Detecting Hateful MEMEs against women and analyzing their popularity.

DEVAL SRIVASTAVA, Virginia Tech, USA

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:

Deval Srivastava. 2018. Detecting Hateful MEMEs against women and analyzing their popularity.. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 MOTIVATION

Demographic studies [1] tell that in the USA, 78% of women use at least one social media site compared to 66% of men. On sites like Pinterest, women account for 75% of the users [3] and on Facebook 77% of users are women, furthermore, women are also more likely to use Instagram [1]. From these figures, one can ascertain that social media has given new freedom of expression to women. However, alongside the growth of women users, online hate and misogyny have also grown [4]. According to a 2017 report from Amnesty International, 23% of women from eight countries have experienced online abuse or harassment at least once, and 41% of these said that on at least one occasion, these online experiences made them feel that their physical safety was threatened. In a different study at least 41% of female users expressed having experienced some form of misogyny and harassment [2]. This may have been through text, image, or another form.

Recently, "memes" have become a popular form of expression on social media amongst young adults. They are images that are superimposed with text. The image and text together are used to convey a joke. However, memes have become more than just ordinary jokes. Often the creators would embed hateful ideas and offensive content against women masquerading as humor.

Reddit is a popular social media platform, especially among the youth. A lot of the posts on Reddit are memes and studies [**?**] have shown that a vast majority of memes on the internet may have originally spread from Reddit. It also offers dedicated online spaces (subreddits) where people can discuss content. Many of these subreddits share memes that contain hate against women. Some of these subreddits exist purely to share hateful content sometimes under the guise of 'dark' memes or other times indiscreetly. There may also be a case that misogynistic memes may also be shared in mainstream harmless subreddits, where they may gain popularity. These memes create a poor environment for women and normalize misogyny.

In this work we are gonna design methods to detect such misogynistic memes and and analyze characterstics of the response they garner on Reddit. I plan to study the popularity of misogynistic memes and more specifically find the

Manuscript submitted to ACM

engagement of these memes on Reddit, find if these memes get moderated, and try to answer the question of what kind of subReddits offer lucrative spaces for these memes.

2 LITERATURE REVIEW

Detection and study of misogynistic memes are encouraged through multiple fields, at its core, it's a form of hate speech detection. Hate speech is defined by the Cambridge Dictionary as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation". Hate-speech detection has become a well-studied discipline, Schmidt et all [15] performed a survey of hate speech detection and they find that fairly competent systems have been developed. As discussed misogyny is rampant on social media hence detecting it has been the main focus for some time. One of the earlier works in that space [11] created a dataset using offensive language used for women as keywords, later on, Anzovino et all [5] released a corpus of labeled misogynist tweets. From these papers, it can be noted that the problems of hate speech and misogyny detection have been modeled as text classification problems. However, A major portion of the internet is visual and images are ubiquitous. Detecting hate speech across images and text can be considered Multimodal hate speech detection.

In the space of multimodal hate-speech detection, Haoti et al [17] collected Instagram posts and comments and had crowd workers annotate the data as being a aform of cyberbullying or not. A bit later yang et [16] al provide preliminary work on detecting multimodal hate speech they collect social media images with text and build a fusion model to classify them as hate speech. This work takes motivation from these and similar studies to build a multimodal classifier. Memes are a focus of this work and there have been a few studies that design models for multimodal memes, Dimitrov et al [9] study memes to detect propaganda, they motivate the use of multimodal transformers. Furthermore, Kiela et al [13] released a multimodal meme dataset for identifying hateful memes. Analysis of memes on Reddit has also been under some research I take inspiration on method design from kate et al [6] who try to predict the popularity of memes on Reddit.

This project takes motivation from all the past and current research in the space of hate speech and misogyny detection and adds to the same. Also addressing the gap in literature we can observe that most of the work is unimodal and that study of misogyny in multimodal data and memes has been less studied.

3 RESEARCH QUESTIONS

In this project, we address the following research questions:

3.1 RQ-1: How can we identify a misogynistic meme?

(1) **RQ-1.A**: How do existing hateful meme detection datasets and models fare on misogyny detection?
(2) **RQ-1.B**: How should the information within memes be represented for best classification results?

