**Possible Revisions for ‘A Bag-of-Sounds Approach to Multimodal Hate Speech Detection’**

**Innovative but lacking**

Official Review by Reviewer Qwkg25 Feb 2025, 09:54 (modified: 28 Feb 2025, 09:27)

**Review:**  
Clarity  
The writing is clear and follows a logical progression. Technical concepts, such as Mel spectrograms and machine learning models, are explained concisely. However, some sections, particularly the discussion on performance metrics and limitations, could benefit from a more structured breakdown to improve readability.  
Even though the limitation section in my opinion is already well structured, what we could do to enhance readability would be to introduce subheadings to clearly distinguish different categories of limitations. For example:  
- Methodological Limitations: Discussing the absence of transformer-based models and reliance on statistical machine learning approaches.  
- Sociolinguistic Factors: Addressing challenges related to speaker variation, dialectal influences, and dataset biases. This will provide a clearer framework for readers and improve comprehension.  
Significance  
The study contributes valuable insights into the feasibility of using speech data for hate speech detection, particularly in Dravidian languages. However, its impact is limited by the poor performance of the models in the test phase. The findings suggest that more work is needed to refine the methodology before this approach can be widely adopted.  
If we want to go a bit more in depth into providing possible things to refine in our methodology, even though it may be out of scope for this study, we can possibly include some further discussion in the last paragraph of the discussion section.

Future research could improve model performance through:  
- Hyperparameter tuning: Employing optimization techniques such as GridSearchCV or RandomizedSearchCV to fine-tune parameters for classifiers like Random Forest, SVM, and Logistic Regression.  
- Exploring alternative machine learning models: More advanced boosting algorithms like XGBoost, CatBoost, or LightGBM could improve classification performance, though their application would shift away from a strictly statistical language model approach.  
- Alternative speech feature representations: While this study uses Mel spectrograms, future work could compare results using MFCCs (Mel-Frequency Cepstral Coefficients), which have been effective in speech-based classification tasks.

**Rating:** 7: Good paper, accept

**Confidence:** 4: The reviewer is confident but not absolutely certain that the evaluation is correct

The study aims to detect hate speech in Dravidian languages (Tamil, Malayalam, Kannada, etc.) using deep learning models and compare performance across different approaches. The models used here are random forest, SVM, logistic regression, naive bayes, BERT and XLM-RoBERTa

Official Review by Reviewer VwQM24 Feb 2025, 08:56 (modified: 28 Feb 2025, 09:27)

**Review**:

* dataset description should be before methodology. Some sections are mixed up.  
  This should be a quick fix for just moving things around.
* Move the Limitations Section: The "Limitations" section should be placed before the references rather than after.  
  No need to do this because the limitations section is not part of the main paper and is just something to have in addition to the paper.
* Improve Dataset Description: The number of data points in each language should be explicitly stated in the dataset section  
  This is already covered in table 1 and table 2.
* Future Work Section: Expand this section to include specific directions for improving model performance, dataset expansion, and addressing the study’s limitations.  
  In a Future Work section if we wanted to add one, we could briefly outline potential directions for improving the methodology while keeping the focus on multimodal hate speech detection. This would help maintain the scope of the study while still acknowledging potential improvements without overloading the methodology with speculative enhancements.
* Ensure proper verb agreement and sentence structuring.  
  There are some grammatical errors that may need to be double checked when reading this over again. “with written and spoken social media language data for from YouTube” (line 82). Overall, I didn’t notice any actual grammatical errors.

The current "Limitations" section does not focus on the key challenges. Dataset Issues , Computational Constraints etc.,  
If we want to, we can possibly expand on the sociolinguistic challenges that come with this speech dataset but the principals that are already elaborated on sort of do this and I don’t think it is entirely necessary. If you want, you can expand on this because this area is not something I am too familiar with.

**Rating**: 6: Marginally above acceptance threshold

**Confidence**: 3: The reviewer is fairly confident that the evaluation is correct

A Bag-of-Sounds Approach to Multimodal Hate Speech Detection

Official Review by Reviewer UKZr22 Feb 2025, 21:37 (modified: 28 Feb 2025, 09:27)

**Review**:

Quality: The paper presents a unique and innovative approach to hate speech detection by focusing on speech signals alongside text-based methods. The use of Mel spectrogram transformations is well-justified and aligns with prior audio-based NLP research. The experimental setup is well-structured, but the reliance on traditional machine learning models instead of deep learning approaches (e.g., transformers) limits the study's ability to compete with state-of-the-art methods. Additionally, class imbalance and data preprocessing choices impact model performance, leading to poor test set results.  
We already addressed this in the methodology section.

Clarity: The paper is clearly written with a logical flow from problem definition to methodology and results. The figures, tables, and explanations effectively convey key findings. However, the lack of in-depth analysis on model failures and comparisons with deep learning alternatives makes it difficult to assess the true potential of the proposed approach. A discussion on why the model underperformed on the test set would add more value.  
In the Discussion section, we could clarify why the models underperformed in the test phase and how future refinements may unlock the full potential of this approach. Specifically:

Performance Discrepancy: The decline from validation to test performance suggests possible overfitting to the training set or dataset-related biases.

Feasibility of Speech-Based Detection: Despite lower performance, the study demonstrates that speech data carries valuable information for hate speech detection, and improvements in data representation and model architecture could yield better results in future work.

Broader Implications: The study provides a foundation for future multimodal hate speech detection systems.

**Rating**: 7: Good paper, accept

**Confidence**: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Some extra content could be initialized for the paper

Official Review by Reviewer ksdf14 Feb 2025, 08:24 (modified: 28 Feb 2025, 09:27)

**Review**:

* A figure illustrating the methodology can be introduced for better understanding the system.  
  Pipeline is already straightforward enough and a diagram would take up too much space.
* GPT or ElMo can be considered for model building  
  In the Discussion section, we could briefly mention dedicated speech-based transformer models as a future research direction. While our study intentionally avoids deep learning models to maintain comparability with statistical approaches, future work could explore using models like ElMo. Adding this would demonstrate our awareness of state-of-the-art methods while still maintaining the scope of the paper.

**Rating**: 6: Marginally above acceptance threshold

**Confidence**: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature