Memes as Hate Speech:
Exploring Multimodality in the Context of Offensive Online Culture

Saga Hansson

Abstract

This paper explores classification of hate speech memes, focusing on two studies on the matter. A classification model developed by Oriol et al. (2019) is tested on a dataset created by Kiela et al. (2020). The model achieves a testing accuracy of 70.1%, compared to the dataset benchmark of ∼64% reported by Kiela et al. (2020), and the validation accuracy of 83.3% noted by Oriol et al. (2019). It is concluded that the dataset does not seem to reach its aim in being multimodal, and that results reported by Oriol et al. (2019) can not be fully reproduced. Other matters pertaining to hate speech in data are also discussed, including potential ethical issues in data collection.

1 Introduction

Comprised of images and captions written across the images, memes are a a widely used tool for expressing humour, as well as opinions and ideas. Memes are used in all corners of the internet, from highly criticised forums to reputable news outlets. As the format is used by a wide variety of people on an equally wide variety of online platforms, it is inevitable that some memes are created with malice and are used to spread hate.

Automatically removing memes that contain hate speech could reduce their spread, thus creating a safer online space for everyone. Fortunately, the task of reducing online hate speech with the help of machine learning is rather popular.

Oriol et al. (2019) has tackled the task of detecting multimodal hate speech by creating a classifier trained on memes downloaded from Google and from the Reddit Memes dataset,[1] concluding that the data is in fact not multimodal as the classifier relies heavily on the image data. A comparable model using similar data is created by Pranesh and Shekhar (2020). Both Gomez et al. (2020) and Suryawanshi et al. (2020) have created multimodal datasets for hate speech detection: Gomez et al. (2020) created the MMHS150K Dataset by manually annotating 150K tweets, where each tweet is composed of text and an image. Suryawanshi et al. (2020) constructed the MultiOFF Dataset by annotating memes from an already existing dataset, the 2016 U.S. Presidential Election memes,[2] as hateful or not. A concise overview of previous work on detecting unimodal offensive content (either text or images) is also provided by Suryawanshi et al. (2020). Another, more extensive overview on detecting hate speech in text is given by Schmidt and Wiegand (2017), mainly focusing on textual features. Particularly interesting work on textual hate speech is conducted by Ross et al. (2017) who examine annotator reliability and binary labelling of hate speech, concluding that agreement between annotators is low, and suggest that using multiple labels for each item or viewing hate speech as a continuous issue might present a more reliable picture. Further previous work on textual hate speech include Fernquist et al. (2019), who examine the possibility of detecting hate speech in Swedish by comparing pre-trained language models with a baseline support vector machine (SVM), finding that the pre-trained models outperform the SVM. In creating their hate speech dataset, a continuous scale was used as opposed to binary labels, following the suggestion of Ross et al. (2017). Finally, Kiela et al. (2020) has developed the Facebook Hateful Memes dataset, aiming to force classification models to be multimodal. A competition was also launched, with the top models achieving accuracies of approximately 70-75%.

In this project, the hate meme classification model by Oriol et al. (2019) is put to the test by ap-

[1] Reddit Memes dataset on Kaggle

[2] 2016 U.S. Presidential Election Memes on Kaggle

plying it to one of the recently developed datasets, The Hateful Memes Dataset (Kiela et al., 2020). The two main goals of this project are 1) to examine whether or not the results reported by Oriol et al. (2019) can be reproduced with another dataset, and 2) to investigate whether or not the dataset created by Kiela et al. (2020) reaches its aim in being multimodal.

The report is structered as follows: in section 2, the approach by Oriol et al. (2019) and the dataset created by Kiela et al. (2020) are presented, followed by a short discussion about hate speech. A description of the model and data variations used in the experiments of this report is then given. Section 3 details the results, while section 4 discusses the results and other matters pertaining to the models and datasets created by Oriol et al. (2019) and Kiela et al. (2020). Finally, the conclusion and some suggestions on further work are presented in section 5.

2 Materials and methods

2.1 Hatespeech in pixels

Oriol et al. (2019) created a model that classifies memes as either hate speech or not hate speech by combining visual and textual features. The meme image is encoded using VGG-16, whereas the meme text is encoded using BERT, which are then concatenated, creating a multimodal representation. Subsequently, this representation is fed through a multi-layer perceptron consisting of two linear layers, each followed by a ReLU activation, and finally through the output layer. The architecture of the model also allows the use of just one modality (either image or text) or both (Oriol et al., 2019).

