

HATEFUL MEME DETECTION FOR SOCIAL MEDIA APPLICATIONS

PROJECT REPORT

Submitted by

**DARSHAN.G (171EC137)
DEEPAK.K (171EC139)**

*In partial fulfilment for the award of the degree
of*

BACHELOR DEGREE

in

**ELECTRONICS AND COMMUNICATION
ENGINEERING**



**BANNARI AMMAN INSTITUTE OF TECHNOLOGY
(An Autonomous Institution Affiliated to Anna University, Chennai)
SATHYAMANGALAM-638401**

ANNA UNIVERSITY: CHENNAI 600 025

MAY 2021

HATEFUL MEME DETECTION FOR SOCIAL MEDIA APPLICATIONS

PROJECT REPORT

Submitted by

**DARSHAN.G (171EC137)
DEEPAK.K (171EC139)**

*In partial fulfilment for the award of the degree
of*

BACHELOR DEGREE

in

**ELECTRONICS AND COMMUNICATION
ENGINEERING**



**BANNARI AMMAN INSTITUTE OF TECHNOLOGY
(An Autonomous Institution Affiliated to Anna University, Chennai)
SATHYAMANGALAM-638401**

ANNA UNIVERSITY: CHENNAI 600 025

MAY 2021

BONAFIDE CERTIFICATE

Certified that this project report **“HATEFUL MEME DETECTION FOR SOCIAL MEDIA APPLICATIONS”** is the bonafide work of **“DARSHAN.G (171EC137) AND DEEPAK.K (171EC139)”** who carried out the project work under my supervision.

SIGNATURE

DR. POONGODI.C
HEAD OF THE DEPARTMENT
Professor,
Department of ECE,
Bannari Amman Institute of Technology,
Sathyamangalam.

SIGNATURE

MR.SURESH.G
SUPERVISOR
Assistant Professor.
Department of ECE,
Bannari Amman Institute of Technology,
Sathyamangalam.

Submitted for Project Viva Voce examination held on

Internal Examiner

External Examiner

DECLARATION

We affirm that the project work titled **“HATEFUL MEME DETECTION FOR SOCIAL MEDIA APPLICATIONS”** being submitted in partial fulfilment for the award of the degree of **BACHELOR OF ENGINEERING IN ELECTRONICS AND COMMUNICATION ENGINEERING** is the record of original work done by us under the guidance of **MR. SURESH G**, Assistant Professor, Department of Electronics and Communication Engineering. It has not formed a part of any other project work(s) submitted for the award of any degree or diploma, either in this or any other University.

DARSHAN.G
(171EC137)

DEEPAK.K
(171EC139)

I certify that the declaration made above by the candidates is true.

MR. SURESH G

ACKNOWLEDGEMENT

We would like to enunciate heartfelt thanks to our esteemed Chairman **Dr.S.V.Balasubramaniam**, and the respected Director **Dr.M.P.Vijaykumar**, for providing excellent facilities and support during the course of study in this institute.

We are grateful to **Dr.C.Poongodi, Professor, Department of Electronics and Communication Engineering** for his / her valuable suggestions to carry out the project work successfully.

We wish to express our sincere thanks to **Mr.V.Baranidharan, Assistant Professor**, for his/her constructive ideas, inspirations, encouragement and much needed technical support extended to complete our project work.

We wish to express our sincere thanks to Faculty guide **Mr.G.Suresh, Assistant Professor, Department of Electronics and Communication Engineering** for his/ her constructive ideas, inspirations, encouragement, excellent guidance and much needed technical support extended to complete our project work.

We would like to thank our friends, faculty and non-teaching staff who have directly and indirectly contributed to the success of this project.

