IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203185|

Data Analytics System for Offensive Memes Text Classification in Social Networks

Divyabharathi G[1], Harinikiruthika.R[2], Divyadharshini M[3], Manisha Kumari[4]

Assistant Professor, Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, Tamilnadu, India[1]

Department of Computer Science and Engineering, Mahendra Institute of Technology, Namakkal, Tamilnadu, India[2,3,4]

ABSTRACT: A meme is a culturally relevant, and brief form of media that raise a content over the internet. Now a days posting a meme is popular communication medium, due to its multimodal nature. Postings of hateful memes or fooling, cyberbullying are growing gradually. Meme takes a major part in forming people's trust and perspective. Meme can be quickly post by anybody, and its integrity stands compromised. Hateful and aggressive matter detection have been largely traversed in a form such as text or image. And Memes complicate the task, because some meme can have a good caption and normal pictures, but if combined in some way, they can become offensive. So, it is required to fuse both modality to identify whether a given meme is hateful or not. So here for text classification we found the sequential model like Bi-LSTM and for image we will go with CNN. Late fusion technique is used to combine the image and text mode with EX-OR method to investigate its effectiveness.

KEYWORDS: Hateful meme, Deep Learning, Natural Language Processing, Late fusion, Image Processing, Bi-LSTM, EX- OR.

I. INTRODUCTION

Now a days social media be it Facebook, Instagram, Twitter or any other social media has gained a lot of enticement since its beginning. Day by day the content generation is growing rapidly. Almost half population of society is present on social media. Most of the present day memes have more impact on people as they are structurally apart from longer form of media. Memes become successful because they are Familiar, short, brief, culturally relevant and quickly understood by the target audience. Memes can be in many forms like

- Image Dominant: where textual data is less.
- Text Dominant: where textual data is huge
- Text and Image Dominant: where textual and image data are same.

Recently the most central usage of memes to be of images combined with text. Irrespective of the type of meme, the meme may get modified, rehashed or recreated while inter- acting through social media network. In this age memes are decorated fancy pictures that are put forward to be diverting, often as a way to freely disrespect individual control. Additional memes can be spoken recordings, and few memes have a more logical subject. So due to the massive use of social networks, negative impacts need to be automatically limited. This project serves to propose a technique which uses text mode and image mode processing to classify the memes as hateful or non-hateful and stop spreading hate across internet. This work aims to conduct thorough approaches in computer vision, deep learning and natural language processing.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203185|

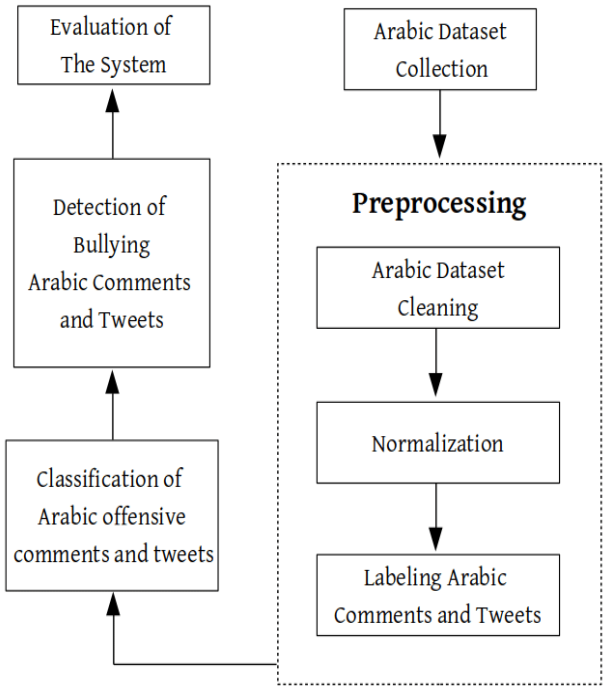

Fig 1: MEMES TEXT CLASSIFICATION

The author created an attention model based on semantic similarity to overcome the limitations of BERT .Developed a BERT-based model that does a better job than most models in the MovieQA problem. In order to solve MCQ, the author first extracts sentences from large text, which makes it easier to answer MCQ questions [3].To differentiate between profanity, hate speech and other texts author has applied Some lexical features and applied a Support Vector Machine classifier to set up a measure. A character 4-gram model works well for this task [4]. Extraction and preprocessing of text, image and face encoding is done. Author performed four different groups of classifiers but LSTM model perform well than other text models and for vision DNN models show better advancement [5].Author propose a method to analyze memes based on sentiments of meme. Memes with emotions like happy, angery, or other kind of emotions are collected. Visual features are extracted MATLAB functions and textual features using OCR.J48 algorithm worked accurately on the memes [6].