Motivated by the research of trying to detect propaganda and hate in memes. In this work we take learnings from past works and address the gap in trying to detect misogyny in multimodal memes. Through the first sub-research question we want to know how effective would be existing methods and datasets on this new task. Multimodal research has grown tremendously over the past few years, through the second sub research question we want to analyze effective would be a multimodal architecture compared to a language-only architecture.

3.2 RQ-2: How do subreddits and their users respond to misogynistic memes?

(1) **RQ-2.A**: What kind of user engagement (represented using upvotes, comments) do these memes capture compared to standard memes?
(2) **RQ-2.B**: Is there a moderation (warnings, takedowns) response to the misogynistic memes?
(3) **RQ-3.C**:What kind of subreddits provide the most habitable environment for misogynistic memes in terms of user engagement and moderation? ?

When a misogynistic meme has been posted on a subreddit,its users and adminstration can respond to it in multiple ways. We quantify these responses and try to find how they compare to responses to non-misogynistic memes. A user can respond to a meme by either up/down-voting or by commenting on it. We express this interaction as user engagement with the meme. A moderator/adminstration can respond by either warnings or takedowns.

The first sub research question is motivated by Kate et al [6] 's work where they define popularity metrics and the second sub question is motivated by the birman et al[7] study of moderation.

Through the last question we want to explore what kind of subreddits would have the most favourable response to misogynistic content, a favourable response can be defined as a high engagement metric and low moderation.

4 METHODOLOGY

In this section we will discuss the methodology followed and the steps taken to answer the two research questions, in the first section we will discuss the modelling methods revolving around misogyny detection and in the second section, the focus is on the analysis of the response surrounding misogynistic memes.

4.1 Methods for RQ1

To answer this research question we will explore a number of methods to detect misogynistic memes and evaluate their performance. Before we discuss the methods, we can define the dataset that is going to be used for training or evaluation of these methods. In this work, I focus on the misogynistic meme dataset[10] released during SemEval 2022 as a challenge. It contains 10,000 memes for training along with their text transcriptions. The class distribution for this dataset is evenly split between misogynistic and non-misogynistic classes. To answer the two sub research questions I take the following steps.

(1) **RQ1.A** : For this RQ we evaluate the effectiveness of existing models on the task of misogyny detection in memes. Motivated by the research in the space of meme classification. I consider the datasets and models trained on hateful meme detection. Specifically, the dataset [13] released by facebook as a challenge.I directly a utilize a baseline released by facebook on this task using MMBT (Multimodal BiTransformer)[12] and evaluate it on the test split of misogyny detection.
(2) **RQ1.B** : For this RQ I train two model architectures to classify misogynistic memes. The first model represents the meme using only its text transcription. Effectively turning the problem into a text sequence classification problem. BERT [8] encoder is used to obtain contextual representations and then a linear layer is added to classify the input. The second model uses a multi-modal representation of the MEME by leveraging both the image and text transcription. For this model, I use CLIP[14] to obtain the encodings for Text and Image content. I fuse these encodings using linear layers and classify them using softmax. For these models, I evaluate and train them on the dataset [10].

4.2 Methods for RQ2

For this RQ, we want to collect, detect and analyze characteristics of misogynistic memes on Reddit. Before diving into the analysis section, we can discuss the data collection and processing part. Since each subreddit has a different high level narrative and will respond differntly to memes, collection and analysis is segregated by subreddits. Reddit has upwards of thousands of subdreddits, but for the purpose of this research we limit analysis to 11 subreddits. These subreddits have been choosen by searching for the query "meme" and filtering results by communities.

The data collection steps are defined below:

(1) For the collection of memes, we consider a two year period. Now for each subreddit, collecting and processing all memes in this period would be very time expensive. To work our way around this, we collect memes in two fashions.
(2) Firstly we collect all memes posted in the two year period that match the query "women" or "woman". This procedure relies on the assumption that a meme matching the query would be relevant to women in some way.
(3) Secondly, to get a general set of memes. We collect a small amount of memes from each subreddit from the two year period. This corpus woudnt neccessarily be relevant to women but would be representative of standard memes posted to the subreddit.
(4) Considering technical details, we utilize PushShift API to obtain posts, relevant metrics and meme images within a time period and PRAW is used for collection of metrics like upvotes and subreddit subscriber counts.
(5) To provide the model consistency in the inputs it has observed during training we add text transcripts to the memes. We utilize Tesseract OCR Tool and custom image processing to generate text transcriptions for the scraped memes.

