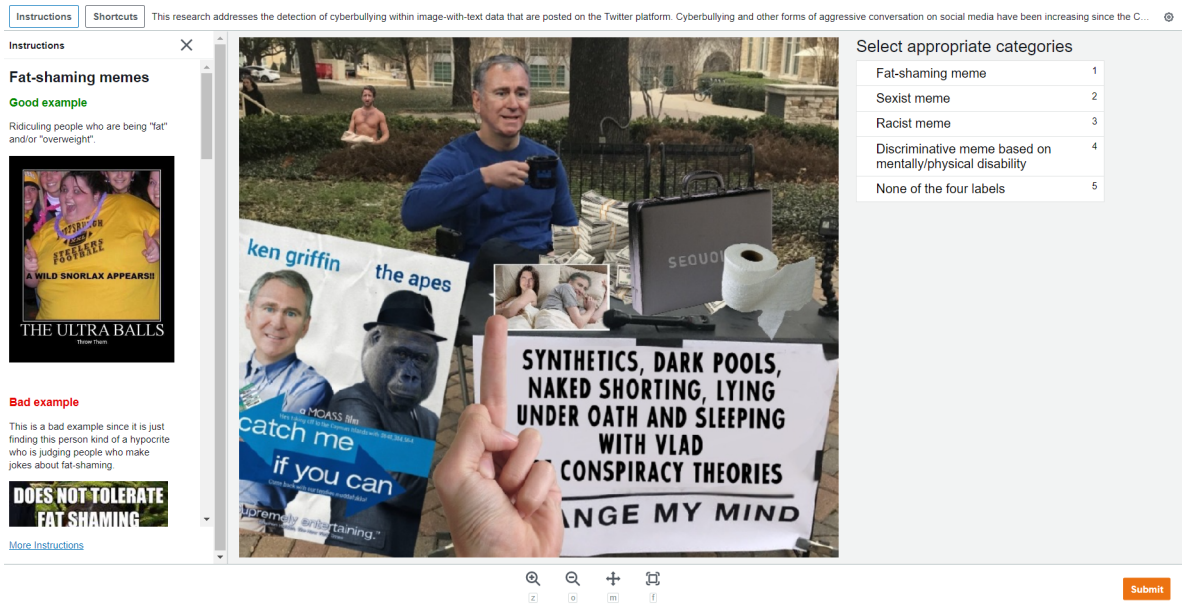

Figure 12. Labelling dashboard in Amazon SageMaker Ground Truth

- **Confidence-map**: the confidence score from the class.
- **Human-annotated**: if the labelling job was annotated automatically or manually (in this case the settings were set on "humans").
- **Creation-date**: when the labelling job was completed for each image-with-text data.

The confidence score was outputted for each image by Amazon SageMaker Ground Truth. A confidence score is a number between 0 and 1 that indicates how confident the tool is in the label for the applicable image. The confidence score is used to compare labelled data objects to each other, and to identify the least or most confident labels of each image.

The highest confidence score would result in the final label for the applicable image-with-text data. If one of the cyberbullying labels were indecisive. In this case, if some workers annotated the meme as racist while others find it a discriminative meme towards people with a mental and/or physical disability, the final label was decided based on the

annotation with the highest confidence/agreement score. This is also known as the Cohen's Kappa level.

C Model Performances

C.1 VisualBERT and Vilio

Final parameters for the optimised Vilio model:

- decay_steps: 13308;19962
- lr_decay_ratio: 0.1
- num_train_steps: 500
- save_steps: 125
- warmup_steps: 50
- batch_size: 4
- valid_steps: 2000
- lr_rate: 1e-5
- weight_decay: 0.001
- max_len: 128

See the next page for the training and validation figures for each model.

