MemeScouts@LT-EDI 2026: Asking the Right Questions - Prompted Weak Supervision for Meme Hate Speech Detection

Ivo Bueno[1,3] **Lea Hirlimann**[2,3] **Enkelejda Kasneci**[1,3]

[1]Technical University of Munich [2]LMU Munich
[3]Munich Center for Machine Learning (MCML)
Correspondence: ivo.bueno@tum.de, hirlimann@cis.lmu.de

Abstract

Detecting hate speech in memes is challenging due to their multimodal nature and subtle, culturally grounded cues such as sarcasm and context. While recent vision-language models (VLMs) enable joint reasoning over text and images, end-to-end prompting can be brittle, as a single prediction must resolve target, stance, implicitness, and irony. These challenges are amplified in multilingual settings. We propose a prompted weak supervision (PWS) approach that decomposes meme understanding into targeted, question-based labeling functions with constrained answer options for homophobia and transphobia detection in the `LT-EDI 2026` shared task. Using a quantized `Qwen3-VLM` to extract features by answering targeted questions, our method outperforms direct VLM classification, with substantial gains for Chinese and Hindi, ranking **1st in English**, **2nd in Chinese**, and **3rd in Hindi**. Iterative refinement via error-driven LF expansion and feature pruning reduces redundancy and improves generalization. Our results highlight the effectiveness of prompted weak supervision for multilingual multimodal hate speech detection.[1]

1 Introduction

Hate speech detection remains challenging due to the complexity and subtlety of such content. Unlike explicit abuse, hate speech is often implicit, requiring an understanding of context, intent, speaker-target relations, and whether the content is self-referential or critiques or endorses harmful views (ElSherief et al., 2021; Zsisku et al., 2024).

These challenges are amplified in memes, a multimodal and culturally grounded form of communication. Memes rely on sarcasm, irony, and shared knowledge, where meaning emerges from image-text interaction. Detecting hate in memes therefore requires multimodal reasoning and sensitivity to cultural and linguistic nuances (Bui et al., 2025; Velioglu and Rose, 2020).

Recent advances in large language models (LLMs) and vision-language models (VLMs) offer new opportunities. These models jointly process text and images and achieve strong zero- and few-shot performance. However, direct VLM-prompting for meme classification remains insufficient, particularly in multilingual and culturally diverse settings where subtle cues are difficult to capture with a single prediction. Moreover, fine-tuning incurs substantial data and computational costs.

To address these limitations, we adopt a prompted weak supervision (PWS) approach that decomposes meme understanding into question-based labeling functions with constrained answers, yielding structured, interpretable features instead of a single end-to-end label. Rather than a single prediction, the model produces structured responses capturing aspects of hate speech such as target identification, implicit bias, and stance. These responses are aggregated into features for downstream classification. This framework improves performance and interpretability, through question-level insight into the model's reasoning.

With this in mind, we address the following research questions:

- (RQ1) Can prompted weak supervision improve meme hate speech detection?
- (RQ2) How do language and cultural differences affect model performance?
- (RQ3) What insights into model behavior emerge from analyzing feature importance and labeling function patterns?

[1]The repository is available on GitHub: https://github.com/ivojuniorx4/LT-EDI-Shared-Task-MemeScouts-with-Prompted-Weak-Supervision

