TechSSN at SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification using Deep Learning Models

Rajalakshmi Sivanaiah, Angel Deborah S, Sakaya Milton R, Mirnalinee T T
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai - 603110, Tamil Nadu, India
rajalakshmis@ssn.edu.in, angeldeborahs@ssn.edu.in,
miltonrs@ssn.edu.in, mirnalineett@ssn.edu.in

Abstract

Research is progressing in a fast manner in the field of offensive, hate speech, abusive and sarcastic data. Tackling hate speech against women is urgent and really needed to give respect to the lady of our life. This paper describes the system used for identifying misogynous content using images and text. The system developed by the team TECHSSN uses transformer models to detect the misogynous content from text and Convolutional Neural Network model for image data. Various models like BERT, ALBERT, XLNET and CNN are explored and the combination of ALBERT and CNN as an ensemble model provides better results than the rest. This system was developed for the task 5 of the competition, SemEval 2022.

1 Introduction

In our society, women are facing lot of challenges in terms of education, employment, career and life. Eventhough women are better off now-a-days, there are not considered as equal to men in many situations. With the invent and rapid usage of social media platforms, offensive images and texts conveying several forms of hate against women are transmitted and spread online in a fast manner. Although opportunities for women have been increased on the Internet, systematic inequality and discrimination offline is continued in online in the form of offensive contents through MEMEs. A meme is essentially an image characterized by a pictorial content with an overlaying text introduced by human on it.

Most of the memes are created mainly for funniness, still it is used for ironic purpose too. The task 5 by Fersini et al. (2022) in SemEval 2022 mainly focused on identifying the misogynous memes using textual and image contents. There are two subtasks in this. Subtask A is to categorize the memes as misogynous or not misogynous. Subtask B is to classify the misogynous memes into one of the categories like stereotype, shaming, objectification and violence.

2 Related Work

The survey on various techniques used for Automatic Misogyny Identification (AMI) tasks happened in EVALITA 2018 and IBERALEVAL 2018 were discussed in detail by Shushkevich and Cardiff (2019). Machine learning (SVM, Naive Bayes, Logistic Regression) and deep learning techniques (CNN, GRU, RNN, LSTM) are used and achieved an accuracy of 90% in IBERALEVAL task and 70% in EVALITA task.

García-Díaz et al. (2021) uses sentiment analysis and social computing technologies with word embeddings and linguistic features for AMI that achieves better results than traditional machine learning algorithms. Hate speech against women is handled by Frenda et al. (2019). Pamungkas et al. (2020) created models using RNN and BERT for multilingual misogyny detection.

We have performed irony and offensive language detection in earlier SemEval workshop tasks Sivanaiah et al. (2021), Sivanaiah et al. (2020), Sivanaiah et al. (2019) and Sivanaiah et al. (2018). Various machine learning techniques like linear regression, logistic regression, naive Bayes, Random forest, Support Vector Machines and deep learning techniques like Recurrent Neural Networks, Convolutional Neural Networks, Long Short Term Memory Networks, BERT, ColBERT models are used in the above tasks. BERT and ColBERT models perfomed better than other machine learning and deep learning models.

3 System Overview

The proposed system uses the Transformer models to classify the textual content of the memes and Convolutional Neural Network (CNN) to classify

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 571 - 574

Dataset	Misogynous		Shaming		Stereotype		Objectification		Violence	
	1	0	1	0	1	0	1	0	1	0
Training (10000)	5000	5000	1271	3729	2810	2190	2201	2799	953	4047
Trial (100)	44	56	0	44	34	10	2	42	9	35
Test (1000)	500	500	146	354	350	150	348	152	153	347

Table 1: Data distribution

the visual content. The majority voting ensemble classifier is used to predict the label for the input sample using the output of textual and visual content classifiers. The data is preprocessed to remove the unwanted information before building the model. The architecture of the system is shown in Figure 1.

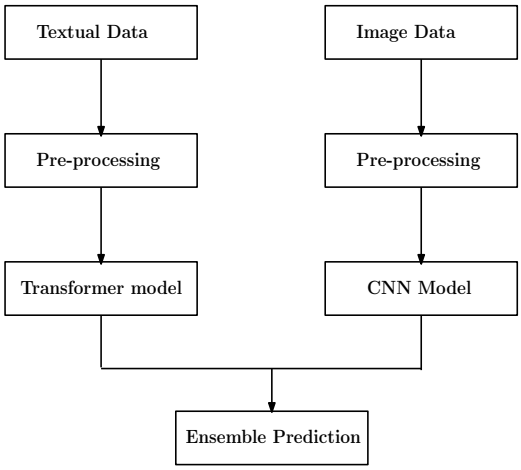

Figure 1: System Architecture

The steps can be summarized as follows:

1. Preprocess the text data and image data separately
2. Separate the image data into folders according to the label category
3. Setup the architecture of the transformer models
4. Train and evaluate the transformer models with text data
5. Use CNN model to train and evaluate the image data
6. Generate the class labels for the test data using transformer model and CNN model
7. Combine the output of both the models using majority voting ensemble classifier

3.1 Dataset

Organizers have provided the data with 10000, 100 and 1000 samples in training, trial and testing datasets. The division of data into various labels are shown in Table 1. Out of 10000 training data, 5000 belongs to misogynous class and 5000 belongs to non-misogynous class. Out of 5000 training misogynous data, 1271 belongs to shaming category and 3729 to not-shaming category. There are 2810 stereotype samples and 2190 non-stereotype samples in 5000 training misogynous data. We have 2201 objectification and 2799 non-objectification samples in misogynous type in training dataset. There are 953 violence and 4047 non-violence samples in misogynous training data.