DARSHAN.G(171EC137)

DEEPAK.K(171EC139)

TABLE OF CONTENTS

CHAPTER NO.	DESCRIPTION	PAGE NO.
	ACKNOWLEDGEMENT	5
	ABSTRACT	9
	TABLE OF CONTENTS	6
	TABLE OF FIGURES	8
1	INTRODUCTON	10
2	LITRATURE REVIEW	11
3	DATASET DESCRIPTION	13
4	PROPOSED WORK	14
4.1	DATA FUSION	14
4.1.1	EARLY FUSION	14
4.1.2	LATE FUSION	15
4.2	TOKENIZATION	15
4.2.1	WORD TOKENIZATION	15
4.2.2	CHARACTER TOKENIZATION	16
4.2.3	SUBWORD TOKENIZATION	17
4.3	IMPROVING INTER AND INTRA MODALITY VISUAL RELATION MODEL FOR LANGUAGE FEATURE EXTRACTION	17
4.4	LANGUAGEANDVISIONCONCAT MODEL ARCHITECTURE	18
4.4.1	LANGUAGEANDVISIONCONCAT MODEL FLOW	18
4.4.2	PARTS OF LANGAGEANDVISIONCONCAT MODEL	19
4.4.2.1	INPUT EMBEDDING	19
4.4.2.1.1	OBJECT DETECTOR	20
4.4.2.1.2	TOKENIZER	20
4.4.2.2	VISION LANGUAGE TRANSFORM	20
4.4.2.2.1	INTRA-MODALITY TRANSFORM	21
4.4.2.2.2	INTER-MODALITY TRANSFORM	21
4.4.2.3	SUPERVISION SIGNALS	21
4.4.2.3.1	VISION LANGUAGE MASKING	21
4.4.2.3.2	IMAGE-SENETENCE MATCHING	21
4.4.2.3.3	VISUAL QUESTION ANSWERING	22
4.5	APPLICATIONS	22
4.6	EXISTING MODEL VS PROPOSED MODEL	22

5	ANALYSIS OF RESULT	23
5.1	RESULT	23
6	CONCLUSION AND REFERENCE	24
6.1	CONCLUSION	24
6.2	REFERENCE	24

TABLE OF FIGURES

FIGURE NO.	DESCRIPTION	PAGE NO.
3.1	HATEFUL MEME SAMPLE	13
4.1	EARLY FUSION	14
4.2	LATE FUSION	15
4.3	WORD TOKENIZATION	16
4.4	CHARACTER TOKENIZATION	16
4.5	SUBWORD TOKENIZATION	17
4.6	IMPROVING INTER AND INTRA MODALITY VISUAL RELATION MODEL FOR LANGUAGE FEATURE EXTRACTION	18
4.7	LANGUAGEANDVISIONCONCAT ARCHITECTURE	18
4.8	LANGUAGEANDVISIONCONCAT MODEL FLOW DIAGRAM	19
4.9	OBJECT WORD ALIGNMENT DECODER	21
6.1	PREDICTED OUTPUT	23
6.2	OUTPUT STORED IN CSV	23

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.

INTRODUCTION

1.1. 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 LanguageAndVisionConcat model which is generally a multimodel algorithm which helps to preform complex tasks like the hateful meme detection.

LITRATURE SURVEY

2.1. LITRATURE SURVEY

1) The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes by Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine

This work 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: difficult examples ("benign confounders") are added to the dataset to make it hard to rely on unimodal signals. The task requires subtle reasoning, yet is straightforward to evaluate as a binary classification problem. We provide baseline performance numbers for unimodal models, as well as for multimodal models with various degrees of sophistication. 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.

2) Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge by Riza Velioglu and Jewgeni Rose

Memes on the Internet are often harmless and sometimes amusing. However, by using certain types of images, text, or combinations of both, the seemingly harmless meme becomes a multimodal type of hate speech -- a hateful meme. The 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. Our approach achieves 0.811 AUROC with an accuracy of 0.765 on the challenge test set and placed third out of 3,173 participants in the Hateful Memes Challenge.

