MULTI CHANNEL CONVOLUTIONAL NEURAL NETWORK FOR MEME SARCASM CLASSIFICATION

[1]S.Prathap, [2]P.Yeshika, [3]K.Thapaswi, [4]S.Harishini, [5]E.Varshini

[1]Assistant Professor, Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad, INDIA.

[2345]UG,Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad , INDIA.

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.

Natural language is a vital information source of human sentiments. Automated sarcasm detection is often described as a natural language processing (NLP) problem as it primarily requires understanding the human expressions, language, and/or emotions articulated via textual or nontextual content. Sarcasm detection has attracted growing interest over the past decade as it facilitates accurate analytics in online comments and reviews [1, 2]. As a figurative literary device, sarcasm makes use of words in a way that deviates from the conventional order and meaning thereby misleading polarity classification results. For example, in a statement “Staying up till 2:30am was a brilliant idea to miss my office meeting,” the positive word “brilliant” along with the adverse situation “miss my office meeting” conveys the sarcasm, because sarcasm has an implied sentiment (negative) that is different from surface sentiment (positive due to presence of “brilliant”). Various rule-based, statistical, machine learning, and deep learning-based approaches have been reported in pertinent literature on automatic sarcasm detection in single sentences that often rely on the content of utterances in isolation. These include a range of techniques such as sense disambiguation [3] to polarity flip detection in text [4] and multimodal (text +image) content [5, 6].

2 LITREATURE 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-toend 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 bottomup mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings.

3 METHODOLOGY

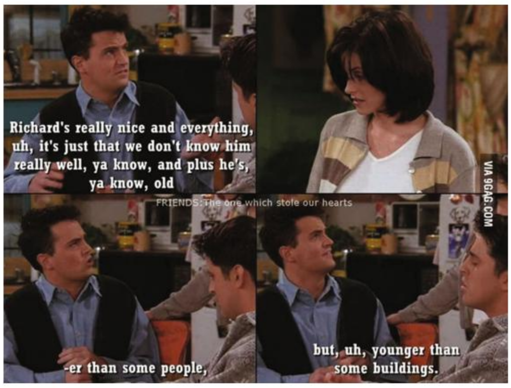
Richard's really nice and everything, uh, it's just that we don't know him really well, ya know, and plus he's, ya know, old
FRIENDS: The one which stole our hearts
VIA 9GAG.COM
-er than some people,
but, uh, younger than some buildings.

Fig 1: Sample Meme

Dataset:

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.

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.

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.

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.

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.

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

asa

4 RESULTS

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 0.5. The results of compared models on the dataset were from [59]. 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 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. 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. Moreover, it may help to fuse different unimodal pre-training models using our method in future studies for similar tasks.

These results indicated that the accuracy of the multi-task model only with the unimodal textual auxiliary or only with the unimodal visual auxiliary task is very similar in hateful meme detection. Furthermore, both the results were also very close compared to the multimodal task, which showed that the accuracy of detecting hateful memes could hardly be improved by adding a single unimodal auxiliary task alone. In contrast, the multi-task learning model was greatly enhanced with the addition of a unimodal textual auxiliary task and a unimodal visual auxiliary task. Moreover, all the cases optimized using equal weights with ω j u = 1 performed worse than the same model using the adaptive weight adjustment strategy. In conclusion, the multi-task learning and the adaptive weight adjustment strategy helped improve the testing accuracy and reduce the generation errors

5 CONCLUSION

Our research aims to improve the accuracy and reduce generalization errors of detecting hateful memes, which are widely available on the Internet and have severe negative impacts. For this purpose, we selected a multimodal dataset of hateful memes published by

Facebook AI as our experimental dataset. Moreover, we designed a multi-task learning model that can generate auxiliary labels self-supervised. A text classification model BERT and an image classification model RESNET were selected as the backbone, and a late fusion method was used. In the multi-task learning network, we added two unimodal auxiliary learning tasks, the textual and the visual auxiliary task, to the primary classification task. In order to solve the problem of lacking labels for the unimodal auxiliary tasks and the high cost of manual labeling, we chose a strategy of self-supervised label generation for the auxiliary tasks. In the phrase of optimization, we added a data-driving adaptive weight adjustment strategy to balance the learning process and reduce the generalization errors. By comparing our multi-task learning model with various advanced models for the detection of hateful memes, we can find that our multi-task learning model achieved more accurate results.

6 REFERENCES

1. Devroye, L.; Györfi, L.; Lugosi, G. A Probabilistic Theory of Pattern Recognition; Springer Science & Business Media: New York, NY,USA, 2013 ; Volume 31.

2. Fan, J.; Li, R.; Zhang, C.H.; Zou, H. Statistical Foundations of Data Science; Chapman and Hall/CRC: New York, NY, USA, 2020 .

3. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction;Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.

4. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. Foundations of Machine Learning; MIT Press: Cambridge, MA, USA, 2018.

5. Bertsekas, D.P. Nonlinear programming. J. Oper. Res. Soc. 1997, 48, 334. [CrossRef]

6. Tewari, A.; Bartlett, P.L. On the Consistency of Multiclass Classification Methods. J. Mach. Learn. Res. 2007, 8, 1007–1025.

7. Zhang, T. Statistical analysis of some multi-category large margin classification methods. J. Mach. Learn. Res. 2004, 5, 1225–1251.

8. Vapnik, V.N. An overview of statistical learning theory. IEEE Trans. Neural Netw. 1999, 10, 988–999. [CrossRef] [PubMed]

9. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. arXiv 2017,arXiv:1707.07250.

10. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy,28 July–2 August 2019; NIH Public Access: Bethesda, MD, USA, 2019; Volume 2019, p. 6558.

11. Poria, S.; Hazarika, D.; Majumder, N.; Mihalcea, R. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. IEEE Trans. Affect. Comput. 2020, 1. [CrossRef]

12. Bartlett, P.L.; Jordan, M.I.; McAuliffe, J.D. Convexity, classification, and risk bounds. J. Am. Stat. Assoc. 2006, 101, 138–156. [CrossRef]

13. i Orts, Ò.G. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 460–463.

14. Burnap, P.; Williams, M.L. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. In Proceedings of the Internet, Policy & Politics Conference,Oxford, UK, 26 September 2014.

15. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 2018, 41, 423–443. [CrossRef]

16. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. IEEE Access 2019, 7, 63373–63394. [CrossRef]

17. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.

18. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8992–8999.

19. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th

Annual Meeting of the Association for Computational Linguistics (ACL 2020), online, 5–10 July 2020; NIH Public Access: Bethesda, MD, USA, 2020; Volume 2020, p. 2359.

20. Wang, S.; Zhang, H.; Wang, H. Object co-segmentation via weakly supervised data fusion. Comput. Vis. Image Underst. 2017, 155, 43–54. [CrossRef]