As the research into multimodal hate was quite limited at the time, Oriol et al. (2019) gathered data in two ways. For the neutral, non-hate memes, a dataset from Reddit[3] of 3,325 memes was used, as it was assumed not to contain any hate speech. The hateful memes were gathered by using a tool[4] that queries and downloads images from Google. Employing the tool to perform 3 queries (*racist meme*, *jew meme*, *muslim meme*), a total of 1,695 memes were obtained and labeled as hateful, leading to a class imbalance of 66%-34%, where the larger class is non-hateful memes. Since the created dataset is on the small side, it was not split into train, validation and test, but only into train (containing 4,266 items) and validation (containing 754 items) (Oriol et al., 2019). The dataset is not made available.

Oriol et al. (2019) report a validation accuracy of 83.3% when using both image and text as input, 83.0% when using only image as input, and 76.1% when using only text as input, drawing the conclusion that text has little beneficial impact on the performance of the model. The difference in performance between the image-only and text-only models is attributed to the higher dimensionality of the image representation (4,096, as opposed to the 768 of the text representation).

2.2 The Hateful Memes Dataset

Launched as a part of a challenge, the dataset[5] created by Kiela et al. from Facebook AI (2020) contains 10k memes, gathered through an extensive process. Starting with over 1 million images, the process involves filtering out images that are not memes, duplicates, memes in languages other than English, memes containing "self injury or suicidal content, child exploitation or nudity, calls to violence, adult sexual content or nudity, invitation to acts of terrorism and human trafficking" (Kiela et al., 2020, p. 4) and memes containing slurs. Memes containing slurs were removed, as these are not multimodal – the meme text would be enough to classify such a meme as hateful.

In order to avoid issues regarding licensing, the memes are reconstructed by switching the images of the memes to images that are as similar as possible. Memes with images that did not have an appropriate replacement image were removed.

Following filtering and reconstruction, the memes were annotated by three annotators with either of the following: "a 1 indicating definitely hateful; a 2 indicating not sure; and a 3 indicating definitely not-hateful" (Kiela et al., 2020, p. 4). If the annotators did not agree on a label, experts were consulted.

In creating the dataset, each offensive meme is, if possible, given at least one "benign confounder" (Kiela et al., 2020). These benign confounders are variations of the offensive meme, made benign by either swapping the image or the text. An example of this, from Kiela et al. (2020) is shown in Figure 1. The purpose of the benign confounders is to force the model to utilize both modalities when

[3] Reddit Memes dataset on Kaggle
[4] Google Images Download on GitHub
[5] Facebook Hateful memes dataset on Kaggle