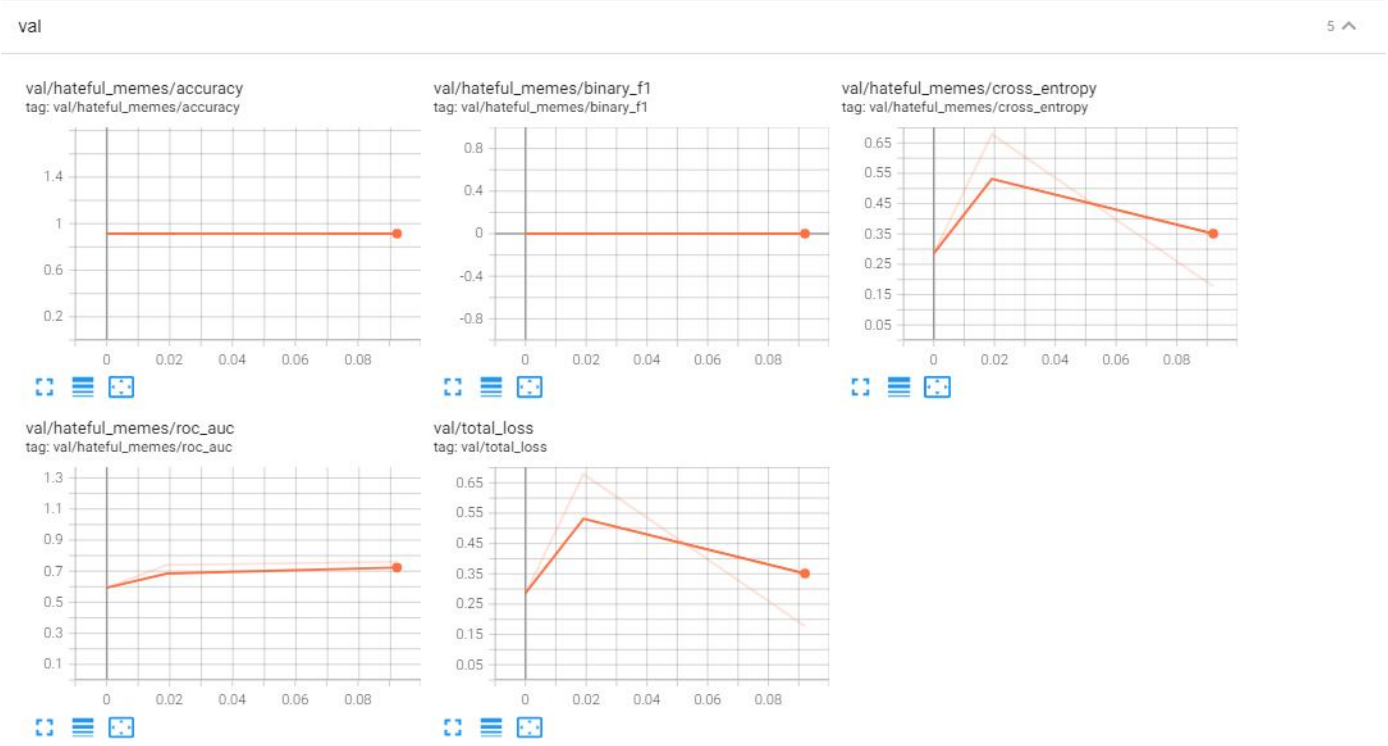

Figure 13. Iteration 1 - VisualBERT

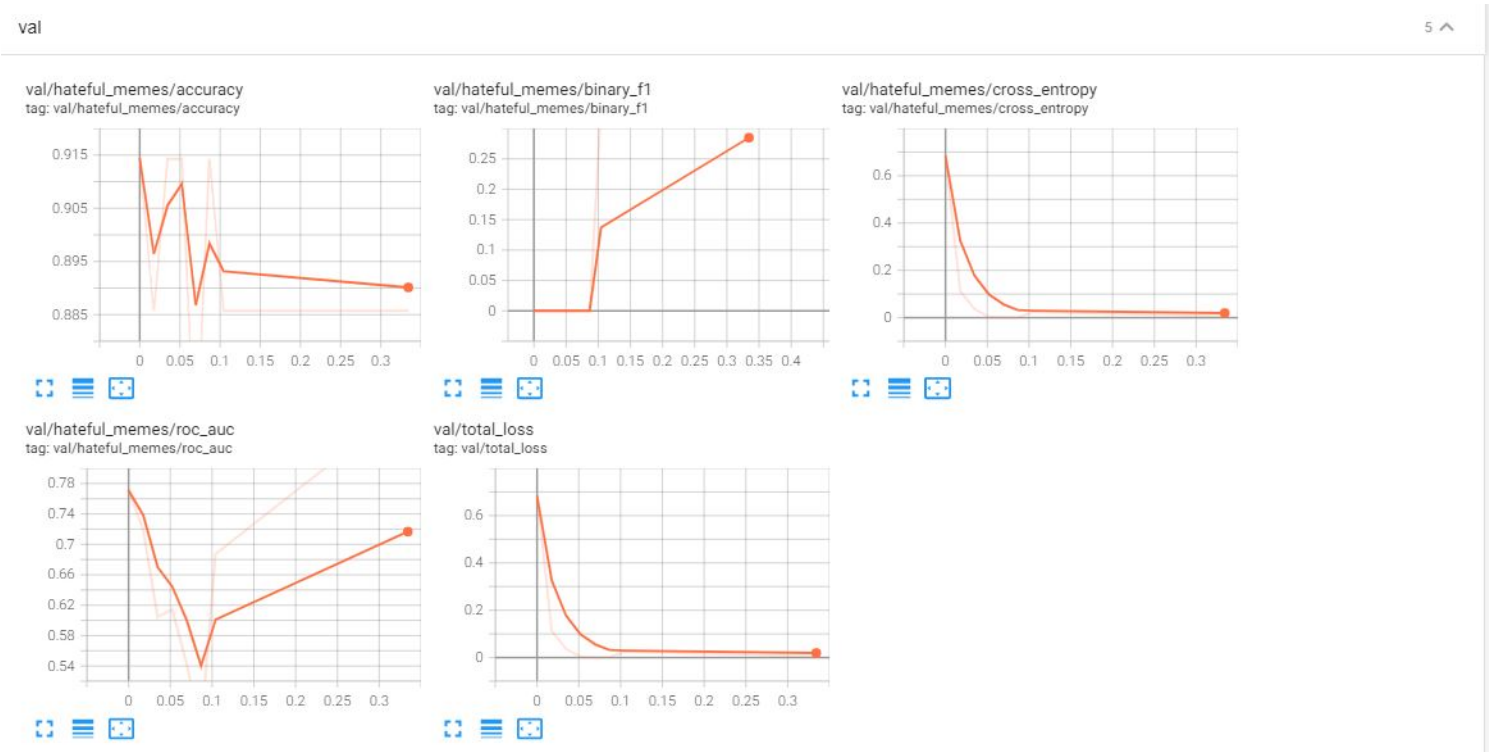

Figure 14. Iteration 2 - VisualBERT

Table 4. Validation Iteration 1: Vilio

Vilio model 1 - Accuracy: 86.50, AUROC: 73.86, F1: 31.58						
	Validation			**Test**		
	Steps	**Loss**	**Accuracy**	**Steps**	**Loss**	**Accuracy**
1	10	0.6456	1.00	260	0.0050	1.00
2	20	0.5727	1.00	270	0.0002	1.00
3	30	0.4174	1.00	280	0.0002	1.00
4	40	0.2628	1.00	290	0.0008	1.00
5	50	0.1060	1.00	300	1.2892	0.75
6	60	1.0992	0.75	310	0.0269	1.00
7	70	1.4968	0.75	320	1.0632	0.75
8	80	1.5769	0.75	330	0.0060	1.00
9	90	0.0222	1.00	340	0.0087	1.00
10	100	2.2430	0.50	350	0.0012	1.00
11	110	0.9855	0.75	360	1.4998	0.50
12	120	0.0035	1.00	370	0.0019	1.00
13	130	0.9950	0.75	380	2.2768	0.75
14	140	1.1411	0.75	390	0.0053	1.00
15	150	0.0736	1.00	400	0.1572	0.75
16	160	0.7642	0.75	410	0.1828	1.00
17	170	0.0367	1.00	420	0.0001	1.00
18	180	0.9684	0.75	430	1.0739	0.75
19	190	1.6699	0.50	440	0.2597	0.75
20	200	0.0201	1.00	450	1.7399	0.75
21	210	0.8610	0.75	460	0.0181	0.50
22	220	0.0043	0.75	470	0.0011	1.00
23	230	0.0141	1.00	480	0.0009	1.00
