# SSN-NLP-ACE@Multimodal Hate Speech Event Detection 2023: Detection of Hate Speech and Targets using Logistic Regression and SVM

**K Avanthika**
SSN College of Engineering
Tamil Nadu, India
avanthika@ssn.edu.in

**Mrithula KL**
SSN College of Engineering
Tamil Nadu, India
mrithula2010075@ssn.edu.in

**Thenmozhi D**
SSN College of Engineering
Tamil Nadu, India
theni_d@ssn.edu.in

## Abstract

Hate speech has become a noteworthy concern in the digital age owing to its ability to brew violence, spread discrimination, and foster a belligerent atmosphere. Identifying and distinguishing hate speech from harmless discourse on online platforms is essential to maintain a safe and inclusive digital environment.

In this research paper, we propose a multimodal approach to hate speech detection, directed towards the identification of hate speech and its related targets. Our method uses logistic regression and support vector machines (SVMs) to analyse textual content extracted from social media platforms. We exploit natural language processing techniques to preprocess and extract relevant features from textual content, capturing linguistic patterns, sentiment, and contextual information.

These features are fed into logistic regression and SVM classifiers and trained on the labelled dataset. In addition, we performed a comparative analysis to evaluate the effectiveness of the multimodal approach compared to the use of existing methods. The proposed method holds promise for automated hate speech detection systems, facilitating censorship, and proactive intervention to mitigate the harmful effects of hate speech on online platforms.

## 1 Introduction

Hate speech is a form of communication that expresses prejudice, hatred, or discrimination against a specific individual or group based on attributes such as race, ethnicity, religion, nationality, gender, sexual orientation, disability, or other characteristics. It is distinguished by its goal to denigrate, belittle, or encourage violence or harm against persons or groups based on perceived differences or qualities.

In this ever-expanding digital landscape, the emergence and proliferation of hate speech represent an alarming concern. This not only promotes prejudice and division, but it also endangers societal cohesion and individual well-being. As a result, it is now more important than ever to build effective methods for its identification and mitigation.

In this paper, we delve into the crucial area of hate speech detection with a specific focus on identifying not only offensive language but also the intended targets. This dual objective addresses a critical gap in the existing literature, as understanding the context and impact of hate speech requires considering both its content and the entities it targets. To this end, we explore the effectiveness of two powerful machine learning algorithms, logistic regression and support vector machines (SVM), in the field of hate speech detection. These algorithms have a rich history of success in text classification tasks and provide valuable insight into the complexity of hate speech identification.

We present a novel approach to multimodal hate speech event detection, focusing on two SubTasks: Hate Speech Detection and Target Detection. The proposed solutions for these subtasks of Multimodal Hate Speech Event Detection at CASE 2023 (Thapa et al., 2023) has been evaluated with the baseline score presented by the work (Bhandari et al., 2023). We were placed seventh on SubTask A and eighth on SubTask B. For Hate Speech Detection, we employ Support Vector Machines (SVM), while for Target Detection, we utilise Logistic Regression.

The first SubTask, Hate Speech Detection, involves distinguishing between hate speech and non-hate speech textual content. Traditional approaches have primarily relied on textual analysis techniques to identify hateful language. We leverage the SVM model on a diverse dataset comprising labelled instances of hate speech and non-hate speech, enabling the model to learn the underlying patterns and discriminatory characteristics of hate speech.



*Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 66–70, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_009

The second SubTask, Target Detection, aims to identify the specific targets of hate speech within three categories: community, individual, and organisation. This is crucial for understanding the impact and potential harm caused by hate speech instances. By training the Logistic Regression model on labelled data, we enable it to predict the target category for a given hate speech instance accurately.

While the focus of this paper is on textual content analysis using SVM for Hate Speech Detection and Logistic Regression for Target Detection, we acknowledge that visual elements, such as images or videos, can also contribute valuable information in detecting and understanding hate speech events. Future research could explore the integration of visual analysis techniques alongside textual analysis to further enhance the accuracy and robustness of hate speech event detection.

To evaluate the effectiveness of our proposed approach, we conduct comprehensive experiments on a diverse dataset comprising hate speech instances from various domains. By comparing our results with existing state-of-the-art hate speech detection techniques, we establish the competitiveness of our methodology.

Beyond academic contributions, our research holds practical implications for content moderation, social media platforms, and online communities. Appropriate measures can be taken to mitigate the spread of harmful content, protect targeted individuals and communities, and foster a more inclusive and respectful online environment.