3) Improving Intra- and Inter-Modality Visual Relation for Image Captioning by Yong Wang, WenKai Zhang, Qing Liu, Zhengyuan Zhang, Xin Gao and Xian Sun

It is widely shared that capturing relationships among multi-modality features would be helpful for representing and ultimately describing an image. In this paper, we present a novel Intra- and Inter-modality visual Relation Transformer to improve connections among visual features, termed I2RT. Firstly, we propose Relation Enhanced Transformer Block (RETB) for image feature learning, which strengthens intra-modality visual relations among objects. Moreover, to bridge the gap between inter-

modality feature representations, we align them explicitly via Visual Guided Alignment (VGA) module. Finally, 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. Extensive ablation experiments are conducted for the comprehensive analysis of the proposed method.

4) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering by Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould and Lei Zhang

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 this work, we propose 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 our 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. Applying this approach to image captioning, our results on the MSCOCO test server establish a new state-of-the-art for the task, achieving CIDEr / SPICE / BLEU-4 scores of 117.9, 21.5 and 36.9, respectively. Demonstrating the broad applicability of the method, applying the same approach to VQA we obtain first place in the 2017 VQA Challenge.

DATASET DESCRIPTION

3.1 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 3.1. HATEFUL MEME SAMPLE

PROPOSED WORK

4.1. 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.

4.1.1. EARLY FUSION

Early fusion can defined as the fusion technique 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.

