ConOffense: Multi-modal multitask Contrastive learning for offensive content identification

Debaditya Shome
School of Electronics engineering
KIIT University
Odisha, India
0000-0001-9168-0379

T. Kar
School of Electronics engineering
KIIT University
Odisha, India
tkarfet@kiit.ac.in

***Abstract*—Hateful or offensive content has been increasingly common on social media platforms in recent years, and the problem is now widespread. There is a pressing need for effective automatic solutions for detecting such content, especially due to the gigantic size of social media data. Although significant progress has been made in the automated identification of offensive content, most of the focus has been on only using textual information. It can be easily noticed that with the rise in visual information shared on these platforms, it is quite common to have hateful content on images rather than in the associated text. Due to this, present day unimodal text-based methods won't be able to cope up with the multimodal hateful content. In this paper, we propose a novel multimodal neural network powered by contrastive learning for identifying offensive posts on social media utilizing both visual and textual information. We design the text and visual encoders with a lightweight architecture to make the solution efficient for real world use. Evaluation on the MMHS150K dataset shows state-of-the-art performance of 82.6 percent test accuracy, making an improvement of approximately +14.1 percent accuracy over the previous best performing benchmark model on the dataset.**

***Index Terms*—Multimodal learning, Contrastive learning, Representation learning, Social media, Offensive content identification**

I. Introduction

The recent rise in use of social media has resulted in a hyper-connected world. However, numerous instances of abusive language and offensive content often outweighs the possible advantages of social media. It has the potential to cause significant harm to our society and to marginalised individuals or groups. The development in online hate organizations has been paralleled by an increase in web-based hate speech, harassment, bullying, and discrimination, which is directed both directly and indirectly through forums, blogs, and emails [1]. This increase in hate speech online is exacerbated by the difficulty in monitoring such actions, as the Internet remains mostly uncontrolled. Due to the massive scale of generated content, manual mining of such content is also impossible. Most common method utilized by social media platforms to detect hate speech includes keyword spotting, which fails to capture the complex patterns required to detect such content. Thus, automatic detection of offensive content has been of prime interest in the research community [2]. Most progress in this area has been possible by leveraging

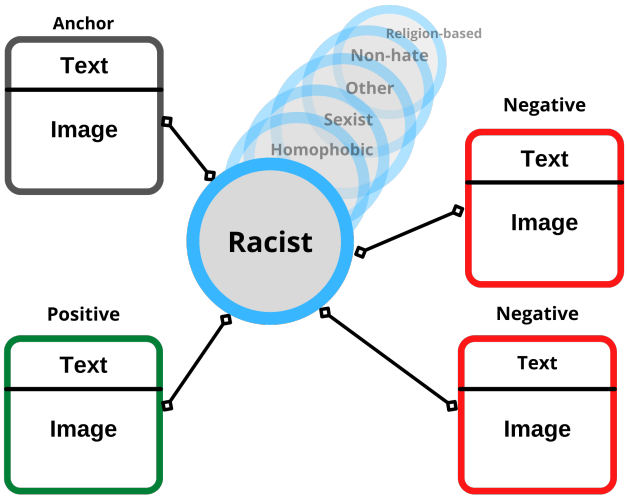

Fig. 1. Multimodal multitask contrastive learning

Natural language processing (NLP) models, and hate speech identification has been considered as a sub-field within NLP [3]. Still, such textual models aren't sufficient in the real world. Users uploading offensive content nowadays tend to use more visual information rather than textual, in order to make such content obscure to hate filters of social media sites. This shows potential for multimodal models which utilize both textual and visual cues to be used as a solution to this challenge. Furthermore, most work has been done on binary classification of hateful content, which limits the scope and impact of such models. Thus, fine-grained multitask classification would be beneficial to distinguish between the different types of offensive content.

In an attempt to solve such challenges, we propose three major contributions.

- We present ConOffense, a multi-modal approach for detecting offensive content on social media by using supervised contrastive learning with fused visual and textual information.
- We extend Supervised Contrastive (SupCon) loss to a multimodal and multitask setting, enabling to perform a fine-grained classification.

- Upon evaluation on the unseen testing split of MMHS150K dataset, our model achieves state of the art performance of 82.6% accuracy, which is an improvement of +14.1% accuracy over previous best baseline.