The subsequent sections of this research paper namely Methodology and Result and Discussion, will provide detailed explanations of our methodology, including data collection and preprocessing, feature extraction techniques, model development using SVM and Logistic Regression, evaluation procedures, and the interpretation of experimental results. We will also discuss the limitations of our approach and suggest potential avenues for future research in the field of multimodal hate speech event detection.

## 2 Related Work

Flow of information is vital to a society, and now with the advent of social media, the need to process them faster, better and in any form is on the rise. Multimodal learning is a type of learning which uses multiple forms of data such as text, audio and images. The obstacles and challenges are clearly articulated by (Cukurova et al., 2020) and (Karan and Šnajder, 2018). The authors of (Blikstein, 2013) present their insights in learning mainly multimodal learning analytics. The works of (Ngiam et al., 2011) and (Ramachandram and Taylor, 2017) discuss deep learning related to multimodal learning. In particular, the work (Ngiam et al., 2011) deals with cross modality feature learnings.The authors of (Ramachandram and Taylor, 2017) have highlighted methods to fuse learned multimodal representations in deep-learning architectures. The authors of (Srivastava and Salakhutdinov, 2012) have presented their model which uses multimodal learning, and also shown a comparison with other deep learning models.

With the increasing amount of data, identifying hate speech has become an important task. A lot of research has taken place regarding the detection and recognition of hate speech.A survey by authors of (Schmidt and Wiegand, 2017) to recognize hate speech uses natural language processing approach. The authors of the work (Parihar et al., 2021) have explored the state-of-the-art algorithms and prospects of AI in the field of Machine Learning and Natural Language Processing. The work (Poletto et al., 2021) analyzes resources available, and discusses the issues and venues for improvement in the field of hate speech. Hate speech recognition not only concerns a single language, but research on multilingual problems have also been undertaken worldwide. For instance, the authors of (Basile et al., 2019) have taken up the problem of hate speech against immigrants and women in different languages, English and Spanish. This work is also targeted, in the sense, it deals with hate speech against a particular community. The authors of the work (Ousidhoum et al., 2019) have considered multi-aspect multilingual hate speech problem and applied state-of-the-art learning models on their dataset for evaluation.

Research has been conducted in the field of hate speech by many, among those,much lesser in number are those that relates to multimodal learning. The authors of (Kiela et al., 2020) propose a new challenge set for multimodal classification, focusing on detecting hate speech in multimodal memes. The work (Fortuna et al., 2021) also deals with hate speech using multimodal learning. This paper highlights multimodal dataset and models to recognising hate speech and the targets of the directed hate.

| Problem | label | Text-embedded images |
|---|---|---|
| Hate Speech | hate | 2,665 |
| | no hate | 2,058 |
| Target | Individual | 1,027 |
| | Organization | 984 |
| | Community | 417 |

Table 1: Dataset distribution.

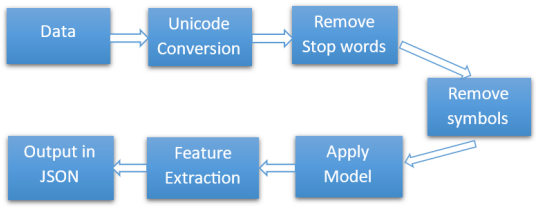


Figure 1: Flow Diagram of the building blocks of the model is shown in the figure

## 3 Dataset and Task

CrisisHateMM (dataset from the task), is multimodal and contains data related to Russia-Ukraine crisis. The work (Bhandari et al., 2023) presented the dataset for the task. It contains social media posts, memes and infographics in the form of text-embedded images which contains some information and context as mentioned above in it.

The first sub task, to detect hate speech, includes a total of 4,723 entries. A hate detected image entry was labelled as 1 and no hate detected entry was labelled as 0.

The second sub task, on the other hand, is to identify the target of hate, which has a total of 2,428 entries. Three different classes were identified as targets namely individual, community and organization. Hate directed towards a individual was labelled as 0, hate directed towards a community as 1 and hate directed towards an organization as 2. The dataset distribution is shown in Table 1.

## 4 Methodology

We present solutions based on classical machine learning models namely SVM and Logistic Regression on this paper. There are advantages to utilising classical ML rather than deep learning. When compared to deep learning models, the need for data is substantially lower, and classical models are often less computationally intensive. In short, classical ML has its own set of benefits, especially when interpretability, data availability, speed or resource constraints are significant factors. The decision to use conventional ML over deep learning was based on the unique situation, data availability, computational resources and the necessity for interpretability and simplicity. In many circumstances, hybrid techniques that include parts of both classical ML and deep learning can be powerful answers.

