Hate Speech Classification in Text-Embedded Images: Integrating Ontology, Contextual Semantics, and Vision-Language Representations

Surendrabikram Thapa[1], Surabhi Adhikari[2], Imran Razzak[3], Roy Ka-Wei Lee[4], and Usman Naseem[5]

[1] Virginia Tech, Blacksburg, VA, USA surendrabikram@vt.edu
[2] Columbia University, New York, NY, USA
[3] University of New South Wales, Sydney, NSW, Australia
[4] Singapore University of Technology and Design, Singapore
[5] Macquarie University, Sydney, NSW, Australia

Abstract. The growing influence of text-embedded images in online communication demands effective strategies for identifying hate speech. The use of hate speech in different contexts makes it necessary to study it in a particular context. Simultaneously, identifying hate speech targets is a crucial research domain as it can offer insights into propagation, impacts, and potential interventions against hate speech. In this article, we address the problem of hate speech detection and target identification in text-embedded images by presenting a comprehensive approach that combines textual and visual cues to accurately detect hate speech and targets within the context of the Russia-Ukraine Crisis. Leveraging a dataset of 4,723 text-embedded images centered around this crisis, we integrate features from the knowledge graph, ontological insights to indicate the presence of hate speech presence, TF-IDF, Named Entity Recognition (NER), and robust vision-language representations. We also provide the rationale behind using different features in our implementation. Our method surpasses existing baselines and methodologies, suggesting the importance of each feature we employ in decision-making.

Keywords: Hate Speech · Multimodal analysis · Vision-Language

1 Introduction

Social media platforms have revolutionized the way we share and consume information, fostering unprecedented connectivity and enabling us to engage with diverse perspectives from around the world. The power of these platforms to rapidly disseminate content has reshaped the way we communicate, learn, and form opinions. However, alongside their undeniable benefits, there exists a flip side to this technological advancement – the issue of hate speech. Hate speech, characterized by harmful, discriminatory, or prejudiced language, poses a significant challenge to maintaining a safe and respectful online environment [1]. In recent years, higher internet penetration and the prevalence of multimodal content have introduced new dimensions to online communication. This fusion of text and images, often seen in the form of text-embedded images, while enhancing expression, has also given rise to challenges, notably the proliferation of hate speech within such medium. These text-embedded images, easily shareable across social media platforms, possess the potential to rapidly disseminate toxic ideologies to a global audience, necessitating urgent and innovative strategies to address this critical issue. However, the scale and pace of content sharing makes manual monitoring insufficient. Hence, automated techniques are necessary to immediately detect and counter hate speech in text-embedded images. These solutions alleviate human moderators' load while fostering a safer online environment conducive to positive interactions [1].

Hate speech becomes especially dangerous amid political tensions between nations, and it's even more concerning during times of invasion [1]. Such speech can fuel division and escalate unwanted situations. To tackle this, research is crucial, specifically to identify hate speech in these contexts and determine its targets. In this research, we propose models for hate speech classification and target identification particularly in the context of Russia-Ukraine crisis.

To build a robust classification system, we exploit the full potential of available textual and visual information within text-embedded content. Our approach is grounded in leveraging both textual and visual dimensions, wherein we employ multimodal-information embeddings from CLIP [18] to jointly understand vision and linguistic concepts. To further enhance our understanding, we incorporate insights from knowledge graphs and ontological data, enriching our analysis with contextual depth. The identification of specific targets is also facilitated by NER, enabling us to pinpoint specific entities involved. In parallel, we integrate traditional features like TF-IDF, for word significance. Through this amalgamation, our model aims to offer a robust solution to moderating hate speech.

2 Related Works

Recently, the efforts to detect instances of hate speech within social media have gained considerable attention, predominantly focusing on textual content. However, efforts dedicated to the classification of hate speech within text-embedded images, a significant aspect of current social media communication, remain relatively constrained. In recent times, a rise in academic interest can be seen, particularly concerning the identification of hate speech within memes or images featuring embedded text. Memes, characterized by their amalgamation of images and text, often intended for humor, have emerged as a popular area of exploration. Moreover, the category of text-embedded images extends beyond memes, encompassing an array of textual-visual content forms, including screenshots extracted from television headlines. In these instances, images provide contextual foundations, complemented by accompanying text that conveys information within that established context. While the study of memes has recently gained attention in academic and industry research, the nuanced examination of hate speech within text-embedded images demands equal scholarly consideration.