II. Related work

There is a large amount of progress in offensive content identification already in literature. In [4], the authors used a list of handcrafted features along with character n-grams and annotated a dataset of 16K tweets with hate speech label. In [5], the authors used a L2 regularized Logistic regression model for classifying offensive content and hate speech using TF-IDF features extracted from textual data. They report a misclassification of 40% due to the complex nature of such content. In [6], authors compared multiple machine learning models for multitask hate speech identification on textual data retreived from Indonesian twitter. Their experimental results show a 73.53% best-case accuracy. In [7], the authors compared support vector machine and Bi-LSTM neural network for the task of text-based hate speech identification. The results show SVM to outperform the Bi-LSTM in terms of average accuracy, which points out that deep learning models aren't yet an optimal solution. In [8], the large language model, GPT-3 was fine-tuned for hate speech identification on texts. Experimental results show that GPT-3 was able to achieve upto 78% accuracy using few-shot learning. However, such a large scale model is also not able to reach higher accuracies which show that textual information is not sufficient. In [9], the authors designed a deep learning model fusing affective features with other features for hate speech detection. Upon evaluation their approach achieves an high accuracy of 95.1% on the Davidson dataset, however, the model fails to perform well on SemEval dataset yielding an accuracy of 65.9% only. In [10], multiple deep learning models were compared by training and evaluating over the "Hate and Abusive Speech on Twitter" dataset. The experiments show that bi-directional GRU models with word-level features and Latent Topic Clustering modules, is the best performing model with an F1 score of 0.805. In [11], a generative modelling approach using GPT-2 was used to create synthetic text data for training hate speech detection models. This approach showed an improvement over previous baselines in terms of generalizability to new data. In [12], a unimodal text-based multitask learning methodology was developed to predict sentiment, hate speech and offensive content by a single CNN-BiLSTM model. They achieved a best case F1 score of 0.877 on offensive content, and 0.76 on hate-speech. In [13], a model was trained on text-based English offensive content dataset and tested on multiple different languages without any additional language specific training. Their results show comparable performance on unseen languages. In [14], the authors developed a Graph neural network based approach for identifying hateful content on social media. They incorporated both textual and graph based features, achieving macro f1-score of 0.791 on Gab dataset and 0.780 on twitter dataset. In [15], BERT was re-trained using SOLID, a large offensive text dataset. According to the experimental results, their model achieved a macro F1-score of 0.596, 0.813 and 0.878 on three benchmark datasets. The authors of [16] introduced a novel large-scale dataset of multimodal tweets and benchmarked multiple deep models for hate speech detection. Still, their results show that visual information wasn't much beneficial as their multimodal models weren't able to outperform text-based baselines.

Based on the literature review above, it can be easily seen that most work has focused on unimodal text-based models. However, there is a pressing need for utilizing multimodal data with both visual and textual information in order to tackle the complexity of detecting offensive content in social media. Moreover, most work has focused on classifying between hate and non-hate posts. It would be much more useful in designing a fine-grained classification of offensive content categories.

III. Data preparation

A. Dataset overview

For our research, we use the MMHS150K dataset [16] which consists of 150K multimodal tweets with fine grained labels of offensive or hateful content. The labels include Racist, Sexist, Homophobic, Religion-based and Other hate, which are further encoded in a multitask manner. Each tweet comprises of text associated with an uploaded image. The dataset also has OCR text extracted from images, however we chose to not use the extracted text in order to make our objective focused on learning robust visual representations. The same training, validation and test split of the data is used in order to effectively evaluate and compare our method with the baselines proposed in [16].

B. Pre-processing

1) **Text pre-processing:** Standard text cleaning methods using Regex are first used on lowercase raw texts. All texts are truncated/padded to a fixed output length of 500 tokens.
2) **Image pre-processing:** We resize each image to 128x128 pixels, and normalize each image to have mean of 0 and variance of 1.

IV. ConOffense framework

A. Model architecture

We design our multimodal model with a simple two-stream architecture. The two streams represent different encoders for text and image data. The image encoder F consists of a Deep Residual architecture inspired from ResNet [17]. The text encoder is a simple 1D convolutional network which learns a word embedding W, tailored for offensive content identification. The output embedding of these encoders is fused together using a concatenation layer, and further passed through a 2048 dimensional dense layer, D_1. The overall structure till now is termed as a multimodal encoder. As illustrated in Figure 2, the model is trained in two stages.

