Hateful Meme Detection for Social Media Applications

G.Darshan[1], K.Deepak[2], G.Suresh[3]

[1,2]B.E students, Department of Electronics and Communication Engineering , Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India - 638401.
[3]Assistant Professor, Department of Electronics and Communication Engineering , Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India - 638401.

---***---

Abstract - *"A direct or indirect meme effect on individuals dependent on characteristics, including ethnicity, race, religion, caste, sex, gender identity and disability or disease. Such meme are considered as violent or denying a group (comparing people to non-human things, e.g., animals) speech, explanations of inadequacy, and calls for prohibition or isolation. Taunting crime is also considered hate speech". In modern world, to make AI a more efficient tool for detecting the hateful speech and hateful images, first AI tool should understand the way of people delivering the content like posting the memes in social media. When a meme is viewed, text and images are not viewed independently by the humans as human's understand the meme only by combined meaning of the text and image. In AI, it is a complex process for combining both text and images for analysing the data for detecting the hateful memes.*

Key Words: **Tokenization, Inter-modality transform, Intra-modality transform, LanguageAndVisionConcat model, Hateful memes, Word tokenization, Subword tokenization**

1. INTRODUCTION

Natural language processing helps computer systems communicate with human beings in their own language and scales other language-related obligations. for example, NLP makes it viable for computer systems to examine text, listen speech, interpret it, measure sentiment and determine which elements are critical. These days's machines can examine extra language-primarily based information than people, without fatigue and in a steady, impartial manner. thinking about the outstanding amount of unstructured statistics that's generated every day, from clinical records to social media, automation can be critical to completely examine text and speech facts effectively. In this project, we are going to classify the content of meme as either hateful or non hateful using our proposed model which is created by examining the existing model and creating a model based on the existing model by overcoming some of the drawbacks of the existing system. The existing model we chose is the Language and Vision Concat model which is generally a multimodel algorithm which helps to perform complex tasks like the hateful meme detection.

2. LITRATURE SURVEY

In [1] paper, the author proposes a new challenge set for multimodal classification, focusing on detecting hate speech in multimodal memes. It is constructed such that unimodal models struggle and only multimodal models can succeed. We find that state-of-the-art methods perform poorly compared to humans (64.73% vs. 84.7% accuracy), illustrating the difficulty of the task and highlighting the challenge that this important problem poses to the community.

In [2] paper, the author states that Hateful Memes Challenge is a first-of-its-kind competition which focuses on detecting hate speech in multimodal memes and it proposes a new data set containing 10,000+ new examples of multimodal content. We utilize VisualBERT -- which meant to be the BERT of vision and language -- that was trained multimodally on images and captions and apply Ensemble Learning.

In [3] paper, It is widely shared that capturing relationships among multi-modality features would be helpful for representing and ultimately describing an image. An end-to-end formulation is adopted to train the whole model jointly. Experiments on the MS-COCO dataset show the effectiveness of our model, leading to improvements on all commonly used metrics on the "Karpathy" test split.

In [4] paper, Top-down visual attention mechanisms have been used extensively in image captioning and visual question answering (VQA) to enable deeper image understanding through fine-grained analysis and even multiple steps of reasoning. In the work, it is proposed that a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. This is the natural basis for attention to be considered. Within the approach, the bottom-up mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings.

3. DATASET DESCRIPTION

Hateful meme dataset consist of nearly 10000 memes which consist of multimodal content. Dataset is created in such a way that unimodal classifiers struggle to predict the outcome accurately. We have also designed the dataset in such a way that it overcomes the challenges of learning to avoid false positive. The data set also contains multimodal memes that are similar to hateful examples but are actually harmless. These examples, known as benign confounders, will help researchers address potential biases in classification systems and build systems that avoid false positives.

Fig -1: Hateful memes

4. DATA FUSION

Fusion is the process of collecting information from multiple source combining the features and collectively giving it as an single entity. The fusion process can be classified into two types namely Early fusion and Late fusion.

4.1. EARLY FUSION

Early fusion can defined as the fusion techique in which we combine all the input data into a single entity and then proceed for further processing. It helps the model to analyse the feature, like humans do.

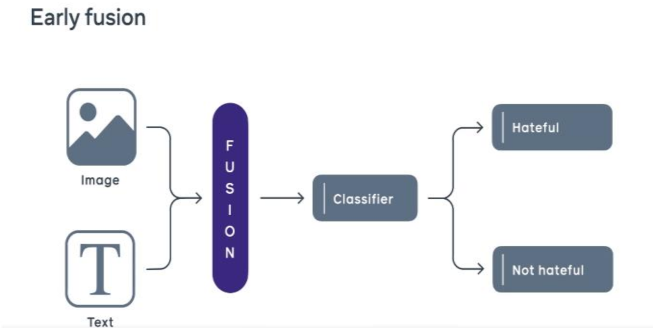


Fig -2: Early fusion

4.2. LATE FUSION

Late fusion can defined as the fusion techique in which we analyse the features seperately and then combine them into a ssingle entity. It contrasts with the early fusion. It is easier to build but is less effective. It makes the model so complex to understand.