Early fusion

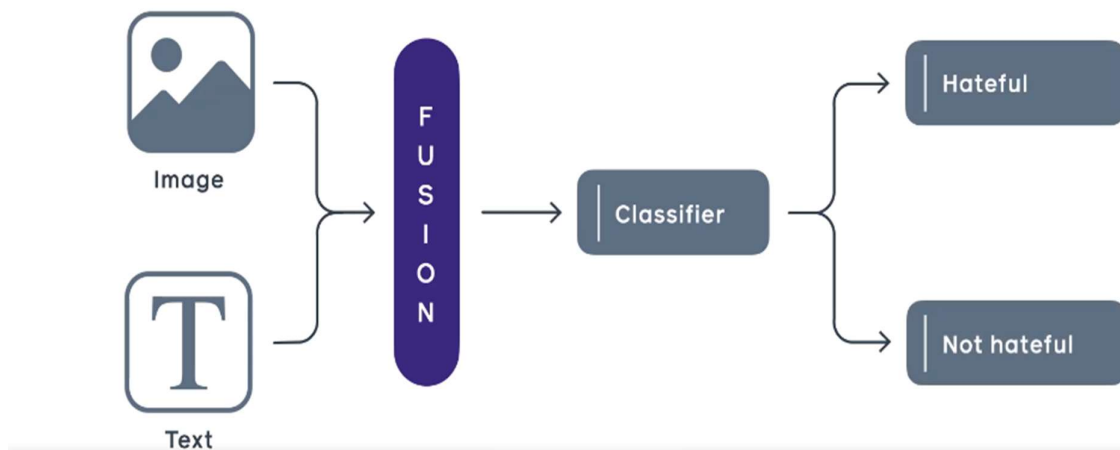


Fig 4.1. EARLY FUSION

4.1.2. LATE FUSION

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

Late fusion

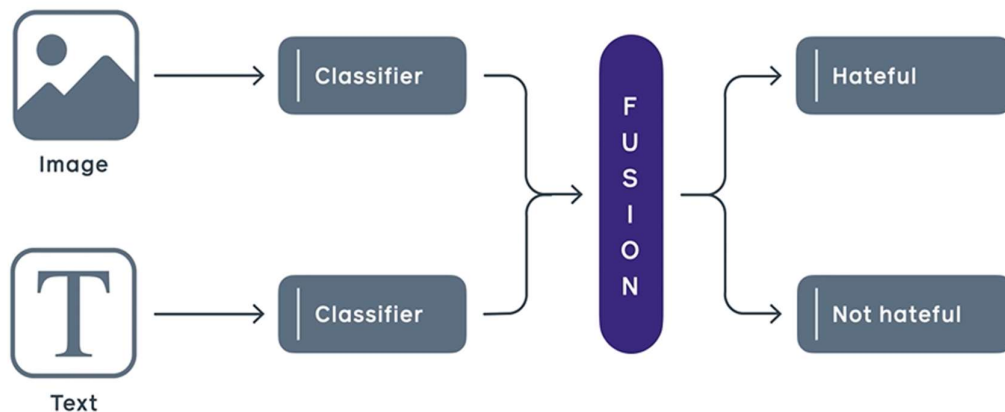


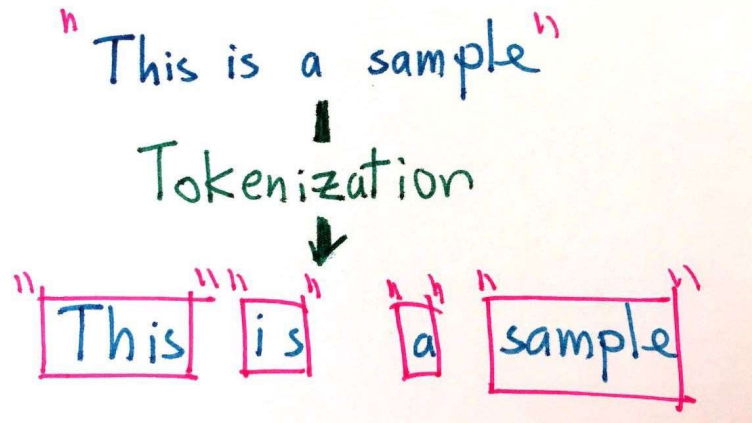
FIG 4.2. LATE FUSION

4.2. 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.

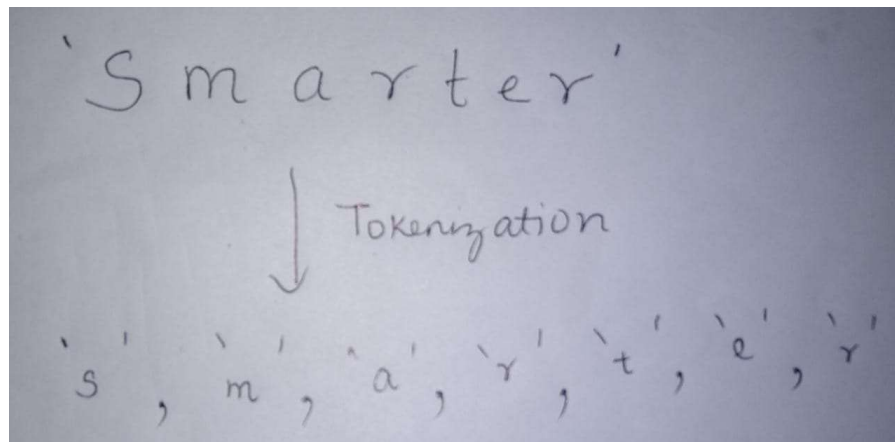
4.2.1. WORD TOKENIZATION

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.

**FIG 4.3. WORD TOKENIZER**

4.2.2. 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, 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.

**FIG 4.4. CHARACTER TOKENIZER**

4.2.3. SUBWORD TOKENIZATION

This tokenizer will split a piece of text into subwords, such that to overcome the disadvantages of word tokenization and character tokenization. The main advantage of a subword tokenizer is that it interpolates between word-based and character-based tokenization. Common words get a slot in the vocabulary, but the tokenizer can fall back to word pieces and individual characters for unknown words.