In the representation learning stage, the multimodal encoder's output is projected to a 128 dimension vector using

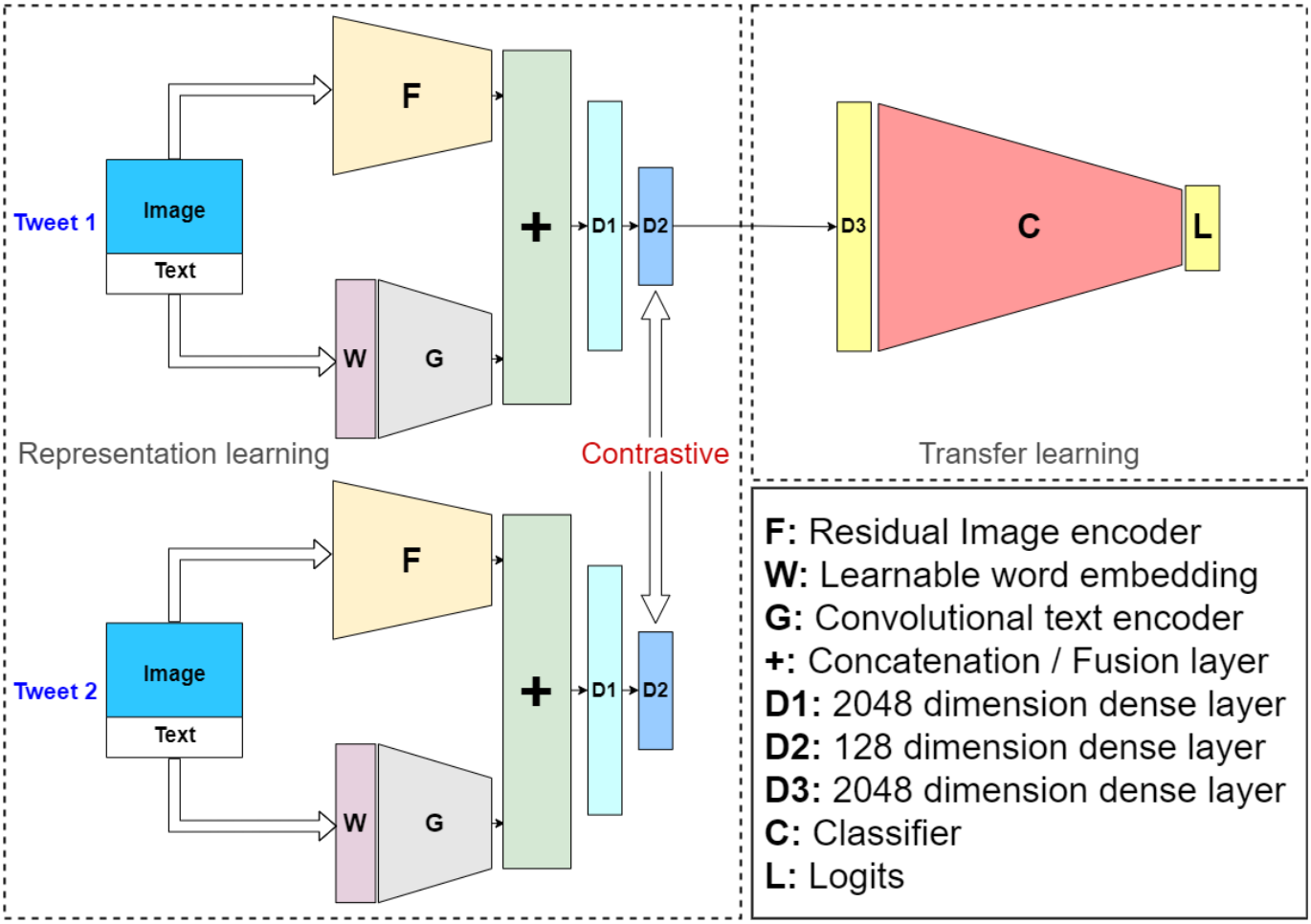

Fig. 2. Model architecture and training

another ReLU activated dense layer, D_2 acting as the projection network. Here, the projection to a smaller dimension is essential to reduce the computation and overfitting.

In the transfer learning stage, a 2048 dimensional ReLU activated Dense layer is attached to the freezed representation learner from the previous stage, followed by a classifier network comprised of few more dense layers.