In addition to the necessity of conducting further research on text-embedded images, there is a pressing need to conduct research in specific contexts and applications. For instance, there exists a pressing demand for studies focused on hate speech within specific contexts like invasion. Also, while the exploration of memes and other forms of multimodal textual-visual content has largely been concentrated within the broader landscape of general social media platforms, the efforts to curate dedicated datasets and undertake research tailored to those distinct contexts remain relatively limited. Recent efforts, however, have shown promising steps in understanding multimodal textual-visual data within specific domains. For example, Pramanick et al. [17] studied harmful memes and their targets during the US election pandemic with curation of the related dataset that involved labeling US election-related memes to indicate harmful content and identifying the specific targets of these harmful memes. Similarly, Naseem et al. [16] introduced a dataset encompassing 10,244 memes that critique vaccines. They also proposed specific models to capture the context within such datasets.

2.1 Feature Extraction for Hate Speech

In the context of multimodal classification tasks, features play an important role [3]. Word references and lexicons serve as straightforward methods for feature extraction in text analysis. Tokenization is a foundational step in both traditional and deep learning models, often using dictionaries/lexicons [14]. Meanwhile, TF-IDF, a widely used technique, assigns importance to terms based on their significance across documents [24]. Features play a critical role in machine learning, hence requiring more robust features to help in better classification. In recent literature, it has been found that TF-IDF is widely used in the analysis of hate speech in the internet. Recent studies demonstrate that integrating knowledge enhances NLP task performance by enriching models with semantic information. In context of hate speech, Maheshappa et al. [13] showed that incorporating insights from knowledge graph helps to improve hate speech detection. In the realm of hate speech detection in text-embedded images, multimodal models like CLIP [18], GroupViT [25] has shown promising directions [1]. Similarly, the incorporation of WordNet information [6] is significant in the classification of hate speech. Furthermore, NER features have played a pivotal role in hate speech and target classification [15, 5].

In our implementation, for precise identification of hate speech pertaining to the Russia-Ukraine crisis, we use wordnet features along with information about the presence of hate speech. We also leverage features from TF-IDF and NER along with the vision-language representation offered by the CLIP model. Thus, our work bridges a crucial research void by providing a tailored approach to hate speech detection in the specific landscape of the Russia-Ukraine crisis.

3 Dataset

Since our objective is to classify between the hate speech and non-hate speech dataset in the context of Russia-Ukraine crisis, we use CrisisHateMM, the dataset prepared by Bhandari et al. [1] on our implementation. The dataset comprises 4,723 text-embedded images, focusing on the Russia-Ukraine crisis. The dataset had nearly an equal ratio of hate and non-hate speech data with 2,058 images containing no instances of hate speech, while the remaining 2,665 exhibited elements of hate speech. Within this subset of 2,665 images with hate speech, 2,428 specifically had instances of targeted hate speech. The targeted hate speech means that the hate was targeted at some individual, organization, or community. For our first task of hate speech identification, we took all images in the dataset whereas for the second task of target identification, we only took 2,428 images with directed hate speech. We only took text-embedded images with directed hate speech for the identification of targets of hate speech.

4 Methodology

Given an image, **I**, the objective is to determine whether the image contains hate speech content. Additionally, the goal is also to identify whether the hateful image targets one of the specified categories, namely, individual, organization, or community. Figure 1 illustrates an overview of the proposed method.