3.2 Data Pre-processing

Data preprocessing is critical for the success of any machine learning solution to remove the irregularities in the data. Normalization is done to flatten the dimensions of data in textual form. The text data is cleaned and processed to remove URLs, annotate emojis, emoticons, convert uppercase to lowercase, remove stopwords, remove special characters, remove accented characters, lemmatize text, and remove extra whitespace. The images are categorized into their corresponding label folders for further processing. Noise in the image data is removed and converted into pixel values of size 128x128.

3.3 Classification with Transformer models

A transformer is a deep learning model that uses self-attention with weights. Transformer models are used in natural language processing and computer vision. It uses encoder-decoder mechanism to learn features from the sequential input data. Encoder learns the relevant features in the input and pass them to next layer. Decoder does the opposite by taking the output of encoder and incorporate the contextual information to generate the output sequence. Both encoder and decoder layer uses attention mechanism to perform these operations. While Recurrent Neural Networks (RNN)

process the data in the sequential order, transformers does not require this. Bidirectional Encoder Representations from Transformers (BERT) is a popular model developed by Google for language modelling (Devlin et al., 2018). Variants of BERT like A Lite BERT (ALBERT) (Lan et al., 2019), XLNET (Yang et al., 2019), Robustly optimized BERT approach (ROBERTa) and COntextualized Late interaction over BERT (ColBERT) are used to perform the learning on text data. XLnet is an extension of the Transformer-XL model pre-trained using an autoregressive method to learn bidirectional contexts.

BERT model works in a better manner for sparse data representations. Instead of training the model from the base, we can take a pretrained model and tune the model to suit our need. The drawback of BERT model is that it generates many parameters for learning which makes the model complex and time consuming. ALBERT model is a lighter version of BERT that reduces the parameters size without much reduction in performance. ALBERT model uses the multi-headed, multi-layer transformer architecture. The number of epochs used to train the models is 5. The embeddings used are albert-base-v2, xlnet-base-uncased and bert-base-uncased.

3.4 Classification using CNN model

Convolutional Neural Network (CNN) (Albawi et al., 2017) is a deep learning technique that takes the image as input, assign importance in the form of learnable weights, learn the various aspects in the input and classify the data. It uses convolutional, maxpooling and dropout layers to learn the features. CNNs have wide applications in computer vision, medical image processing, NLP and recommender systems. We have used 3 convolutional layer with ReLU activation function, maxpooling layer after each convolutional layer, fully connected layer with ReLU activation function and final output layer with softmax activation. Adam optimizer is used with learning rate 0.001 for 30 epochs.

3.5 Ensemble Prediction

Performance of text model and image model are analyzed with the ensemble technique. Majority voting is done with the output of transformer models and CNN to predict the final output class label. Ensemble method will predict the output class label as misogynous if both transformer and CNN models predict the label as misogynous.

4 Results and Discussions

We have used variants of BERT models for analyzing text data and CNN for image data. The results are tabulated in Table 2. The models are executed for various epochs and the results are listed for 5 epochs.

Models	Subtask A	Subtask B
AlBERT	0.5128	-
XLNet	0.3524	-
BERT	0.4584	0.4137
CNN Model	0.4756	-
Ensemble model	0.5223	0.4673

Table 2: F1-measure Results

In subtask A, we have used AlBERT, XLNet and basic BERT techniques to classify the given sample as misogynous or non-misogynous using the text data. Albert model gives better results than other BERT variants. CNN model is used for the classification of image data. The combined ensemble model (AlBERT + CNN) gives better F1 score than using them separately. For the subtask B, we have worked on two models BERT and ensemble model. Ensemble model with multi-label classification model in transformers and CNN is used for predicting the subcategories. The ensemble model (image and text) achieved better score than the BERT model (text alone) and the performance can be increased by tweaking the hyper parameters.

5 Conclusion

User generated content in social media is rapidly increasing day by day that detecting and limiting the diffusion of sarcastic and hate speech content against women is tedious. Automatic identification and removal of these contents is the current topic of research. Many shared tasks are conducted in various conferences for Automatic Misogyny Identification (AMI).

Task 5 in SemEval 2022 focuses on Multimedia Automatic Misogyny Identification (MAMI) with the help of image and text data. Transformer models are used to identify the misogynous content from the text and CNN model is used to detect from the image content. Performance of the system can be increased by tweeking the parameters in the transformer models. The current system trains the text and visual modalities independently on

the labels. However, memes portray misogynistic content though a combination of text and image cues, and training these models independently might miss out on the context provided by the other modality, which might be crucial to further improve the performance of this system.

References

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4):1159–1164.

Rajalakshmi Sivanaiah, Angel Deborah S, S Milton Rajendram, and Mirnalinee T T. 2018. SSN MLRG1 at SemEval-2018 task 3: Irony detection in English tweets using MultiLayer perceptron. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 633–637, New Orleans, Louisiana. Association for Computational Linguistics.

Rajalakshmi Sivanaiah, Angel Deborah S, S Milton Rajendram, Mirnalinee TT, Abrit Pal Singh, Aviansh Gupta, and Ayush Nanda. 2021. TECHSSN at SemEval-2021 task 7: Humor and offense detection and classification using ColBERT embeddings. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1185–1189, Online. Association for Computational Linguistics.

Rajalakshmi Sivanaiah, Angel Suseelan, Logesh B, Harshini S, Geetika B, Dyaneswaran S, S Milton Rajendram, and Mirnalinee T T. 2019. TECHSSN at SemEval-2019 task 6: Identifying and categorizing offensive language in tweets using deep neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 753–758, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Rajalakshmi Sivanaiah, Angel Suseelan, S Milton Rajendram, and Mirnalinee T.t. 2020. TECHSSN at SemEval-2020 task 12: Offensive language detection using BERT embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2190–2196, Barcelona (online). International Committee for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.