B. Contrastive representation learning

The first stage in our model's training mechanism involves Contrastive learning. We sample N tweet and label pairs per minibatch, where each tweet comprises of a pair of image I and text T. Each such pair of image and text is passed as input to the multimodal encoder as mentioned in Equation 1 to obtain intermediate representations r_I and r_T, and we obtain a fused representation c_{fusion} by concatenating the intermediate representations. Then, c_{fusion} is passed through D_1 to obtain a 2048 dimension vector, e. Finally, a dense layer D_2 projects e to a 128-dimensional vector z representing the learnt latent representation of the multimodal tweet.

$$\begin{aligned} r_I &= F(I) \\ r_T &= G(W(T)) \\ c_{fusion} &= r_I + r_T \\ e &= D_1(c_{fusion}) \\ z &= D_2(e) \end{aligned} \tag{1}$$

Inspired from the success of SupCon loss [18], we extend it to formulate MMSupCon loss, our multimodal multitask constrastive loss as seen in Equation 2.

$$L = \sum_{k \in K} \sum_{t \in T} \frac{-1}{|P(t,k)|} \sum_{p \in P(t,k)} log \frac{exp(z_t \cdot z_p / \tau)}{\sum_{n \in N(t,k)} exp(z_t \cdot z_n / \tau)} \tag{2}$$

Here, K refers to the set of classes in the dataset, τ is the temperature parameter and $P(t,k)$ represents a two-dimensional matrix in which t refers to indices of tweets with positive label in the particular class k. In $|P(t,k)|$ is the number of positive samples in the class k. Similarly, $N(t,k)$ represents a two-dimensional matrix with all negatively labeled samples of the particular class k. In our problem, we have $k = 6$.

For each class k, z_p and z_n are the projected embeddings of positive and negative samples respectively in that minibatch by our multimodal encoder and projection layer. Our loss contrasts each positive embedding with the embedding of the anchor tweet to reduce their distance in representation space, and all the negative sample embeddings in the minibatch are contrasted in order to increase their distance with the anchor in representation space. A model trained with MMSupCon loss can effectively learn robust representations which can be used for transfer learning with any simple feed-forward classifier. It can achieve state-of-the-art (SOTA) performance on complex social media posts with both text and image data using transfer learning with the representations learnt by the multimodal encoder.

C. Transfer learning

The second stage of our training mechanism involves transfer learning. We freeze the multimodal encoder and attaching a dense layer D_3, followed by a classification network C made up of ReLU activated fully-connected layers. The final layer is sigmoid activated and outputs a logit vector representing the multitask predictions. Each of the elements in the logit vector is a binary prediction of the particular class and the standard cross-entropy loss is used for training.

V. Results and Evaluation

A. Training performance

As discussed in the previous section, our model is trained in a two-step process. In the first stage, the model was trained to optimize the MMSupCon loss as in Equation 2. We used a batch size of 256, a learning rate of 0.0001, and 500 epochs. The training in both stages utilizes the Adam optimizer [19]. In the second stage, the following metrics were used in evaluating the training and validation performance of our model for a total of 50 epochs.

1) **Accuracy:** The accuracy metric is the most commonly used performance metric in any classification task. For the multitask classification, binary accuracy was chosen and the average accuracy over all the six classes of tweets in the dataset resembles the overall accuracy.
2) **AUC score:** The area under the ROC curve (AUC) score indicates how well predictions are rated across all classes and how well the model can differentiate between them. It guarantees that performance is aggregated over all possible classification criteria.

As seen in Figure 3, the model starts with a initial training accuracy as high as 87% due to the previous supervised contrastive pre-training phase. The classifier is a 5-layer fully connected neural network with ReLU activations. The training accuracy and AUC approximately reaches upto 92% and 0.96 respectively. The validation accuracy and AUC goes upto 84% and 0.88 respectively. The validation accuracy doesn't go up after a while, which leaves room for experiments with larger and sophisticated classifiers in future.