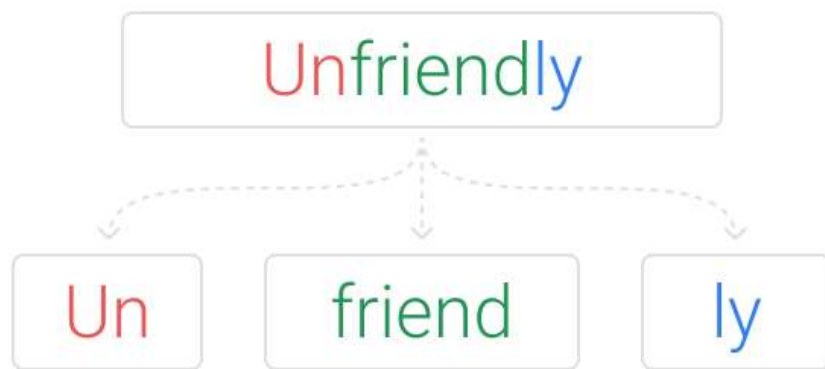


FIG 4.5. SUBWORD TOKENIZER

4.3. IMPROVING INTER AND INTRA MODALITY VISUAL RELATION MODEL FOR 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.

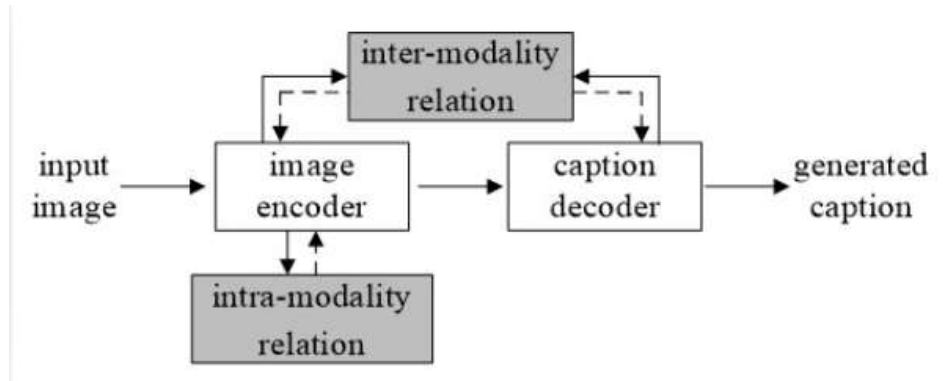


FIG 4.6. IMPROVING INTER AND INTRA MODALITY VISUAL RELATION MODEL FOR LANGUAGE FEATURE EXTRACTION MODEL

4.4. LANGUAGEANDVISIONCONCAT MODEL ARCHITECTURE

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 multimodal hateful meme detector

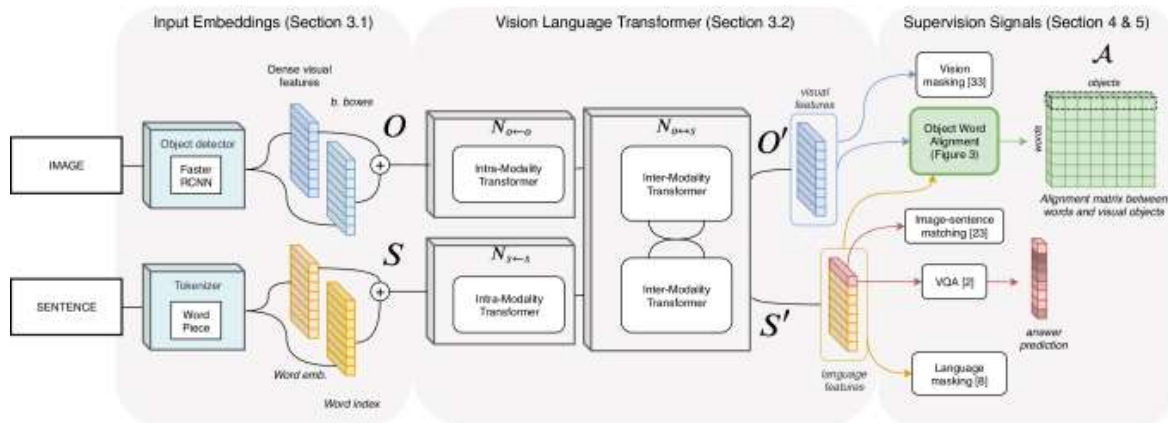


FIG 4.7. LANGUAGEANDVISIONCONCAT ARCHITECTURE

4.4.1. LANGUAGEANDVISIONCONCAT MODEL FLOW

The flow diagram of LanguageAndVisionConcat model is defined below.