This paper shows a various ways like concatenation bilinear transformation and gated summation to fuse text and photo signal with further improvement [7].Two architectures are evaluated for detection of objects in text for visual question answering. A late fusion architecture where text and image are encoded separately before fuse result. , And early fusion B2T2 model where visual features are placed on the same level as input word tokens. B2T2 is highly effective [8]. This paper presents a new challenging dataset for identification of hateful speech in multimodal memes. It is established by adding some meme example in such a way that unimodal classifier would fight to classify them. Also some baseline are provided for both models. But existing methods fail to reach at performance. So this pointing the challenge to community [9].

Related Work

The work done to identify objectionable content in images is covered in the related part. It also discusses the multimodality and meme analysis study that has been done. The CNN based approach has better performance in terms of recall when image features are considered For the purpose of locating objectionable information in a picture, nudity recognition based on CNN approaches has been proposed (Arentz and Olstad, 2004; Kakumanu et al., 2007; Tian et al., 2018). Convolutional neural networks (CNNs) have been proposed as a method to categorize images for children as appropriate or inappropriate in a number of publications. 2018 (Connie et al.). A research on offensive visuals and non-compliant logos was undertaken by Gandhi et al. in 2019. They developed an algorithm to recognize offensive content in non compliant and objectionable photos.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203185|

A photograph is deemed offensive if it contains nudity, sexually explicit content, weapons or other symbols of violence, or if it is racially insulting. By comparing the embeddings of the images, the authors were able to identify similar images and construct the dataset that they used. To determine the kind of item in the image, the classifier uses a previously trained object detector. Object detection is heavily used in this study. In our study, we depend on automatically generated features through a CNN that has already been trained to classify memes with comparatively fewer resources. A novel framework for categorizing pornographic web sites using images was put forth by Hu et al. in 2007. To categorize Websites into discrete images, the authors used a decision tree.

The algorithm fuses the output from the image classifier to identify inappropriate content in accordance with content representations. This work depends on CNN to find pornographic material on the webpage. Unlike our study, their attempt to identify the content is less cryptic and more explicit. He and colleagues (2016) suggested a meme extraction algorithm that uses data posted during occasions like the anti-vaccination movement to automatically extract textual features2 The extraction process is carried out by identifying isolated sentences and combining the mutation variation of each phrase associated with the meme. This research examines the peaks and points of confluence of memes. Drakett et al. (2018) used thematic analysis of 240 example memes to address the issue of online harassment of marginalized groups through the use of memes. This study examines memes from a psycho linguistic angle.

II. METHODS

Data Collection and Annotation Obtaining and Annotating Information We produced the dataset by manually categorizing the data into contentious and non-offensive groups. The reviewer must classify the shown image either as offensive or inoffensive depending on the image that supported it. The annotators were tasked to identify a specific image as being offensive or not offensive depending on the image that supported it.. Memes that target: (a) personal assault(b) homophobia may be deemed offensive. racial epithets(c)Attack on a minority group(d) non-offensive in any other situation. Most memes consist of an image and a caption. The reviewer needs to be conscious that the context and importance of the images in conveying their meaning are important points to note. Consequently, a picture might rarely have no meaning. The meme should be labeled offensive and granted the benefit of the doubt if there is any doubt as to whether it is being taken seriously.

When annotating the data, Annotators should consider the population's total exposure to the meme's content. Only six male annotators provided their assistance once pre-processing and annotation guidelines were established. The gender distribution of the annotation distribution was balanced in order to avoid gender bias. Eight annotators (six men and two women) eventually consented to take part in the annotation effort. Two steps were taken to finish the annotating process. Every one of the eight annotators received 50 memes in the first round. The vast majority vote has been used as the gold standard although there was no support for truth provided, and the Fleiss' kappa was determined for this majority vote. Initially, "fair agreement" between the annotators was indicated by the spread of kappa's highest and lowest values being between 0.2 and 0.3. The challenges that annotators faced when recording the data included the following: varying annotators had varying interpretations of satirical ideas. Most sarcastic memes were annotated by people who differed from each other.

Is one illustration of such a picture. The annotators were merely classifying the images as offensive if they had been offended by them due to their lack of knowledge of US politics. We updated the annotation standards and added numbers V and VI to lists that were previously received in an effort to solve the difficulties brought up by the annotators. Each annotator received 50 brand-new memes when the annotation criteria were updated. When Kappa was determined, it revealed average agreement between the annotators, similar to the previous set of annotations. (0.4 and 0.5). We forwarded every joke to the annotators after we got a good lot of consensus, A reported audience response to the content can serve as the gold standard for gauging sentiments, according to psychology (Gilbert, 2006), and this response may be interpreted as the actual joke. Since the dataset's memes truly undermine the sentiment analysis results, data annotation is a difficult and emotionally taxing task for the annotators.