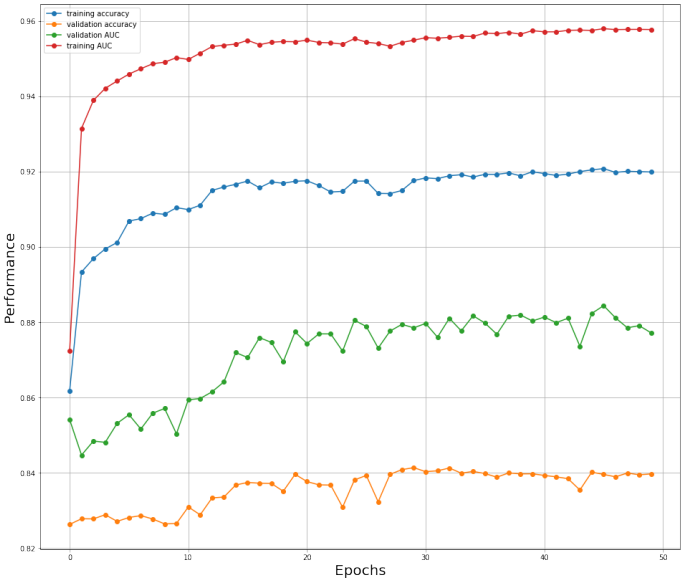

Fig. 3. Training and validation performance during transfer learning

TABLE I
Performance comparison with SOTA on test dataset

Model	Accuracy	AUC	F1-Score
Davidson (text) [5]	68.4%	0.732	0.666
LSTM (text) [16]	68.3%	0.732	0.703
FCM (text) [16]	67.8%	0.727	0.697
FCM (image) [16]	56.8%	0.589	0.667
FCM (image+text) [16]	68.4%	0.734	0.704
SCM (image+text) [16]	68.5%	0.732	0.702
TKM (image+text) [16]	68.2%	0.731	0.701
ConOffense (Ours)	**82.6%**	**0.849**	**0.804**

B. Testing performance

We evaluate our model on the same testing split used in [16] to compare our approach with their baselines efficiently. As seen in Table I, our proposed model outperforms all of the previous baselines by a significant margin. This shows the effectiveness of our contrastive learning approach. It has been pointed out in literature that large number of hard negatives are crucial in learning robust representations with contrastive learning [20]. In our dataset there are a large number of non- hateful tweets present, which thus helps in boosting the performance of our model.

C. Visualization of learnt representations

After the first stage, we freeze the multimodal encoder and use it to extract feature vectors from tweets in the dataset. The t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm is utilized to visualize the high-dimensional feature vectors in two-dimensional plane. As illustrated in Figure 4, the multimodal encoder learns representations which are quite separable between positives and negatives in case of Sexist, Homophobe, Other hate and Racist classes. However, it fails