To perform analysis we define following two metrics:

(1) **User engagement metric**: Motivated by Kate et al[6] 's work of developing the popularity metric as,

$$popularity = \frac{Number\ Of\ Upvotes}{Subscriber\ count} \tag{1}$$

We define similar metric sfor user engagement, based on the upvote count, comment count normalized by the subscriber counts. we call them the upvote engagement metric, and the comment engagement metric

$$upvote\ engagement = \frac{Number\ Of\ Upvotes}{Subscriber\ count} \tag{2}$$

$$comment\ engagement = \frac{Number\ Of\ comments}{Subscriber\ count} \tag{3}$$

(2) **Moderation Response metric**: Motivated by Birman et al[7] we define a moderation response in a given subreddit as the number of posts that have been taken down either by Reddit adminstration or the moderators.

$$moderation\ response = \frac{Number\ of\ takendown\ posts}{All\ posts} \tag{4}$$

5 FINDINGS

In this section we can discuss the results obtained after following the methodology defined above and see how our findings answer the research questions we have posed.

5.1 Results for RQ1

5.2 RQ1.A: Evaluation of A MMBT model trained on Facebook hateful meme detection dataset challenge

Accuracy : 62%

F1 Score : 0.57

Since this is a binary classification task, random accuracy would be at 50%. It can be said that any model that exceeds threshold understands the data to some extent. However 62% accuracy is not acceptable accuracy, it would be considered to be inaccurate model in academia and industry. It should be noted that this model hasn't been trained on the task of misogyny detection, but still manages to reach some performance, showing overlap between the tasks of hateful meme detection and misogynistic meme detection.

RQ1.B: Evaluation of Language only model and multimodal architecture:

Performance of Language only Model:

Accuracy: 79%

F1 Score: 0.78

Performance of Multimodal architecture with KFold Cross validation k=5:

Avg Accuracy: 87.4%

Avg F1 score: 0.87

From the performance of these models we can reach a conclusion on what is the best way of representing information within a meme for modelling. Firstly considering the language only model. Its able to outperform the MMBT model baseline, hence it can be established that task specific training does help here. Secondly looking at the results of the multimodal architecture, its able to get the highest performance within our set of models. Demonstrating that image data within a meme is not redundant to text and adds more information which helps in classification. To conclude, multimodal encoding of image and text would be the best way to represent a meme.

5.3 Results for RQ2

We can first look at the data collection results, I collected data from the following subreddits in 1.From the data we can observe that this set of subreddits includes some of the most popular subreddits and some more nicher smaller subreddits, this distribution will provide us a balanced result.

5.4 RQ2.A: Expressing user engagement of a meme using upvotes,comments and awards

User engagement can be expressed using the number of upvotes a meme gets normalized by the subscribers in that subreddit, we perform this analysis for all selected subreddits and report the results in the plot 1.

The higher the bar in this plot, higher is the engagement expressed using upvotes. We observe that mainstream and bigger subreddits generally have poor user engagement for misogynistic memes compared to the smaller subreddits. We can also express user engagement as a ratio of engagement metrics of misogynistic memes compared to non misogynistic memes. This result is plotted 2.

Table 1. Subreddit data

Subreddit	Subscriber Count	Misogynistic Memes	Non-Misogynistic Memes
memes	18.7M	1199	2400
meme	2M	200	2400
dankememe	5.5M	X	1200
Memes_Of_The_Dank	825K	72	2400
terriblefacebookmemes	1.8M	175	1300
ComedyCemetery	980K	85	1299
Okbuddyretard	975K	250	1300
dankmeme	62k	354	2400
MemesIRL	97K	5	1000
Grimdank	257k	15	2400
Darkmemers	78k	10	2400