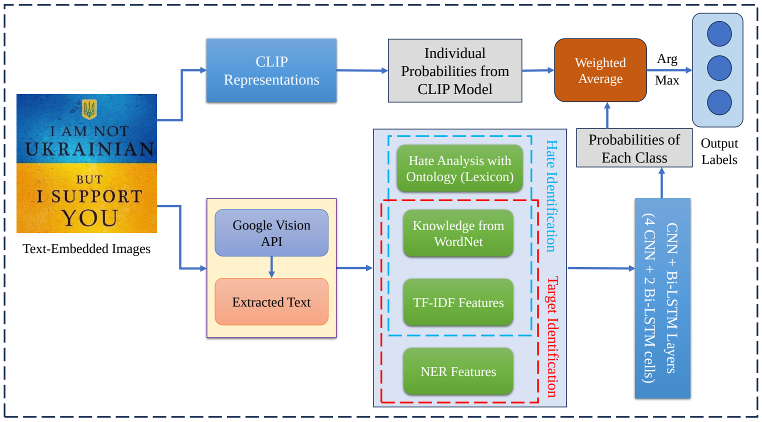

Fig. 1. Our proposed framework leveraging various vision-language representations.

4.1 Textual Features

The textual features from the text-embedded images were extracted using Google Vision API. In this section, we outline the textual features integrated into our methodology. These features collectively contribute to our analysis aimed at accurate hate speech and its target detection. It is important to note that we perform standard preprocessing techniques before calculating the features.

Ontological Insights for Presence of Hate Speech Leveraging ontology is a significant aspect of our approach. To harness the power of ontology, we utilize the expanded lexicon of abusive words provided by Wiegand et al. [23]. This lexicon comprises an expanded list of 8,478 words, encompassing abusive aspects of a language. Among these, 2,989 words are classified as abusive, while the remaining 5,489 words are categorized as non-abusive. Our utilization of this lexicon involves analyzing text from text-embedded images to investigate their content. For each text-embedded image, we represent the ontological information using a two-dimensional vector. The first value denotes the number of words within the text in image that match the abusive lexicon, and the second value represents the number of words that match with the non-abusive lexicon. This vectorized representation serves as a representation to uncover patterns and relationships in the data, enabling us to effectively recognize and classify the content within text-embedded images. We used this information only for the classification of hate speech and did not use this feature in the target identification model.

TF-IDF Features TF-IDF features are integral components of our approach. As shown in equation 1, it considers both the frequency of a term within a document (term frequency) and its scarcity across the entire corpus (inverse document frequency). This helps in highlighting significant terms that can aid in hate speech identification. Bhandari et al. [1] showed that, in context of Russia-Ukraine crisis, there are some words in hate speech which are more significant than the others. TF-IDF helps to leverage this information.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{1}$$

where, $\text{TF}(t, d)$ represents the Term Frequency of term t in document d, which measures how often term t appears in document d. $\text{IDF}(t, D)$ represents the Inverse Document Frequency of term t across the entire corpus D, which measures how unique or rare the term is across the corpus.

Leveraging Knowledge Graph Features We harness the power of the WordNet knowledge graph [6] to enhance our approach. Specifically, we used the hyponymy and hypernymy relationships within WordNet. This enables us to establish connections between specific instances (hyponyms) and their more general concepts (hypernyms), aiding in a deeper understanding of semantic nuances. To identify the top relevant keywords in hate speech, non-hate speech, and specific targets, we employ the SAGE topic model [4]. For hate speech classification, we select 16 keywords, encompassing 8 nonoverlapping words in each category that hold greater relevance in hate speech and non-hate speech posts. Likewise, in target identification, we identify 21 words, including the top 7 non-overlapping words in each target class. In our approach, for each word, we determine the hyponyms and hypernyms, capped at a maximum of 10 per word. By aggregating the resulting hyponyms and hypernyms, we create a combined dictionary. Subsequently, we employ vectorization techniques, essentially assessing whether the text within the text-embedded images contains words from this combined dictionary. This helps to reveal underlying patterns and semantic associations.

Named Entity Recognition (NER) In our target classification process, we also capitalize on NER features. In the context of NER, our approach focuses on leveraging specific entity categories, namely NORP (Nationalities, Religious or Political Groups), PER (Persons), and ORG (Organizations). By recognizing and counting instances of these entities, we extract meaningful information regarding the individuals, groups, and affiliations mentioned within text-embedded images. To achieve this, we employ a counting mechanism for each of the NORP, PER, and ORG entities present within the text. This results in a three-dimensional vector, wherein each dimension corresponds to the count of NORP, PER, and ORG entities, respectively. The resulting vector thus encapsulates the prevalence of these specific entities within the text for reliable target detection.