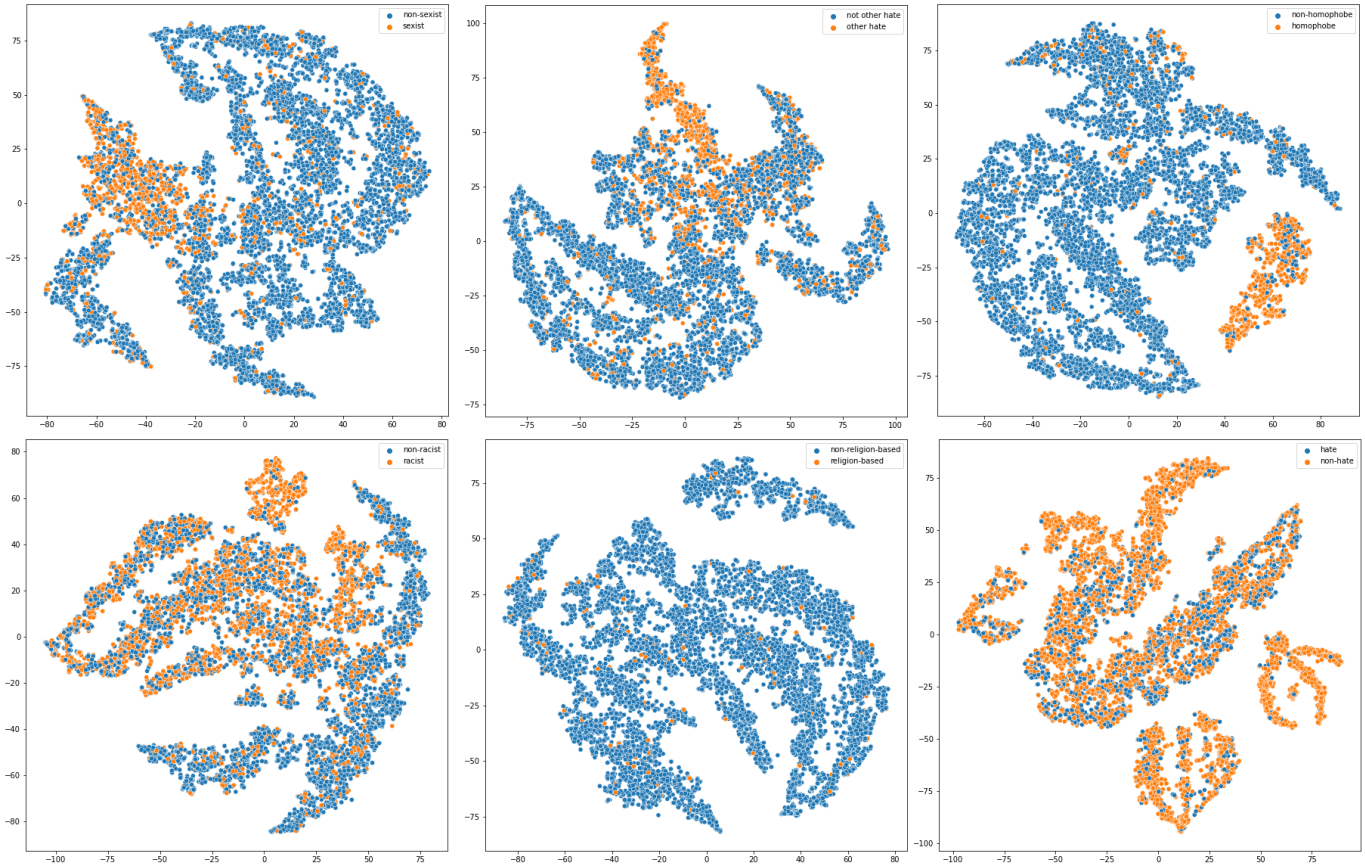

Fig. 4. t-SNE vector representations learnt by contrastive pre-training

to distinguish between Religion-based and Non religious hate due to the subjectivity and less number of samples.

VI. Conclusion

We introduced ConOffense, a multimodal framework to tackle the problem of offensive content detection on raw social media data. ConOffense is trained with a novel multitask contrastive learning approach utilizing both visual and textual information fused together. Experimental results validate our approach with performance improvement over all metrics compared to the previous state-of-the-art on MMHS150K dataset. In the future, we aim to experiment with more sophisticated model architectures trained with our framework in an attempt to boost performance.

References

[1] J. Banks, "Regulating hate speech online," *International Review of Law, Computers & Technology*, vol. 24, no. 3, pp. 233–239, 2010.

[2] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10, 2017.

[3] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: a review on obstacles and solutions," *PeerJ Computer Science*, vol. 7, p. e598, 2021.

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[6] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in indonesian twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, pp. 46–57, 2019.

[7] I. Vogel and M. Meghana, "Profiling hate speech spreaders on twitter: Svm vs. bi-lstm," in *CLEF*, 2021.

[8] K.-L. Chiu and R. Alexander, "Detecting hate speech with gpt-3," *arXiv preprint arXiv:2103.12407*, 2021.

[9] X. Zhou, Y. Yong, X. Fan, G. Ren, Y. Song, Y. Diao, L. Yang, and H. Lin, "Hate speech detection based on sentiment knowledge sharing," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7158–7166, 2021.

[10] Y. Lee, S. Yoon, and K. Jung, "Comparative studies of detecting abusive language on twitter," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 101–106, 2018.

[11] T. Wullach, A. Adler, and E. Minkov, "Towards hate speech detection at large via deep generative modeling," *IEEE Internet Computing*, vol. 25, no. 2, pp. 48–57, 2020.

[12] I. A. Farha and W. Magdy, "Multitask learning for arabic offensive language and hate-speech detection," in *Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection*, pp. 86–90, 2020.

[13] A. Pelicon, R. Shekhar, M. Martinc, B. Škrlj, M. Purver, and S. Pollak, "Zero-shot cross-lingual content filtering: Offensive language and hate speech detection," in *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pp. 30–34, 2021.

[14] M. Das, P. Saha, R. Dutt, P. Goyal, A. Mukherjee, and B. Mathew, "You

too brutus! trapping hateful users in social media: Challenges, solutions & insights," in *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pp. 79–89, 2021.

[15] D. Sarkar, M. Zampieri, T. Ranasinghe, and A. Ororbia, "Fbert: A neural transformer for identifying offensive content," *arXiv preprint arXiv:2109.05074*, 2021.

[16] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1470–1478, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[18] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *arXiv preprint arXiv:2010.01028*, 2020.