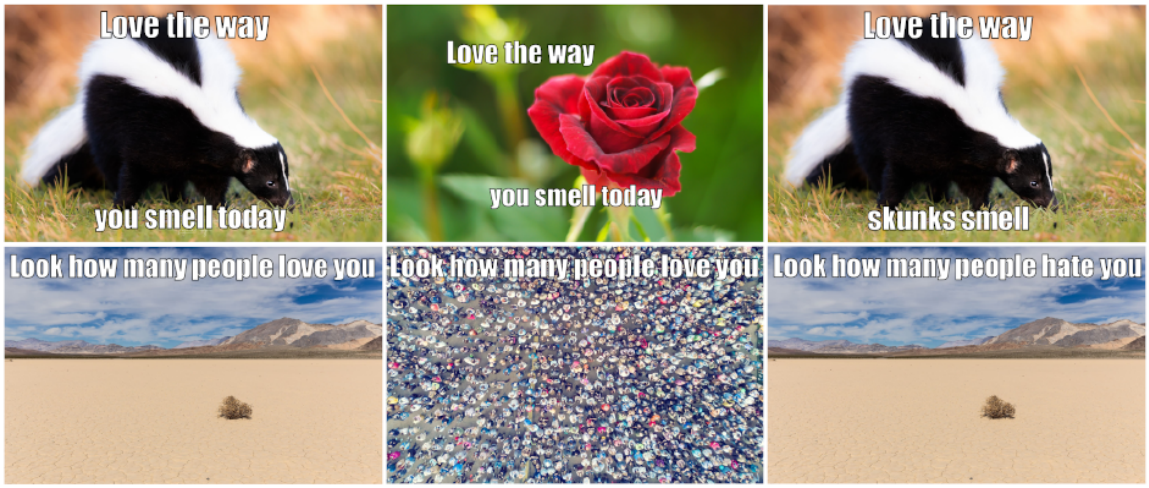

Figure 1: Creation of benign confounders as shown by Kiela et al. (2020). The memes on the left are mean memes (since, as put by Kiela et al. (2020, p. 2), "featuring real hate speech examples [...] would be distasteful"), while the memes in the middle are benign confounders where the image is swapped, and the memes on the right are benign confounders with the text swapped.

determining the label of the meme.

The dataset is split into a labeled train and validation set, and an unlabeled test set, out of which the train set is the only set that is not balanced (∼63% items are non-hateful, ∼37% are hateful). Due to the model created by Oriol et al. (2019) expecting a labeled test set, the original train and validation sets are concatenated and split into new train (containing 70% of the data), validation and test (each containing 15% of the data) sets. These new sets are then made available in a balanced version, containing 6600 items, and the imbalanced version containing 9000 items.

2.3 Defining hate

In order to examine hate speech, it must be defined. Kiela et al. (2020, p. 3) define hate speech in the following way:

> 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. We define attack as violent or dehumanizing (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.

The definition is comprehensive, allowing attacks for many different reasons to be included. For the purpose of this report, the definition supplied by Kiela et al. (2020) is adopted.

2.4 Method

The original model, with hyperparameters as defined in Oriol et al. (2019), shown in Table 1, is first trained on the imbalanced data (version 1), followed by the balanced data (version 1b). As these settings are highly time-consuming, the hyperparameters are adjusted to address this: the number of epochs is decreased, whereas the batch size is increased. This version of the model is trained and tested on both the imbalanced and the balanced data (models 2 and 2b, respectively).

The hyperparameters are then adjusted once more, decreasing the size of the hidden layers in the multi-layer perceptron and the number of epochs. After training and testing this version of the model on both imbalanced and balanced data (versions 3 and 3b, respectively), an image-only and a text-only version of this model are tested on the balanced data (models 3b10 and 3b01, respectively).

3 Results

Using the imbalanced data to train and test the original model by Oriol et al. (2019), the test accuracy is 70.1% (model version 1), compared to the validation accuracy of 83.3% reported by Oriol et al. (2019) when using their data. Conversely, the test accuracy for the best performing model tested by Kiela et al. (2020), Visual BERT COCO, achieves an accuracy of ∼64%. When using the balanced

Version	Hidden size	Epochs	Batch size	Balanced	Image	Text	Val. acc.	Test acc.
1	50	100	25	No	Yes	Yes	70.4	70.1
1b	50	100	25	Yes	Yes	Yes	65.6	62.4
2	50	50	50	No	Yes	Yes	69.5	67.6
2b	50	50	50	Yes	Yes	Yes	65.3	62.3
3	20	40	50	No	Yes	Yes	69.4	68.6
3b	20	40	50	Yes	Yes	Yes	65.1	61.9
3b10	20	40	50	Yes	Yes	No	57.7	59.5
3b01	20	40	50	Yes	No	Yes	65.9	65.4

Table 1: Settings and results for the tested model variations.

data, the accuracy decreases (compare model versions 1 and 1b; 2 and 2; 3 and 3b in Table 1).

By increasing the batch size and decreasing the number of epochs, the duration of the model is reduced from approximately 12 hours to just over 3 hours (excluding testing), not affecting the accuracy of the model significantly (compare model versions 1 and 2, or 1b and 2b in Table 1).

Comparing the model versions that were trained and tested on balanced with the model versions trained and tested on imbalanced data in Table 1, the use of balanced data clearly has some effect. However, the effect of using balanced data is rather small, considering that the imbalanced model versions would achieve an accuracy of 63% by simply classifying all items as non-hateful, due to the class imbalance.

The image-only and text-only versions of the model (versions 3b10 and 3b01, respectively, in Table 1) resulted in some interesting outcomes. While the image-only version has a lower accuracy than the multimodal version, the text-only version outperforms its multimodal counterpart.