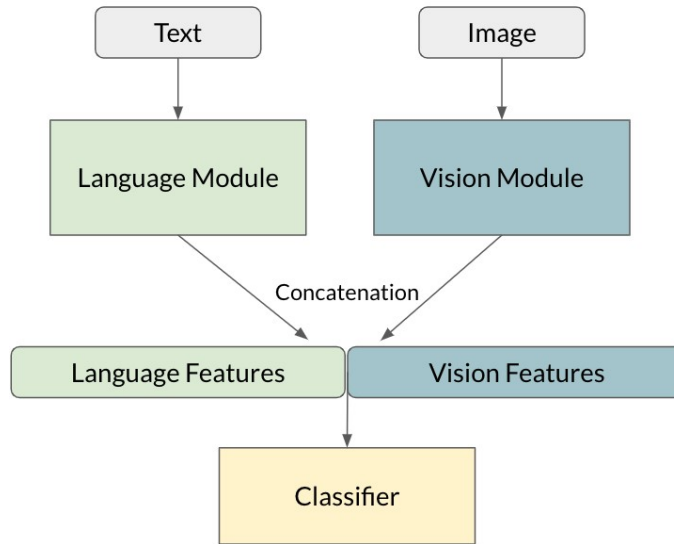


FIG 4.8. LANGUAGEANDVISIONCONCAT MODEL FLOW DIAGRAM

4.4.2. PARTS OF LANGUAGEANDVISIONCONCAT MODEL

The LanguageAndVisionConcat model consists of three major subblocks: 1) Input embedding, 2) Vision language transform, and 3) Superior signals. The language module uses fasttext embeddings as input, computed as the `text_transform` in our data generator (we'll keep the embeddings fixed for simplicity, although they are fit to our training data). The outputs of the language module will come from a trainable Linear layer, as a way of fine-tuning the embedding representation during training. The vision module inputs will be normalized images, computed as the `image_transform` in our data generator, and the outputs will be the outputs of a ResNet model.

4.4.2.1. 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 Object detector and Tokenizer.

4.4.2.1.1. OBJECT DETECTOR

This 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). On the vision side, we use an object detector – Faster-RCNN to extract object level-visual features from the input image. Similar to hard attention mechanisms, this enforces the system to reason on object level rather than on the pixel level or global level.

4.4.2.1.2. TOKENIZER

This block processes the input text data and divide the input sentence into smaller tokens or units with the help of word tokenizer. This smaller units are then sent to next layers for further processing. On the language side, sentences are tokenized using the WordPiece tokenizer. As common in language processing, a special token [CLS] is added at the beginning of the tokenized sentence, which encodes the multimodal information of the image and sentence. The transformation of this token, performed during the forward pass through the network, corresponds to the prediction of the answer to the task. Tokens are embedded into d-dimensional vectors using a look-up table learned during a pre-training phase which concentrates on language only. The index position of the word is added to the dense vector as a positional encoding in order to obtain index-aware word level embedding.

4.4.2.2. VISION LANGUAGE TRANSFORM

This subblock is the middle layer of the model. The vision language transform layer is further divided into Inter-modality transform and Intra-modality transform. The vision-language transformer is composed of two self-attention modules of type intra-modality transformer and inter-modality transformer. They take as input one input sequence (in case of intra-modality) or two input sequences (in case of inter-modality).

4.4.2.2.1. INTRA-MODALITY TRANSFORM

It 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.

4.4.2.2.2. INTER-MODALITY TRANSFORM

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

4.4.2.3. SUPERVISION SIGNALS

We train the vision-language encoder following the recently widely-adapted strategy of combining BERT-like self-supervised signals with task-specific supervision signals, which has been applied to various problems in vision and language. We select four supervision signals: vision masking, language masking, image-sentence matching and visual question answering.