4.2 Vision-Language Representations

In our methodology, we leverage the powerful capabilities of vision-language representations to enhance our hate speech detection framework. Specifically, we employ the CLIP (Contrastive Language-Image Pretraining) model, which bridges the gap between textual and visual information, facilitating a comprehensive analysis of text-embedded images. The CLIP model possesses the unique ability to understand both images and text in a shared embedding space. It thus provides our framework with a more holistic understanding of the content within text-embedded images. The classifications made by the CLIP model are later incorporated into a weighted ensemble, synergizing with the insights drawn from other textual features. This approach enhances the overall accuracy of hate speech detection and target classification, resulting in a robust and comprehensive solution that aligns with the intricacies of text-embedded image analysis.

4.3 CNN + Bi-LSTM

Convolutional Neural Networks (CNNs) are adept at capturing local text features, while Recurrent Neural Networks (RNNs) excel at capturing long-term dependencies. Combining these architectures can yield improved performance across a spectrum of NLP tasks, including sentiment analysis and text classification [26]. In our experimentation, we employ a model featuring four convolutional layers in series and two Bidirectional Long Short-Term Memory (BiLSTM) cells. In our model design, we feed word embeddings into the convolutional layer. Following every two convolutional layers, we apply max-pooling with a window size of three to capture salient information. To mitigate overfitting, L2 regularization is incorporated into both networks. The activation function used for BiLSTM cells is Tanh. Subsequent to the first BiLSTM cell, we employ batch normalization [20]. The optimization is handled by the Adam optimizer. The output of the BiLSTM cell is connected to a fully connected layer with ReLU as the activation function. As shown in Fig. 1, we provide the model with a combination of features including ontological features, TF-IDF, and knowledge graph insights for hate speech identification. Similarly, for target identification, we integrate TF-IDF, knowledge graph information, and NER features. This comprehensive approach leverages multiple aspects of linguistic and semantic understanding, contributing to our model's ability to accurately detect hate speech instances

Table 1. Performance Comparison of our model different unimodal and multimodal algorithms on CrisisHateMM dataset

Modality	Model	Hate Classification			Target Classification		
		Accuracy ↑	F1-score ↑	MMAE ↓	Accuracy ↑	F1-score ↑	MMAE ↓
Textual	BERT	0.779	0.767	0.240	0.629	0.427	0.998
	DistilBERT	0.754	0.750	0.247	0.637	0.423	1.008
	DistilRoBERTa	0.777	0.769	0.233	0.654	0.440	0.919
	CNN + BiLSTM	0.781	0.773	0.227	0.687	0.619	0.520
Visual	DenseNet-161	0.741	0.739	0.259	0.538	0.425	0.774
	Visformer_small	0.741	0.739	0.257	0.451	0.407	0.772
	MViTv2_base	0.731	0.726	0.276	0.576	0.422	0.657
	VGG19	0.686	0.679	0.305	0.525	0.395	0.785
Multimodal	CLIP	0.798	0.786	0.204	0.684	0.615	0.579
	GroupViT	0.792	0.785	0.214	0.598	0.451	0.763
	MOMENTA	0.787	0.772	0.229	0.688	0.629	0.531
	DisMultiHate	0.771	0.756	0.244	0.624	0.588	0.591
	CogVLM	0.701	0.677	0.349	0.493	0.411	0.740
	MultimodalGPT	0.663	0.610	0.451	0.478	0.383	0.729
	Our Approach	**0.848**	**0.833**	**0.126**	**0.778**	**0.753**	**0.261**

and identify specific targets within text-embedded images. The probability of each classes is taken to form a weighted average and make final decision.