24	240	1.0398	0.75	490	0.3116	0.75
25	250	0.0506	1.00	500	0.0002	1.00

Table 5. Validation Iteration 2: Vilio

Vilio model 2 - Accuracy: 90.50, AUROC: 81.06. F1: 40.00						
	Validation			**Test**		
	Steps	**Loss**	**Accuracy**	**Steps**	**Loss**	**Accuracy**
1	10	0.7389	1.00	260	0.0031	1.00
2	20	0.4891	1.00	270	1.3818	0.50
3	30	0.1553	1.00	280	0.0313	1.00
4	40	0.0147	1.00	290	0.7805	0.75
5	50	0.0033	1.00	300	1.2222	0.75
6	60	0.0038	1.00	310	0.0149	1.00
7	70	0.7589	0.75	320	0.7564	0.50
8	80	0.0270	1.00	330	0.0047	1.00
9	90	0.0004	1.00	340	0.0022	1.00
10	100	0.0058	1.00	350	0.1647	0.75
11	110	1.2551	0.50	360	0.0005	1.00
12	120	0.0182	1.00	370	0.0043	1.00
13	130	1.4508	0.75	380	0.0002	1.00
14	140	0.0091	1.00	390	0.0004	1.00
15	150	1.2646	0.75	400	0.5388	0.50
16	160	0.0252	1.00	410	1.0212	1.00
17	170	1.3599	0.75	420	0.0001	1.00
18	180	0.0069	1.00	430	0.0001	1.00
19	190	0.0037	1.00	440	0.0017	0.75
20	200	0.0034	1.00	450	0.0018	0.75
21	210	0.8013	0.75	460	0.0008	1.00
22	220	0.0060	1.00	470	0.1271	0.75
23	230	1.4389	0.75	480	0.0001	1.00
24	240	0.0182	1.00	490	0.0001	1.00
25	250	0.0057	1.00	500	0.0006	0.75