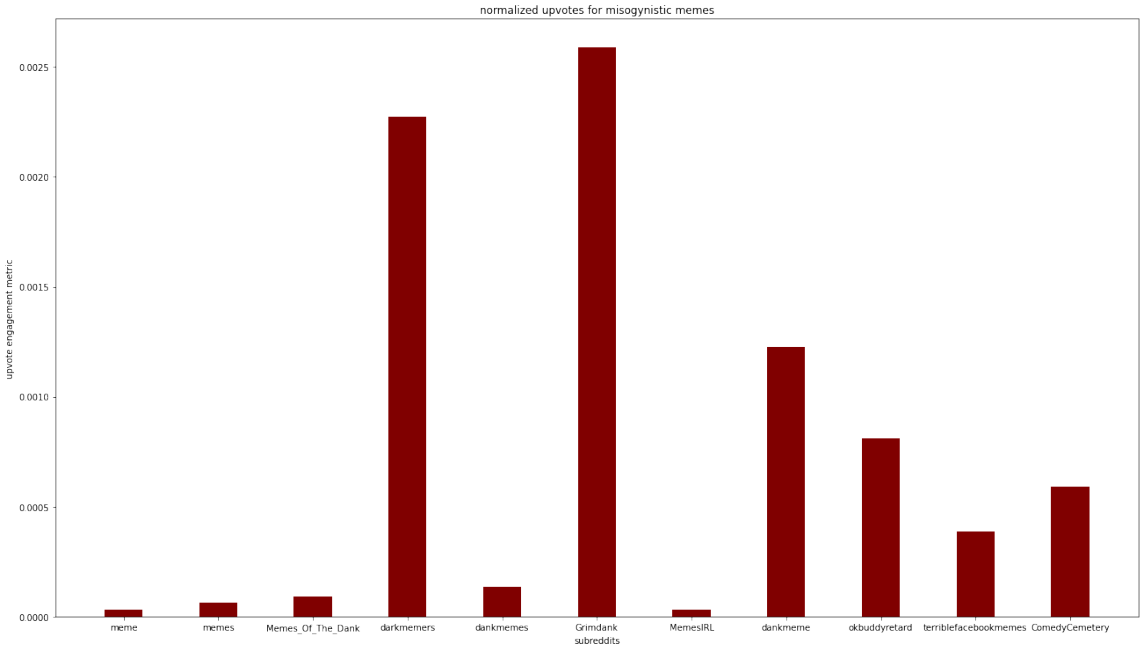

Fig. 1. Upvote engagement for all subredits for misogynistic memes

In this plot the higher the bar, that many times more engagement did non-misogynistic memes receive compared to misogynistic ones. Again in this plot we observe that for bigger subreddits there is a huge gap in user engagement compared to much lesser gap for smaller subreddits.

User engagement can also be expressed using number of comments. For a comment engagement metric I report two plots, the first one showing the average user engagement metric for misogynistic memes in figure **??** and a second plot showing the ratio of how how much more comment engagement did the non-misogynistic memes get compared to misogynistic.

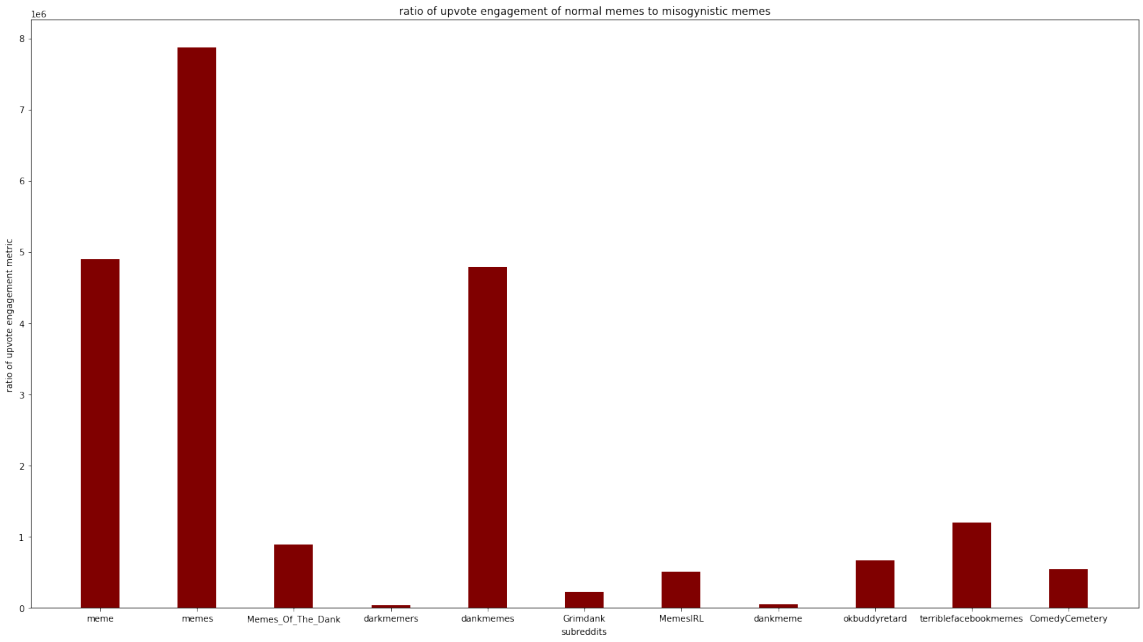

Fig. 2. ratio of upvote engagement of normal memes to misogynistic meme for each subreddit