III. RESULT ANALYSIS

We compared the results of our model with all kinds of unimodal and multimodal models on the hateful memes dataset. The activation function in our model is selected as ReLU, and the threshold value to calculate the hateful/not-hatful class center is set as 0.50.5. The results of compared models on the dataset were from. For the unimodal models, it can be found that their performance is generally less satisfactory. In addition, the unimodal text model outperformed the unimodal image model, reflecting the fact that the text features may contain more information. For the multimodal

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203185|

models, they outperformed the unimodal models. We also found that the fusion method affects their performance, while models using early fusion methods outperformed those using later fusion methods. For the multimodal pretrained process, there was little difference between the multimodal pretrained model and the unimodal pretrained model.

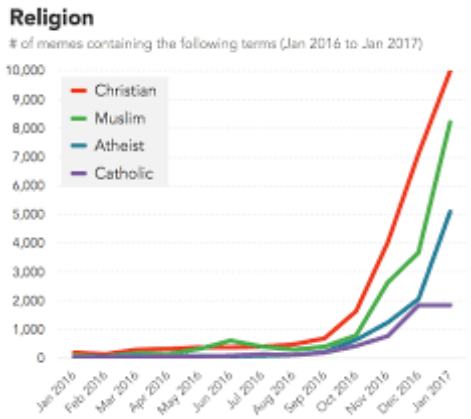

Fig 3: Result analysis

In contrast to the models mentioned above, our model used a late fusion method and two unimodal pre-training models. Although the late fusion method generally performed worse than the early fusion method, our model outperformed those early fusion models. Thanks to the additional auxiliary learning, which validated the idea that adding multi-task learning to hateful meme detection can improve the accuracy of the task. As the hateful memes data are complicated with two modalities, we designed multi-task learning to make the statistical inference. When we optimize the model, the extracted information may be fuzzy if we pay too much attention to the multimodal part. However, if we pay too much attention to the unimodal part, the extracted information may be much unilateral and weaken our primary task. In addition, the gradient magnitudes of the backpropagation of several tasks' losses may differ. When backpropagating to the shared bottom part, the task with a small gradient magnitude has less weight to update the model parameters, making the shared bottom not learn enough for that task. Of course, we can simply introduce static weights to balance the gradients for different tasks. However, this does not work well. If we assigned a fixed weight for a task with a large gradient magnitude at the beginning of training, this small weight would keep limiting this task by the end of the training, making this task not learned enough and enhancing the generalization errors

IV. CONCLUSION

We proposed to investigate the important problem of memes classification system using computer vision and NLP techniques. Suggests a method that could be beneficial to classify memes with fix visual and textual features. Late Fusion will classify content for both text and image before trying to fuse the results. Eventually, further research and work to identify hateful meme is in progress and will give a more refined classification scheme for meme. In the future, we plan to extend this work to other multimodal feature extraction methods to improve training on specific data sets. In addition, social media trends and patterns are changing rapidly, so it is necessary to capture memes in real time with respect to a particular domain so as to find the influential entities. This work can be extended to collect this data in real time and train deep learning models to identify hateful memes.

REFERENCES

[1] Bharathi Raja Chakravarthi,Shardul Suryawanshi Mihael Arcan, Paul Buitelaar, "Multimodal Meme Dataset (MultiOFF) for Identifying Of- fensive Content in Image and Text,"Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020.

[2] William Yang Wang, Miaomiao Wen," I Can Has Cheezburger? A nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions" In: The 2015 Annual Conference of the North American Chapter of the ACL, pp. 355– 365 (2015).

[3] Omar Mossad, Amgad Ahmed, et al., "FAT ALBERT: Finding Answers in Large Texts using Semantic Similarity Attention Layer based on BERT."

[4] Shervin Malmasi, MarcosZampieri "Detecting Hate Speech in Social- Media"

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203185|

[5] M. Beskow,Sumeet Kumar, Kathleen M. Carley," The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning".
[6] E. S. Smitha , S. Sendhilkumar , and G. S. Mahalaksmi,"Meme Classi- fication Using Textual and Visual Features".
[7] Gargi Ghosh Reshef Shilon, Fan Yang, Xiaochang Peng,Hao Ma ,Eider Moore, Goran Predovic,"Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification".
[8] Alberti, C., Ling, J., Collins, M., Reitter, D.: Fusion of detected objects in text for visual question answering. arXiv preprint arXiv:1908.05054 (2019).
[9] Goswami, V.,Kiela, D., Firooz, H., Mohan, A., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. arXiv preprint arXiv:2005.04790 (2020).
[10] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. arxiv:1708.01967, 2017.
[11] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. 2017.
[12] Sergey Smetanin. , EmoSense at SemEval-2019, Task 3: Bidirectional LSTM Network for Contextual Emotion Detection in Textual Con- versations. (International Workshop on Semantic Evaluation held in conjunction with NAACL-2019 in New Orleans, LA, USA.)
[13] C. C. Park and G. Kim. Expressing an image stream with a sequence of natural sentences. In Advances in neural information processing systems, pages 73–81, 2015.Paula Fortuna and Sergio Nunes. A survey on automatic detection of hate speech in text. 51(4), 2018.
[14] Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356 (2016

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details