Training loss and validation accuracy and loss for model versions 3, 3b, 3b10 and 3b01 is visualized in Figure 2 and Figure 3 in the appendix. With regards to the training loss, we can see that all four model variations seem to learn at various rates and are decreasing until the point when training stops, indicating that further training could be beneficial. Examining the validation loss of model 3b10 (red), the image-only model, it does not seem to be learning particularly much at all, and is the only model version that achieves a higher test accuracy than validation accuracy (see Table 1).

4 Discussion

Reporting the performance of a model without providing the necessary tools to replicate the model and its results requires a lot of trust from the reader. Oriol et al. (2019) supply the queries and the downloading tool used to collect the hateful memes, but not the actual memes, which is insufficient. Using the same downloading tool and the same queries will not produce the same results, making it impossible to exactly replicate the results produced by Oriol et al. (2019). Furthermore, the chosen queries are not discussed by Oriol et al. (2019), implying that there is no issue with this method. This is a rather naive approach, as a simple search on Google Images shows that there are memes that do not fall under the category of hateful memes, despite the search query being *racist meme*. Disproving this claim is just as impossible as reproducing the results by Oriol et al. (2019), since I do not supply the results from my search query, which further showcases the problem of not providing your data.

The Hateful memes dataset by Kiela et al. (2020) is collected in a less problematic way, but also has its issues. It is important to note and to problematize the fact that the labelling of the dataset is done manually by a small number of people. Having only a few annotators will not be a good representation of neither the world, nor the people who are affected by hate speech (women and all people from the BIPOC and LGBTQ+ communities). Fernquist et al. (2019, p. 2) notes that "what is considered hate by one person might not be considered hate by another", highlighting the issue of having only a handful of annotators. As we know that such a small group of annotators is not representative, it is important to realize the problem with giving them the power to decide what is considered hate speech and what is not. Although it is not clear who should be entrusted with this, it needs to be considered and discussed when creating a dataset on such a sensitive and important topic as hate speech.

I argue that noting both the representativeness of

a group of annotators and how the creators of the annotation task are working for diversity should be standard procedure, as the level of representativeness has an effect on the confidence we can have in their work as a group. We should also be careful about giving the same authority to a white person as we give a person from the BIPOC community regarding deciding whether or not something is racist, as the opinions of people from the BIPOC community should always be prioritized in these matters. The same goes for sexism, antisemitism and attacks on people from the LGBTQ+ community; the opinions of women, Jewish people, and people from the LGBTQ+ community should be given priority. A possible solution is giving people from marginalized groups heavier weighing votes in matters that pertain to their group.

In contrast to the results presented by Oriol et al. (2019), the text-only version of the model (3b01 in Table 1) performs better than both the image-only version and the multimodal version. This suggests that the Facebook Hateful Memes dataset does not require a multimodal model, proving one of the main goals of the dataset would be unsuccessful. However, to verify this claim, further, more extensive investigations would have to be conducted. As these results are discordant with those presented by Oriol et al. (2019), it is clear that the model does not exclusively rely on the image representation, but is flexible enough to give more weight to the modality that provides more insight, depending on the data.

Predictably, using balanced data causes the accuracy to decrease. However, the decrease in accuracy is not proportional to the decrease of one of the labels, indicating that the model does not exclusively depend on the ratio of the labels when classifying the memes.

As mentioned in section 3, the most successful model version, version 1 in 1, with a test accuracy of 70.1%, took just over 10 hours to train, while the second best performing version, version 3, with a test accuracy of 68.6%, took just over three hours. Comparing the balanced model versions, versions 1b and 3b, which took just shy of 8 and 2 hours respectively, the difference in accuracy is even smaller. The slightly higher accuracy is not, in my opinion, enough to warrant the increase in execution time. The execution time is reduced further for the text-only version, taking approximately 70 minutes.

The script by Oriol et al. (2019) writes a file containing image names of the 20 memes with the lowest squared error, and the 20 memes with the highest squared error. Examining these memes for a number of the model versions (some did not produce this file) shows that it seems easier for the model to predict memes containing hate speech as those that produced the lowest error are all of label 1, hate speech. Reviewing the images of these memes did not shed any light as to why this is the case, however, briefly examining the meme text reveals that *Muslim/-s*, *Islam* and *religion* are among the most common content words.