For the comment engagement plots we also observe that misogynistic memes in smaller subreddits received higher comment engagement compared to bigger subreddits and looking at the ratio of comment engagement, we see a similar story.

5.5 Results for RQ2

For this RQ we measure the moderation response by the subreddit when a misogynistic meme has been posted.The moderation response can be measure d as the percentage of posts that have been moderated. We plot this information for each subreddit in figure 5.

From this plot we can observe that smaller subreddits have significantly lesser moderation compared to the larger subreddits. For some of the subreddits like GrimDank, memesIRL and darkmemers. There is no absolutely no moderation on misogynistic memes. On bigger subreddits moderation exsists however that is also only around 15%-20%. Some of these bigger subreddits are mainstream so in an ideal case the moderation should be even higher. This result highlights how Reddit has a moderation problem.

5.6 Results for RQ3

The results in RQ1 and RQ2 suggest that subreddit with lesser subscribers woul offer the most lucrative spaces for misogynistic memes. Due to the lesser moderation and higher engagement. However that doesnt completely answer this question. In future work, we can study niche subreddits and the discourse around misogynistic memes to answer this question.

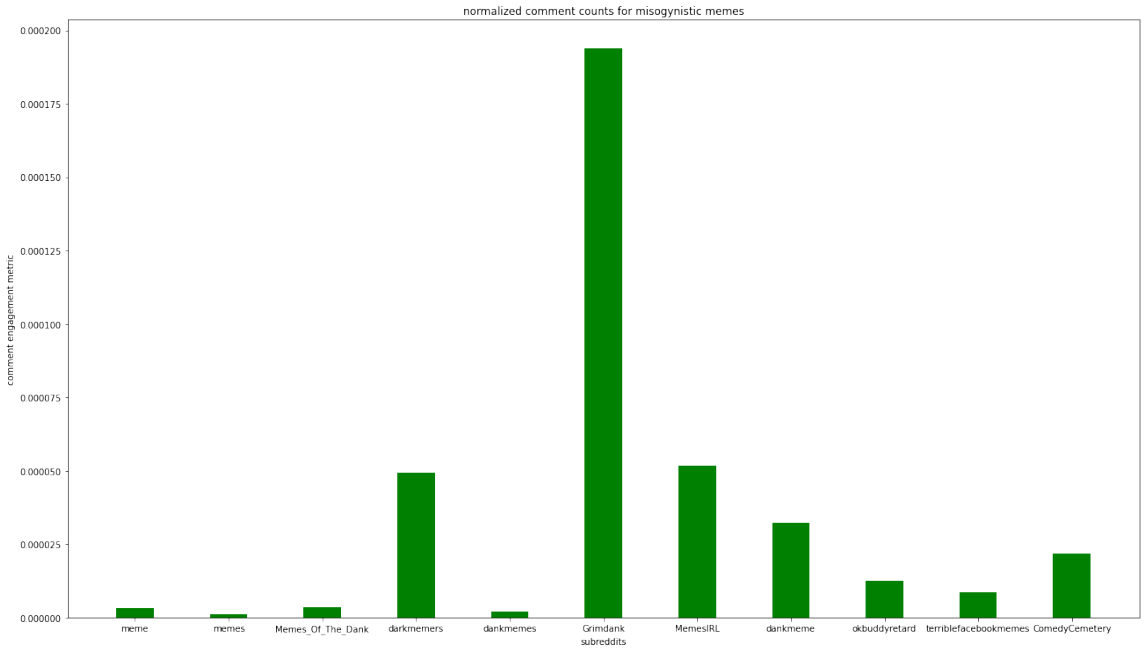

Fig. 3. comment engagement for each subreddit for misogynistic memes

??