4.4 Weighted Average and Final Predictions

In our methodology's final stages, we employ a weighted average technique to amalgamate the predictions from both the CNN+BiLSTM model and the CLIP model. The weighted average is executed by summing up the predicted probabilities from both models for each class and subsequently dividing the sum by two. Effectively, this process computes the average probability for each class, facilitating a comprehensive and balanced assessment. To ensure the adaptability and effectiveness of the weighted average, we introduce adaptive weights based on the performance of the models on the validation set as shown in equation 2. This approach allows us to dynamically adjust the contributions of each model's predictions based on their respective capabilities and accuracies. Adaptive weighting aims to give more weight to the more accurate model. The weights are normalized so that they sum up to 1. Our train-test-validation split follows a ratio of 70/15/15. The weighted average of probabilities is obtained and the label is assigned to the one with the highest probability.

$$\text{Weighted Average} = \frac{\sum_{i=1}^{n} w_i \cdot \text{Prob}_{\text{Model}_i}}{\sum_{i=1}^{n} w_i} \tag{2}$$

where, n is the number of models (2 for hate speech detection, 3 for target identification). w_i are the adaptive weights for each model based on validation performance. $\text{Prob}_{\text{Model}_i}$ is the predicted probabilities from the i-th model.

5 Experiments

In our experiments, we have performed various experiments with different unimodal and multimodal models.
Unimodal Models: We used the following textual and visual models:

- **Textual Models:** Among textual models, we used BERT [9], DistillBERT [19], DistilRoBERTa [19] and our CNN + BiLSTM model.
- **Visual Models:** In order to assess the performance of unimodal visual models, we used DenseNet-161 [8], Visformer [2], Multiscale Vision Transformers (MViTv2) [11], and VGG-19 [21].

Multimodal Models: These models are important to gauge how jointly learning multiple modalities can help in detection of hate speech and targets. In the case of multimodal models, we use CLIP [18], and GroupViT (Grouping Vision Transformer) [25]. We also experimented with hate speech-specific models such as MOMENTA [17] and DisMultiHate [10]. We also use recent large vision language models (LVLMs) like CogVLM [22], and MultimodalGPT [7].

5.1 Experimental Settings

For baselines, we assessed the model performance by using accuracy, F1-score (macro), and MMAE (Macro Mean Absolute Error) as performance metrics.
Text Preprocessing: The text retrieved from OCR was preprocessed along with the image filtering criteria. We removed non-alphanumeric elements, including special characters, hyperlinks, symbols, and non-English characters that may contribute to noise in the data, which could ultimately distort analysis results. Further, non-English words were removed using the English corpus from the NLTK library [12].

6 Results and Analysis

6.1 Comparison With Baselines

Table 1 shows the overall performance of our proposed model against the state-of-the-art baselines. The performance of our model is higher than all the baselines used in our study. As seen in Table 1, among the unimodal models, the textual models have relatively higher accuracy than the visual models. However, this performance is expected as the memes or text-embedded images have more information within the text than the visual modality alone. Furthermore, it is seen that the non-LVLM multimodal models used in our baselines mostly outperform the unimodal models. This is also as expected as such multimodal models can better understand the context by leveraging both textual and visual information. Among the multimodal and unimodal models used, our proposed model performs the highest with an F1-score of 0.833 in the case of hate speech classification and an F1-score of 0.753 in the case of the targets of hate speech identification. The results show that our method was able to perform the best with a nearly 5-point increase in the F1-score in hate speech classification as compared to the second-best performing model. Our model was able to achieve an F1-score of over 83 percent. Additionally, there was a huge jump in performance in target classification with an increase of nearly 14 points in terms of F1-score. Our model was able to achieve an F1-score of 75.3 percent. This shows that our features along with the weighted average technique yielded better results.

6.2 Ablation Analysis

We conducted an ablation analysis to evaluate the impact of individual components on the performance of our proposed framework for hate speech and target classification tasks. The results are summarized in Table 2.

Table 2. Results of the proposed framework with and without individual components used.

Task	Model	Accuracy	F1-score
Hate Classification	Proposed	**0.848**	**0.833**
	Proposed - Lexicon	0.811	0.809
	Proposed - WordNet	0.820	0.815
	Proposed - TFIDF	0.802	0.788
Target Classification	Proposed	**0.778**	**0.753**
	Proposed - WordNet	0.755	0.732
	Proposed - TFIDF	0.751	0.717
	Proposed - NER	0.703	0.686

For hate speech classification, our proposed framework achieved an accuracy of 0.848 and an F1-score of 0.833, with performance degradation observed when individual components were removed, particularly a decrease of accuracy to 0.811 when excluding lexicon-based features. For target classification, the framework achieved an accuracy of 0.778 and an F1-score of 0.753, with the most significant accuracy decrease to 0.703 observed when excluding NER features, highlighting the importance of these components. These results highlight the effectiveness of incorporating multiple components, including lexicon-based features, semantic knowledge, textual representations, and entity recognition, in our proposed framework for hate speech classification and target classification tasks. Thus, our model benefits from the collective advantages of different components used in our framework.