5 Conclusions and future work

The main goals of this study, mentioned in section 1, are 1) to examine whether or not the results reported by Oriol et al. (2019) can be reproduced with another dataset, and 2) to investigate whether or not the dataset created by Kiela et al. (2020) reaches its aim in being multimodal.

As for point 1), the results could not be fully reproduced with another dataset. Due to the original dataset not being available, comparisons between this and the Hateful memes dataset (Kiela et al., 2020) could not be made, leading to difficulties in improving the model. Other issues regarding the dataset used by Oriol et al. (2019), including trustworthiness and naivety, were also discussed. However, the model did outperform the baseline model reported by Kiela et al. (2020).

Regarding point 2), it was established that the dataset is not multimodal – the model can seemingly rely on the textual representation and achieve a slightly higher accuracy than when utilizing both modalities.

Further and more extensive exploration of the memes that garnered the lowest and highest squared error could potentially reveal the reasons as to why the model seems to depend on the textual representation to a higher degree. Testing more different model variations on the dataset could also contribute to this.

An interesting task would be to examine how similar the labels created by a larger group of annotators would be to the already available labels of the Hateful Memes dataset, given the same instructions. Additionally, giving another group of annotators the task of labelling the data with a continuous scale, instead of binary labels, could also provide some valuable insight. These two tasks

could either give more or less credibility to the dataset, depending on the outcomes of the second and third annotations.

References

Johan Fernquist, Oskar Lindholm, Lisa Kaati, and Nazar Akrami. 2019. A study on the feasibility to detect hate speech in swedish. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*, pages 4724–4729. IEEE.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

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.

Benet Oriol, Cristian Canton-Ferrer, and Xavier Giró-i-Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *CoRR*, abs/1910.02334.

R. R. Pranesh and Ambesh Shekhar. 2020. Memesem:a multi-modal framework for sentimental analysis of meme via transfer learning.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Appendix A Training loss

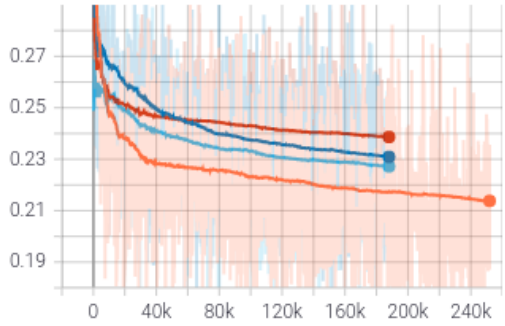

Figure 2: Smoothed training loss (smoothing constant = 0.999) for model variants 3 (orange), 3b (dark blue), 3b10 (red) and 3b01 (light blue). The x-axis represents number of batches.

Appendix B Validation accuracy and loss

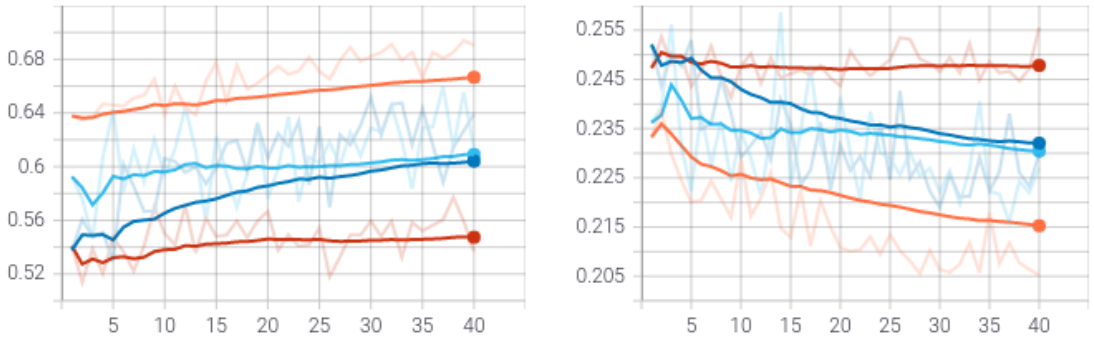

Figure 3: Validation accuracy and loss (smoothing constant = 0.999) for model variants 3 (orange), 3b (dark blue), 3b10 (red) and 3b01 (light blue). The x-axis represents number of epochs.