6 IMPLICATIONS & DISCUSSION

In this section we discuss the implications of this work on the field and how the contributions shape society and academia.

6.1 Implications of RQ1

In RQ1 we explore models for the task of detecting misogynistic memes. Firstly, from the literature, we had observe that multimodal meme detection as a whole has been a relatively less studied field and is still growing. Through this work, this field gains more attention and may also possibly motivate future works on similar topics. We also show that misogyny detection is a problem of its own.This result also motivate the development of better models for misogynistic meme detection, that treat the problem independently. RQ1 results show that multimodal architectures allow for greater performance gains on meme classification. We observe that a simple multimodal architecture that just uses CLIP contextual encodings can outperform a language-only model. These results motivate future work in the domain of multimodal architectures that are capable of really solving misogynistic meme detection.

The development of a model that is able to detect misogynistic memes allows for auto-moderation methods to be built. Subreddits on Reddit can receive upwards of 5000 memes a day. This volume of data cannot be feasibly analyzed and filtered by even a big group of individuals working full time, let alone by few people who are moderators as a hobby on Reddit. If a high performing model could be built that can classify misogynistic memes, such a model could be wrapped as an automated check that has to be cleared before a meme reaches the general public.

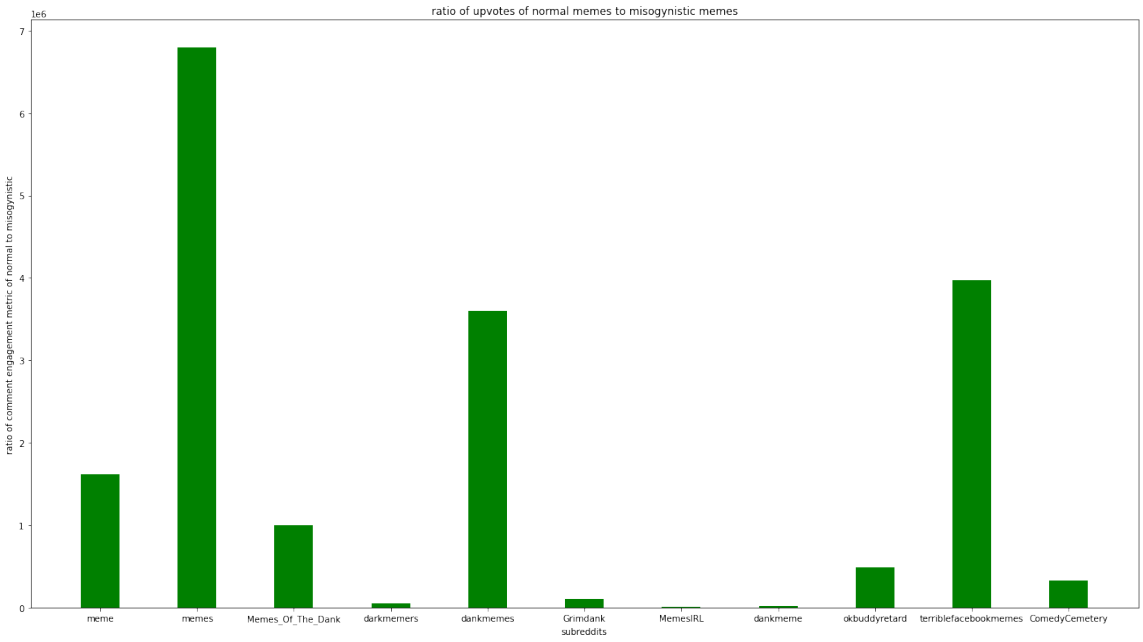

Fig. 4. ratio of comment engagement of normal memes to misogynistic meme for each subreddit

6.2 Implications of RQ2

In the results for RQ2, we find that smaller subreddits provide a more habitable environment for misogynistic memes. This is due to a lower moderation response and a similar if not higher user engagement metric compared to bigger subreddits. This result supports a popular notion on Reddit of how smaller subreddits are prone to believing in extreme ideas.

We highlight that Reddit has a moderation problem. In Smaller subreddits, there is significantly less moderation compared to bigger subreddits and even bigger subreddits only have moderation of about 20%. This makes you think that either the moderators are understaffed and are not able to meet the required output or they are they have a different idea of what they consider to be a violation. Multiple directions of future work can be motivated by this RQ, from further analysis of memes to analysis of moderation.