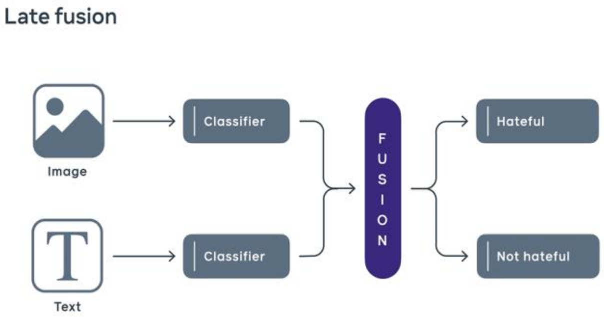


Fig -3: Late fusion

5. TOKENIZATION

Tokenization is nothing but a very common task performed in Natural Language Processing. It is one of the common steps in both traditional NLP's and also in our latest Deep Learning algorithms. Tokenization can be defined as the process of dividing the input text into smaller subunits called tokens. Tokens may be a word, a character or even may be a subword.

5.1. WORD TOKENIZATION

This tokenizer is one of the most commonly used technique. In this a text is split into words based on certain delimiters. Based on delimiters different level of words are formed. One of the main disadvantage of this is dealing with the Out of Vocabulary words. Out of Vocabulary words are nothing but the new words that are encountered when testing. Even though we can overcome this concern by using a small trick called Unknown tokens. Another concern with this approach is that the size of vocabulary which it should process.

5.2. CHARACTER TOKENIZATION

This tokenizer will split the words into certain set of characters and it also overcomes the disadvantages of word tokenization. As we are going to use only 26 unique set of characters to represent tokens, it helps to overcome the concern over huge vocabulary size. It overcomes the concern over Out of Vocabulary words by splitting them down into characters and represent the word in terms of the characters. Even though it overcomes the concerns of word tokenization,

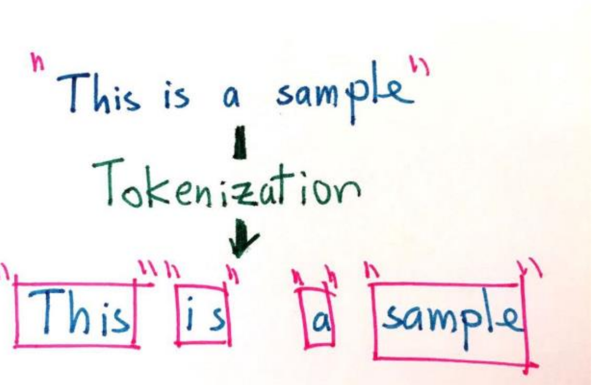


Fig -4: Word tokenization

it also has few concerns. The main concern is that the length of characters increases abruptly as we are representing words in term of characters making it complex to learn the relationship between the characters so that to form meaningful words.

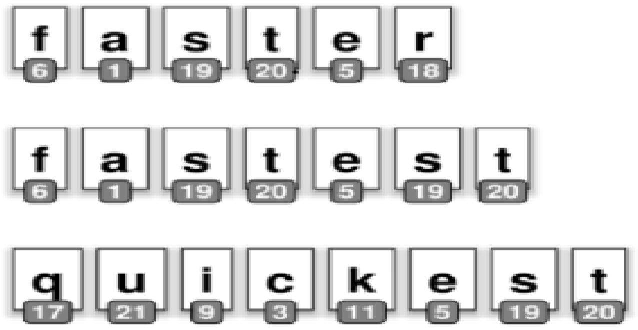

f 6 a 1 s 19 t 20 e 5 r 18
f 6 a 1 s 19 t 20 e 5 s 19 t 20
q 17 u 21 i 9 c 3 k 11 e 5 s 19 t 20

Fig -5: Character tokenization

5.3. SUBWORD TOKENIZATION

This tokenizer will split a piece of text into subwords, like lower into low-er and smallest to small-est, such that to overcome the disadvantages of word tokenization and character tokenization.

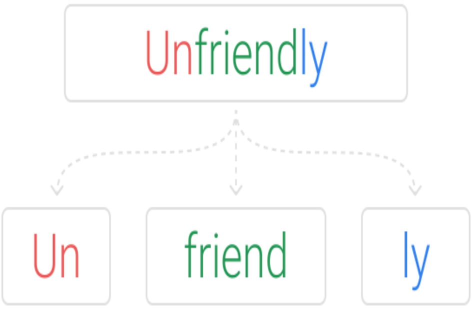


Fig -6: Subword tokenization

6. PROPOSED MODEL

6.1. LANGUAGEANDVISIONCONCAT MODEL

In this model we are going to use torchvision model to extract meme image features and then use fasttext to extract the language features from the meme and finally we concatenate them to form a multimodel hateful meme detector.

