# Luminaries@CASE 2025: Multimodal Hate Speech, Target, Stance and Humor Detection using ALBERT and Classical Models

**Akshay Esackimuthu**
Department of Computer Science and Engineering
Sathyabama Institute of Science And Technology
Chennai - 600119 , Tamil Nadu , India
akshayesackimuthu@gmail.com

## Abstract

In recent years, the detection of harmful and socially impactful content in multimodal online data has emerged as a critical area of research, driven by the increasing prevalence of text-embedded images and memes on social media platforms. These multimodal artifacts serve as powerful vehicles for expressing solidarity, resistance, humor, and sometimes hate, especially within the context of marginalized socio-political movements. To address these challenges, this shared task introduces a comprehensive, fine-grained classification framework consisting of four subtasks: (A) detection of hate speech, (B) identification of hate speech targets, (C) classification of topical stance toward marginalized movements, and (D) detection of intended humor. By focusing on the nuanced interplay between text and image modalities, this task aims to push the boundaries of automated socio-political event understanding and moderation. Using state-of-the-art deep learning and multimodal modeling approaches, this work seeks to enable a more effective detection of complex online phenomena, thus contributing to safer and more inclusive digital environments.

## 1 Introduction

Hate speech detection has become an essential component in fostering a safer and more inclusive digital ecosystem. In today's highly connected world, where social media and online platforms shape public discourse, the rapid dissemination of hateful content can lead to severe social and psychological harm, particularly against marginalized communities. Effectively identifying and mitigating such content not only protects vulnerable groups but also promotes constructive dialogue and reduces the risk of conflict escalation.

Recent advancements in natural language processing (NLP) and computer vision (Parihar et al., 2021) have significantly enhanced the capabilities of hate speech detection systems, particularly in multimodal contexts where images are embedded with textual content. By jointly analyzing both modalities, it is possible to capture subtle nuances, such as sarcasm or implied hostility, that would otherwise be missed in unimodal approaches. This is particularly critical in the context of memes and other visual artifacts commonly used to spread hateful or harmful narratives.

In line with this vision, the shared task introduced in CASE 2025 (Thapa et al., 2025) as part of workshop (Hürriyetoğlu et al., 2025) focuses on the detection of hate speech, identification of targeted entities, stance classification towards marginalized movements, and detection of humor in multimodal social media content. Building upon this framework, our study explores the integration of transformer-based models and classical machine learning techniques to tackle these challenges.This analysis has base references from (Thapa et al., 2024) and (Thapa et al., 2023).

Specifically, we employ the ALBERT base transformer model, known for its parameter efficiency and strong performance in semantic understanding tasks. In addition, we incorporate classical models such as XGBoost, LightGBM, Gradient Boosting, and MLP classifiers, which allow for diverse feature perspectives and robust ensembling strategies. Our approach combines traditional feature engineering (e.g., syntactic and TF-IDF features) with deep contextual embeddings to capture both surface-level and deep semantic cues.

Through weighted ensembling and subtask-specific optimizations, we aim to improve the fine-grained detection of hate speech and its associated attributes, ultimately contributing to more effective content moderation and fostering healthier online interactions.



*Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025*, pages 71–75, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-008

## 2 Dataset & Task Description

### 2.1 Overview

In the evolving digital landscape, text-embedded images, such as memes and infographics, have emerged as powerful tools of expression, particularly in social and political discourse. These images often blend textual and visual cues, creating a complex multimodal environment that challenges traditional content moderation and hate speech detection methods. Within the context of the marginalized movement, such images can serve dual roles: amplifying voices of solidarity and simultaneously perpetuating harmful stereotypes or hostility. The nuanced interplay between humor and offense further complicates moderation efforts, as satire often straddles the delicate boundary between critique and hate.

Recognizing this complexity, the shared task CASE2025 proposes a comprehensive classification framework, focusing on four distinct yet interrelated subtasks: detection of hate speech, identification of hate speech targets, classification of stances toward marginalized movement, and humor detection. The data set used for this study consists of meticulously annotated text-embedded images for each subtask, enabling a detailed exploration of online discourse. The dataset is curated from (Shah et al., 2024) and (Bhandari et al., 2023).The features of the dataset is given in the table 1.

Table 1: Features of the dataset

| Field | Description |
|---|---|
| filename | Name of the file with index value |
| text | Text extracted from text-embedded images |
| label | Ground truth label or category associated with the text/image |

#### 2.1.1 Subtask A: Detection of Hate Speech

The primary objective of this subtask is to determine whether an image contains hateful content. Images are annotated with binary labels: **Hate** and **No Hate**. This binary categorization simplifies initial screening yet serves as a critical foundation for deeper analysis in subsequent subtasks.

| Label | Count |
|---|---|
| No Hate | 2,065 |
| Hate | 1,985 |
| **Total** | **4,050** |

Table 2: Distribution of labels in Subtask A for binary hate speech detection.