REFERENCES

[1] 2017. pew research. (2017). https://www.pewresearch.org/internet/fact-sheet/social-media/
[2] 2017. pew research state of online harrasment. (2017). https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
[3] 2017. statista resarch. (2017). https://www.statista.com/statistics/248168/gender-distribution-of-pinterest-users/
[4] 2017. ucsd misogyny. (2017). https://gehweb.ucsd.edu/social-media-sexist-online-gender-based-violence/
[5] Mary E. Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *NLDB*.
[6] 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 (Mar 2021). https://doi.org/10.1007/s41109-021-00358-7
[7] Iris Birman. 2018. Moderation in different communities on Reddit – A qualitative analysis study.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

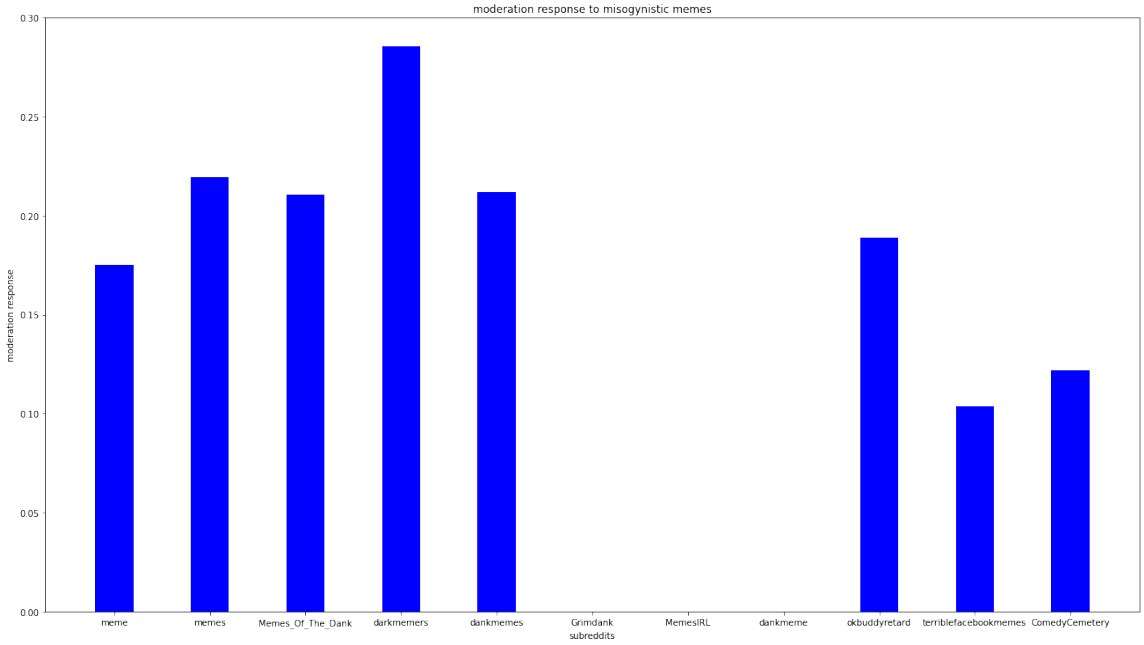

Fig. 5. Moderation Response for misogynistic memes.

[9] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting Propaganda Techniques in Memes. (2021). https://doi.org/10.48550/ARXIV.2109.08013

[10] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2021. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. https://doi.org/10.48550/ARXIV.2106.08409

[11] Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The Problem of Identifying Misogynist Language on Twitter (and Other Online Social Spaces). In *Proceedings of the 8th ACM Conference on Web Science* (Hannover, Germany) *(WebSci '16)*. Association for Computing Machinery, New York, NY, USA, 333–335. https://doi.org/10.1145/2908131.2908183

[12] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. https://doi.org/10.48550/ARXIV.1909.02950

[13] 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. https://doi.org/10.48550/ARXIV.2005.04790

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[15] 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.* Association for Computational Linguistics, Valencia, Spain, 1–10. https://doi.org/10.18653/v1/W17-1101

[16] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification. In *Proceedings of the Third Workshop on Abusive Language Online.* Association for Computational Linguistics, Florence, Italy, 11–18. https://doi.org/10.18653/v1/W19-3502

[17] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 3952–3958.