The LangaugeAndVisionConcat model consist of three major subblock namely 1) Input embedding, 2) Vision language transform and 3) Superior signals.

6.1.1. INPUT EMBEDDING

This subblock is the first layer of the model and it acts as an input layer. The Input embedding layer is further divided into 1) Object detector and 2) Tokenizer.

1) The Object detector block will extract the image level features from the input image feeded to it with the help of fast Regional based Convolutional Neural Network (fast RCNN).

2) The Tokenizer block processes the input text data and divide the input sentence into smaller tokens or units with the help of word tokenizer in the existing model. In this layer we are going to implement subword tokenizer instead of word tokenizer as word tokenizer has many drawbacks like the Out of Vocabulary (OOV) words which will be addressed by subword tokenization.

6.1.2. VISION LANGUAGE TRANSFORM

This subblock is the middle layer of the model. The vision language transform layer is further divided into 1) Intra-modality transform and 2) Inter-modality transform.

1) The intra-modality transform allows the model, the interaction inside a single modality (either language or vision). Thus the query, key and value processed comes from the same modality.

2) In the Inter-modality transform information is flowing between the two modalities (Vision and Langauage). This layer is defined similarly to that of intra-modality transform but the key and value vectors are cross used between the modalitites.

To improve language feature extraction it is widely shared that capturing relationships among multi-modality features would be helpful for representing and ultimately describing an image. To bridge the gap between inter-modality feature representations, we align them explicitly via Visual Guided Alignment (VGA) module. VGA is devised to compensate for the lack of visual information in sequence self-attention so that the model can produce more reasonable attention results for accurate and image-associated descriptions.

6.1.2. OUTPUT LAYER

The features or information of the image and sentence from previous layers are processed and it comes finally to the output layer, where the output will be like if 0 is the output then the meme is classified as Non-hateful and if the output is 1 then he meme is classified as hateful meme according to our classifier.

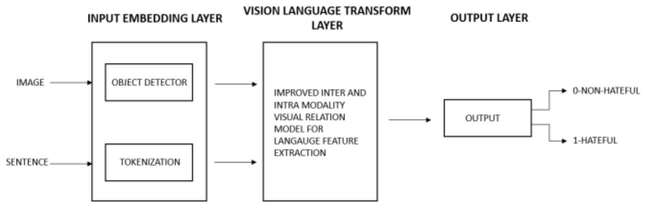
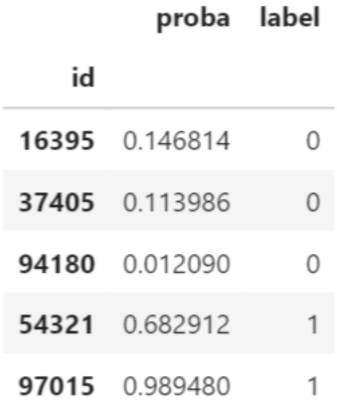

INPUT EMBEDDING LAYER
VISION LANGUAGE TRANSFORM LAYER
OUTPUT LAYER
IMAGE
SENTENCE
OBJECT DETECTOR
TOKENIZATION
IMPROVED INTER AND INTRA MODALITY VISUAL RELATION MODEL FOR LANGAUGE FEATURE EXTRACTION
OUTPUT
0-NON-HATEFUL
1-HATEFUL

Fig -7:Proposed model



	proba	label
id		
16395	0.146814	0
37405	0.113986	0
94180	0.012090	0
54321	0.682912	1
97015	0.989480	1

Fig -8:Sample Output

7. CONCLUSION

Take an image, add some text: you have a meme. Internet memes are often harmless and sometimes hilarious. However, by using certain types of images, text, or combinations of each of these data modalities, the seemingly non-hateful meme becomes a multimodal type of hate speech, a hateful meme. So this type of hateful memes in social media must be identified and should be warned as hateful as this may affect the whole society. So our project can be applied to detect hateful memes.

REFERENCES

[1] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," arXiv:2005.04790.

[2] Riza Velioglu and Jewgeni Rose, "Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge", arXiv:2012.12975

[3] Yong Wang, WenKai Zhang, Qing Liu, Zhengyuan Zhang, Xin Gao and Xian Sun, "Improving Intra- and Inter-Modality Visual Relation for Image Captioning MM '20: Proceedings of the 28th ACM International Conference on Multimedia.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould and Lei Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," arXiv:1707.07998.