4.4.2.3.1. VISION/LANGAUGE MASKING

This signal aims to supervise the encoder's ability to reconstruct missing information in language and vision. More precisely, we randomly mask each language token (resp. visual object) with a probability of 0.15 and ask the model to predict the missing words (resp. objects). Therefore we add two classifiers – for vision masking⁵ and language masking – on top of the vision language encoder and supervise via a cross-entropy loss. [33] proposes to take the object detector prediction as ground truth in order to get over the disparity of visual annotation. Additionally, we also supervise the model to regress the masked objects' FasterRCNN features via L2 loss.

4.4.2.3.2. IMAGE-SENTENCE MATCHING

BERT [8] proposes next sentence prediction supervision by asking to predict if two sentences are consecutive in a given text, or randomly sampled from a corpus. Its vision-language equivalent is image-sentence matching, where the model has to predict whether a given sentence matches a given image or not. Thus, we randomly replace the image in each sentence-image pair with a probability of 0.5. We add a feed-forward layer on top of the [CLS] output embedding to predict whether the pair matches or not. This global matching is supervised using a binary cross-entropy loss.

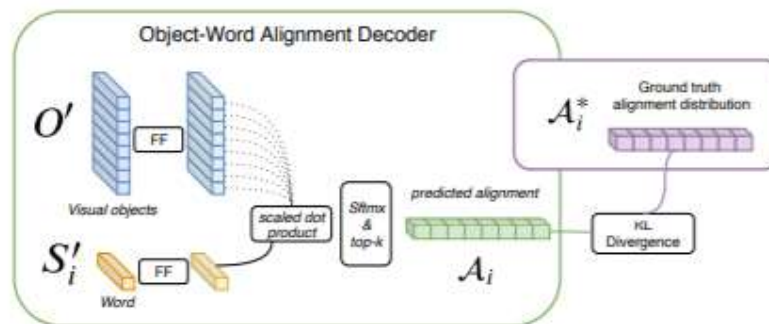


FIG 4.9. OBJECT WORD ALIGNMENT DECODER

4.4.2.3.3. VISUAL QUESTION ANSWERING

Our model is applicable to a wide range of vision-language problems (in Section 6 we evaluate it to two different tasks, namely VQA and Language-driven Comparison of Images). At the same time, independently of the target visionlanguage task, pretraining on VQA helps reasoning as shown in [33]. The VQA task is defined as a classification problem over a set of most frequent answers. In this work, we perform this classification from a prediction head attached to the [CLS] token and supervise it using a cross-entropy loss.

4.5. APPLICATIONS

- 1) Useful in finding Hateful Memes that are shared in social medias
- 2) Subset of this is the hateful comments detection which can be useful in apps like Instagram, Youtube and also in Ecommerce apps where they record customer comments to improve their services.
- 3) Can also used in Search menu of browser to prevent searching online for offensive contents likes photos that abuses a caste or skin tone or a religion.

4.6. EXISTING MODEL VS PROPOSED MODEL

Existing model uses word tokenizer to split the sentence into smaller subunits and uses normal Inter and Intra modality transforms. But in our proposed system the model uses subword tokenizer to overcome concerns of word tokenizer and uses a improved Inter and Intra modality transform for Language Feature Extraction.

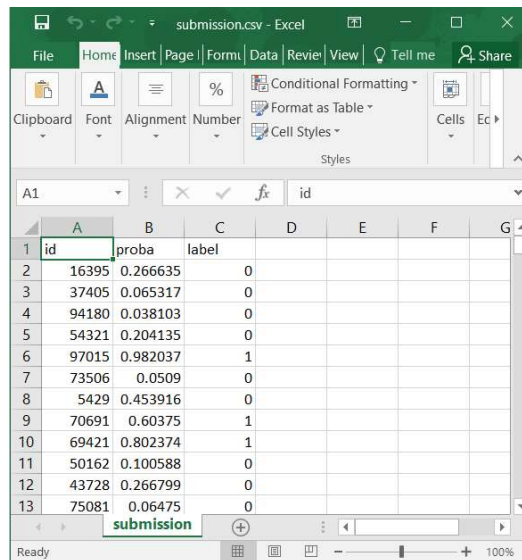
ANALAYSIS OF RESULT

6.1. RESULT

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

FIG 6.1. PREDICTED OUTPUT
(0-NON HATEFUL, 1-HATEFUL)

The output from the model will be stored in a csv file and can be downloaded whenever required.



id	proba	label
16395	0.266635	0
37405	0.065317	0
94180	0.038103	0
54321	0.204135	0
97015	0.982037	1
73506	0.0509	0
5429	0.453916	0
70691	0.60375	1
69421	0.802374	1
50162	0.100588	0
43728	0.266799	0
75081	0.06475	0

FIG 6.2. OUTPUT STORED IN CSV

CONCLUSION

7.1. 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.