#### 2.1.2 Subtask B: Classification of Targets of Hate Speech

For images identified as hateful, the next step is to pinpoint the specific target of hate. The dataset categorizes targets into four classes: **Undirected**, **Individual**, **Community**, and **Organization**. This fine-grained categorization enables a better understanding of hate speech dynamics and the intended victim groups.

| Label | Count |
|---|---|
| Undirected | 617 |
| Individual | 199 |
| Community | 931 |
| Organization | 238 |
| **Total** | **1,985** |

Table 3: Label-wise distribution for Subtask B, focused on hateful images only.

#### 2.1.3 Subtask C: Classification of Topical Stance

This subtask focuses on identifying the stance expressed by the image towards the marginalized movement. Stance classification is crucial for understanding the broader sentiment landscape and distinguishing supportive content from oppositional narratives. The dataset includes three stance labels: **Neutral**, **Support**, and **Oppose**.

| Label | Count |
|---|---|
| Neutral | 1,166 |
| Support | 1,527 |
| Oppose | 1,357 |
| **Total** | **4,050** |

Table 4: Distribution of stances towards the marginalized movement in Subtask C.

#### 2.1.4 Subtask D: Detection of Intended Humor

The final subtask involves determining whether the image is intended to convey humor, sarcasm, or satire. Humor plays a significant role in shaping public perceptions and often acts as a vehicle for veiled hostility. Detecting such elements is essential for nuanced content moderation. The dataset labels images as **Humor** or **No Humor**.

| **Label** | **Count** |
|---|---|
| Humor | 2,737 |
| No Humor | 1,313 |
| **Total** | **4,050** |

Table 5: Distribution of humor-related labels in Subtask D.

## 3 Methodologies Used

### 3.1 Preprocessing

To ensure the textual content extracted from images is clean and analysis-ready, extensive preprocessing steps were implemented:

- Conversion to lowercase to normalize textual patterns.
- Removal of punctuation, stop words, URLs, emojis, and special symbols to minimize noise and irrelevant cues.
- Lemmatization using the NLTK library to reduce words to their base forms, improving semantic understanding.
- Tokenization using built-in mechanisms in TF-IDF and transformer models to prepare the text for vector-based analysis.

### 3.2 Feature Engineering

Several feature engineering strategies were employed to enhance the representational capacity of the text:

- **TF-IDF vectors** for classical machine learning models, capturing term importance and contextual relevance.
- **Syntactic features**, including:
  - Word count, which helps assess verbosity and potential aggressiveness.
  - Stopword ratio, indicating content density.
  - Frequency of punctuation and uppercase letters, often correlated with emotional intensity.
  - Average word length, providing additional stylistic insights.

### 3.3 Models Used

**Transformer-Based Model: ALBERT** The **ALBERT (A Lite BERT) base v2** model was utilized as a primary deep learning approach due to its efficiency and superior performance in text classification tasks. ALBERT leverages self-attention mechanisms to capture complex token relationships, enabling it to understand nuanced semantic and syntactic patterns present in text-embedded images. It was fine-tuned on each subtask-specific labeled dataset, allowing it to adapt to different classification objectives.The flow of the process is shown in Figure 1

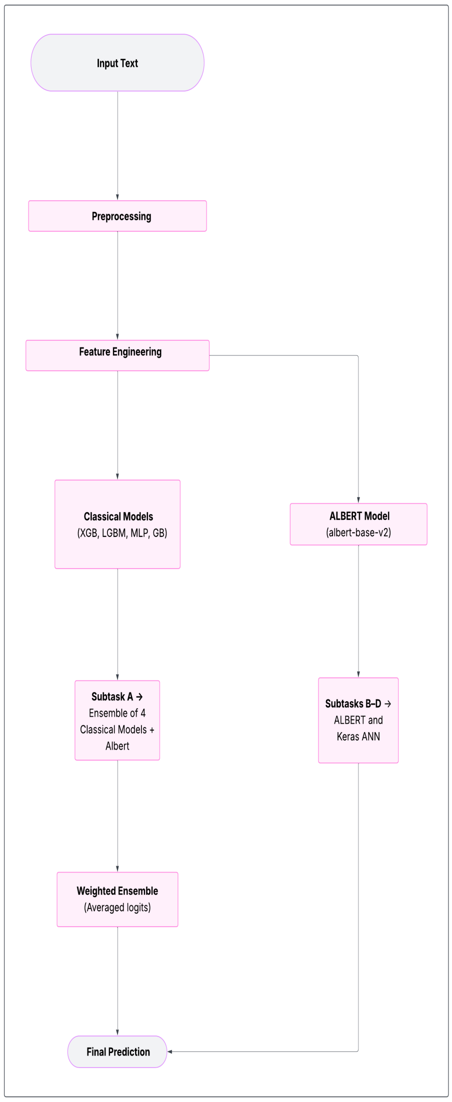


Figure 1: Text-embedded image

## 4 Results & Discussion

This section presents the implementation details and comprehensive analysis of the results obtained for each subtask of the CASE2025 Multimodal Hate Speech Detection Shared Task. The evaluation was carried out using standard metrics such as accuracy, precision, recall, and F1-score, and the results are discussed in depth below.