### 4.1 Preprocessing

Data available may contain noise, missing values or unusable format. Cleaning of raw data helps in the model performance. Preprocessing is an important step which transforms unstructured data to a consistent format, paving way to good working models.

The data given is in the form of text-embedded images. The information from the text-embedded images was collected using Google OCR Vision API.

Textual information obtained from OCR extraction also underwent filtering. This process removed stop words, that is, filtering out words which were considered insignificant. The preprocessing also removed non-alphabetical characters from the text. This included a removal of hyperlinks, symbols and quotes.

Feature extraction was done, to convert the raw data into numeric form. TF-IDF (Term Frequency-Inverse Document Frequency) was used to extract features from the text. It converts a collection of documents to TF-IDF features, which helps in reducing the amount of input features during model building.

### 4.2 SubTask A

Considering the SubTask A is to be a binary classification problem, SVM (Support Vector Machine) was employed.

SVM is a supervised learning algorithm which finds the optimal hyperplane separating the data points of different classes. The hyperplane maximizes the margin between the closest data points from different classes. These data points are the support vectors in finding the optimal hyperplane. This algorithm was chosen owing to its ability to handle both linear and non-linear relationship between the features and the target variables.

We have applied RBF (Radial Basis Function) kernel and tuned the parameters with the objective of maximising the F1-score. The output of the model was converted to JSON format for evaluation.

| Problem | F1 score | Accuracy |
|---|---|---|
| Hate Speech | 76.06 | 76.11 |
| Target | 64.46 | 64.26 |

Table 2: Training Performance.

| Problem | F1 score | Accuracy |
|---|---|---|
| Hate Speech | 78.6 | 79.8 |
| Target | 61.5 | 68.4 |

Table 3: Baseline scores.

### 4.3 SubTask B

The SubTask B is identified to be a classification problem with three classes. Hence we opted for a cost-sensitive logistic regression model.

Logistic Regression is yet another supervised learning technique for classification. It is a statistical analysis method which used probability estimation. Cost-sensitive logistic regression takes misclassification into consideration. This technique was used so as to improve the performance on the imbalanced dataset given. Weights for model building were considered according to the data distribution.

Again, the output generated by model was converted to JSON format for evaluation.

## 5 Result and Discussion

The main evaluation parameter for performance was the F1-score. The training performance parameters of different SubTasks are shown in table 2. On the training dataset, F1-scores of 76% and 64% were obtained in hate speech detection (SubTask A) and target identification (SubTask B) respectively. On the test dataset, our model achieved F1-scores of 78.80% in SubTask A and 52.58% in SubTask B. The details are shown in the table 4. The baseline score from the task paper (Bhandari et al., 2023) are F1 -scores of 78.6% for SubTask A and 61.5% for SubTask B. The table 3 presents the baseline scores.

The SubTask A used SVM for handling complex nonlinear relationships and SubTask B model used cost-sensitive logistic regression to account for misclassification and imbalanced dataset. Our model does not perform better than baseline scores in sub task B. It performed slightly better than random guess, on the other hand, our model was able to improve the score of sub task A from the baseline by a slight margin.

| Problem | F1 score | Accuracy |
|---|---|---|
| Hate Speech | 78.8 | 79.01 |
| Target | 52.58 | 64.05 |

Table 4: Testing Performance.

In any problem, the dataset plays a major role. The imbalance in dataset could be one of the reasons for misclassification. It could also be attributed to the fact that hate directed images themselves might not be directed explicitly, thus making it hard for models to recognise and learn them. Preprocessing the available forms of data always plays a significant role in learning. All the above factors indicate the need for better performing models in the field of multimodal data.

## 6 Conclusion

With the rising need to process data in different forms like opinions and perspectives in social media, identification of hate speech and its targets has become vital. In this paper, we have presented solutions to the task Multimodal Hate Speech Event Detection - CASE 2023. The paper proposes solutions to the task of detecting hate speech in multimodal dataset and identifying the target of the hate as individual,community or organization. The performance metrics includes precision, recall, accuracy with F1-score as the key parameter. Although the results presented herein are good, there remains potential for improvement. Future research can focus on fine-tuning parameters for hate speech recognition. Additional investigation may be undertaken to enhance the performance of existing models and to choose superior models.

## References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. 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 Pat-*

*tern Recognition (CVPR) Workshops*, pages 1993–2002.

Paulo Blikstein. 2013. Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 102–106.

Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 2020. The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology*, 51(5):1441–1449.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing Management*, 58(3):102524.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Nitish Srivastava and Russ R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.