7.2. REFERENCES

1. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision.
2. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
3. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).
4. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).Google Scholar
5. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.
6. Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. 2019. Improving image captioning with conditional generative adversarial nets. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 8142--8150.
7. Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015).

8. Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In Advances in neural information processing systems. 3844--3852.
9. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. Ieee, 248--255.
10. Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3270--3277.
11. Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse+: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017).
12. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770--778.
13. Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In Advances in Neural Information Processing Systems. 11135--11145.
14. Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on Attention for Image Captioning. (2019), 4634--4643.
15. Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015).
16. Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV). 499--515.
17. Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yuwing Tai. 2019. Reflective Decoding Network for Image Captioning. (2019), 8888--8897.
18. Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

19. Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out. 74--81.
20. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740--755.
21. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition. 375--383.
22. Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7219--7228.
23. Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 11--20.
24. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 311--318.
25. Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732 (2015).
26. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7008--7024.
27. Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In CVPR 2011. IEEE, 1745--1752.
28. Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 815--823.
29. Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In Proceedings of the

- IEEE Conference on Computer Vision and Pattern Recognition. 8317--8326.
30. Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016).
 31. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. 5998--6008.
 32. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4566--4575.
 33. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156--3164.
 34. Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems?. In Proceedings of the IEEE conference on computer vision and pattern recognition. 203--212.
 35. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning. 2048--2057.
 36. Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 10685--10694.
 37. Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV). 684--699.
 38. Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy Parsing for Image Captioning. (2019), 2621--2629.
 39. Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision. 4894--4902.

40. Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision. 19—27
41. <https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set>

WORK CONTRIBUTION

Project Title: Hateful Meme Detection for Social media applications

Student Name: Darshan G

Register Number: 171EC137

In order for AI to become a more effective tool for detecting hate speech, it must be able to understand content the way people do: holistically. When viewing a meme, for example, we don't think about the words and photo independently of each other; we understand the combined meaning together. This is extremely challenging for machines, however, because it means they can't just analyze the text and the image separately. They must combine these different modalities and understand how the meaning changes when they are presented together. We have identified a multimodel called LanguageandVisionConcat model which performs well in identifying hateful memes which uses word tokenizer which has some drawbacks like identifying Out of Vocabulary(OOV) words. OOV words are nothing but the model faces new vocabulary which it has not come across in previous iterations which I have overcome by proposing a model which uses Subword tokenizer which helps to over the disadvantages of Word Tokenizer. Another issue with word tokens is connected to the size of the vocabulary. Generally, pre-trained models are trained on a large volume of the text corpus. So, just imagine building the vocabulary with all the unique words in such a large corpus. This explodes the vocabulary. This opens the door for usage of Subword Tokenizer.

DARSHAN G

Name and Signature of Student

WORK CONTRIBUTION

Project Title: Hateful Meme Detection for Social media applications

Student Name: Deepak K

Register Number: 171EC139

“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.” This work 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: difficult examples ("benign confounders") are added to the dataset to make it hard to rely on unimodal signals. The task requires subtle reasoning, yet is straightforward to evaluate as a binary classification problem. We have identified a multimodel called LanguageandVisionConcat model which performs well in identifying hateful memes which uses Inter-Modality and Intra-Modality Transforms that has low efficiency in effective generation of captions for the meme passed. So I have improved the performance of the transform by replacing it with Inter and Intra modality visual relation model for language feature extraction model. So by using this model I have improved the performance of the model.

DEEPAK K

Name and Signature of Student