### 4.1 Subtask A: Hate Speech Detection

For the primary subtask of determining whether a given text contains hate speech, a combination of transformer-based and classical machine learning models was explored. The ALBERT (albert-base-v2) model was fine-tuned using the simple-transformers library, while classical models including XGBoost, LightGBM, GradientBoostingClassifier, and MLPClassifier were trained using TF-IDF and syntactic features. A weighted ensembling approach was adopted to integrate predictions from these models.

The model ensemble achieved an F1-score of **0.7234**, with a recall of **0.7225**, precision of **0.7217**, and accuracy of **0.7219**, securing a competitive rank (14th) on the leaderboard. The results demonstrate that leveraging ensemble strategies can effectively balance the strengths of transformer-based deep representations with classical feature-driven approaches. However, the slight margin for improvement suggests potential benefits from further fine-tuning ensemble weights and incorporating additional linguistic features.

| Metric | Score |
|---|---|
| Recall | 0.7225 |
| Precision | 0.7217 |
| F1-Score | 0.7234 |
| Accuracy | 0.7219 |

Table 6: Subtask A: Hate Speech Detection

### 4.2 Subtask B: Hate Speech Target Identification

In this subtask, the goal was to classify the target of hate speech into four categories: undirected, individual, community, or organization. The ALBERT model was fine-tuned for multiclass classification, and a separate feedforward ANN was developed using Keras Sequential API.

The ALBERT model achieved an F1-score of **0.4984**, with a recall of **0.4869**, precision of **0.5289**, and accuracy of **0.5542**, ranking 6th. These results highlight the inherent challenge of accurately distinguishing nuanced targets within hate speech. While the transformer model effectively captured contextual dependencies, the relatively lower scores compared to subtask A suggest that future work could incorporate more sophisticated target-specific features or additional multimodal cues.

| Metric | Score |
|---|---|
| Recall | 0.4869 |
| Precision | 0.5289 |
| F1-Score | 0.4984 |
| Accuracy | 0.5542 |

Table 7: Subtask B: Target Identification

### 4.3 Subtask C: Stance Classification

The task of stance classification involved categorizing posts as hate-supporting, neutral, or counter-hate. The ALBERT model and a Keras-based ANN were trained independently without ensembling.

The ALBERT model yielded an F1-score of **0.5305**, with a recall of **0.5355**, precision of **0.5434**, and an accuracy of **0.5523**, placing 9th overall. These moderate scores indicate the complexity of stance interpretation, which often depends on subtle linguistic cues and contextual nuances. Integrating additional context-aware features or user-level metadata could potentially enhance performance in future iterations.

| Metric | Score |
|---|---|
| Recall | 0.5355 |
| Precision | 0.5434 |
| F1-Score | 0.5305 |
| Accuracy | 0.5523 |

Table 8: Subtask C: Stance Classification

### 4.4 Subtask D: Humor Detection

In the humor detection subtask, the aim was to determine whether a hateful post contained humorous or sarcastic elements. The ALBERT model and ANN were both trained separately for this binary classification task.

The ALBERT model achieved an F1-score of **0.6070**, recall of **0.6030**, precision of **0.6274**, and accuracy of **0.6844**, resulting in a 15th place ranking. These results underscore the challenge of detecting humor, which is often subjective and culturally dependent. Despite reasonable performance, further improvement could be obtained by integrating multimodal features such as emoji usage, stylistic patterns, or contextual image data.

| Metric | Score |
|---|---|
| Recall | 0.6030 |
| Precision | 0.6274 |
| F1-Score | 0.6070 |
| Accuracy | 0.6844 |

Table 9: Subtask D: Humor Detection

### 4.5 Comparative Analysis

Across all subtasks, the ALBERT (albert-base-v2) model consistently outperformed the ANN-based approaches, demonstrating the strong contextual learning capabilities of transformer architectures. While classical models and ANN methods showed promising trends in certain tasks, they generally lagged behind the fine-tuned transformer in overall performance.

The application of preprocessing techniques such as lemmatization, stopword removal, and syntactic feature engineering contributed significantly to model robustness. Furthermore, the ensembling strategy employed in subtask A highlighted the effectiveness of combining diverse models to improve predictive performance.

## 5 Conclusion

Our approach to the CASE 2025 shared task combined the interpretability of classical machine learning models with the representational power of transformers. Ensembling methods improved performance in hate speech detection (Subtask A), and even single-model approaches worked effectively for the remaining subtasks. Future work includes integrating image features and extending ensemble methods to all subtasks.

## Limitations

- We did not incorporate the image modality or multimodal fusion
- Our ensemble approach was limited to Subtask A due to time and resource constraints.
- We did not explore data augmentation or advanced fusion techniques.

## References

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 Pattern Recognition*, pages 1994–2003.

Ali Hürriyetoğlu, Surendrabikram Thapa, and Hristo and Tanev. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

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

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. pages 17320–17332.

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. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.