7 Conclusion

In this paper, we have delved into the critical domain of hate speech identification within text-embedded images, an increasingly influential mode of communication in today's online landscape. Through rigorous experimentation, we show the importance of a diverse set of features, including ontological insights, TF-IDF analysis, knowledge graph utilization, NER entity recognition, and vision-language representations. Our findings demonstrate the potency of combining these features, enhancing the accuracy of hate speech classification and target identification. The adaptive weighted ensemble technique further bolsters our approach, enabling the fusion of diverse models and their outputs, leveraging the strengths of each. While our approach presents promising results, we acknowledge areas for further refinement. In particular, the challenge of identifying satirical yet hateful content remains, suggesting avenues for deeper explorations into nuanced expressions of hate speech. Moreover, our focus on the Russia-Ukraine crisis context prompts consideration of applying our methodology to other contexts and domains, each with its distinct linguistic and visual characteristics. In conclusion, our research fills a critical void in the study of hate speech within text-embedded images, offering a multifaceted methodology that harnesses the synergy of various features and models. As the digital landscape continues to evolve, our findings contribute to fostering a more respectful and inclusive online environment, where hate speech's harmful impact is mitigated, and meaningful interactions prevail.

References

1. Bhandari, A., Shah, S.B., Thapa, S., Naseem, U., Nasim, M.: Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1993–2002 (2023)
2. Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., Tian, Q.: Visformer: The vision-friendly transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 589–598 (2021)
3. Chhabra, A., Vishwakarma, D.K.: A literature survey on multimodal and multilingual automatic hate speech identification. Multimedia Systems pp. 1–28 (2023)
4. Eisenstein, J., Ahmed, A., Xing, E.P.: Sparse additive generative models of text. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 1041–1048 (2011)
5. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y., Belding, E.: Hate lingo: A target-based linguistic analysis of hate speech in social media. In: Proceedings of the international AAAI conference on web and social media. vol. 12 (2018)
6. Fellbaum, C.: Wordnet. In: Theory and applications of ontology: computer applications, pp. 231–243. Springer (2010)
7. Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790 (2023)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
9. Kenton, J.D., Chang, M.W., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
10. Lee, R.K.W., Cao, R., Fan, Z., Jiang, J., Chong, W.H.: Disentangling hate in online memes. In: Proceedings of the 29th ACM international conference on multimedia. pp. 5138–5147 (2021)

11. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
12. Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. arXiv preprint cs/0205028 (2002)
13. Maheshappa, P., Mathew, B., Saha, P.: Using knowledge graphs to improve hate speech detection. In: Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD). pp. 430–430 (2021)
14. Mekki, A., Zribi, I., Ellouze, M., Belguith, L.H.: Tokenization of tunisian arabic: a comparison between three machine learning models. ACM Transactions on Asian and Low-Resource Language Information Processing (2023)
15. Montariol, S., Riabi, A., Seddah, D.: Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022. pp. 347–363 (2022)
16. Naseem, U., Kim, J., Khushi, M., Dunn, A.G.: A multimodal framework for the identification of vaccine critical memes on twitter. In: Proceedings of the 16th ACM International Conference on Web Search and Data Mining. pp. 706–714 (2023)
17. Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M.S., Nakov, P., Chakraborty, T.: Momenta: A multimodal framework for detecting harmful memes and their targets. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4439–4455 (2021)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
20. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? Advances in neural information processing systems **31** (2018)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
23. Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C.: Inducing a lexicon of abusive words–a feature-based approach. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1046–1056 (2018)
24. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) **26**(3), 1–37 (2008)
25. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022)
26. Zhao, N., Gao, H., Wen, X., Li, H.: Combination of convolutional neural network and gated recurrent unit for aspect-based sentiment analysis. IEEE Access **9**, 15561–15569 (2021)