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 198–205

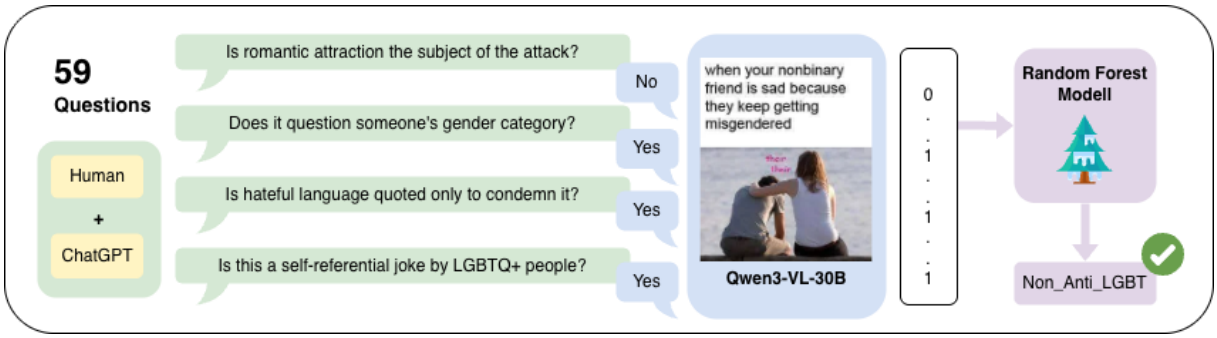

Figure 1: Prompted weak supervision pipeline for homophobia and transphobia detection in memes.

2 Related Work

Detecting hate speech in multimodal data, such as memes, poses a unique challenge, requiring joint reasoning over visual and textual cues as well as human interpretation in context. Both positive and hateful intent in memes are hidden beneath the same layer of irony, sarcasm, and social or cultural references (Velioglu and Rose, 2020). Across languages and cultures, understanding of hate speech in memes varies, as shown by Bui et al. (2025) in their parallel dataset *Multi3Hate*, featuring memes and annotator decisions in five languages. These findings motivate approaches that explicitly represent intermediate judgments (e.g., target, stance, irony) rather than relying on a single end-to-end prediction.

Beyond the multimodal setting, multilingual text remains a challenge for hate speech detection, even for large language models with reasoning capabilities. For harmful content targeting LGBTQ+ communities, slang and culturally specific expressions hinder stable performance across multiple scripts and languages, requiring careful fine-tuning with copious amounts of labeled data (Chan et al., 2024). Detecting homophobia and transphobia in memes thus encompasses these challenges, calling for careful methodological choices that balance performance, interpretability, and computational cost.

Weak supervision combines multiple noisy labeling functions (LFs) to generate training labels. Recent work replaces programmatic LFs (Zhang et al., 2022) with natural language prompts answered by large language models (LLMs), enabling flexible and expressive supervision (Smith et al., 2024). Smith et al. demonstrate that prompted LFs, coupled with a label mapping step, outperforms zero-shot prompting and capture complex heuristics difficult to encode manually. In our setting, prompted LFs are attractive because they can express meme-specific phenomena (e.g., sarcasm reversal, narrator identity) that keyword or surface-form heuristics fail to capture.

A key limitation of prompted LFs is their tendency to correlate due to shared model biases. Su et al. (2023) address this by modeling LF dependencies using prompt representations, and propose pruning and structure learning to reduce redundancy and improve label quality. Because our LFs are answered by a single VLM, correlation and redundancy are expected in the feature space; we therefore include pruning as a central pipeline component and analyze cross-lingual overlap to identify transferable versus language-specific signals.

We adopt this paradigm by designing question-based prompted LFs for multimodal hate speech detection. Unlike prior work, we focus on a multilingual meme setting and emphasize iterative LF refinement and feature selection to improve robustness. Unlike classical weak-supervision pipelines that learn a dedicated label model to aggregate LF votes, we treat prompted LF outputs as structured features consumed by a lightweight supervised classifier. This design suits shared-task settings: it supports rapid LF iteration, maintains question-level interpretability, and leverages labeled data to down-weight unreliable signals.

3 Method

Dataset. We evaluate our approach on the dataset from the Homophobia and Transphobia Meme Classification shared task at `LT-EDI@ACL 2026` (Chakravarthi, 2024). It contains annotated social media memes labeled as *Homophobic*, *Transphobic*, or *Non-Anti-LGBT* in English, Hindi, and Chinese, forming a multilingual benchmark. Class distributions are imbalanced and vary by language, motivating macro-F1 and balanced class weights.

The dataset is split into train/test sets per language, comprising 560/141 (train/test) memes in English, 798/200 in Hindi, and 956/239 in Chinese. Overall, it provides a challenging multimodal and multilingual setting for evaluating robust LGBTQ+ hate speech detection.

Question Generation. As shown in Figure 1 our pipeline begins by constructing questions that serve as LFs for PWS. These questions capture complementary aspects of hate speech, including target identification, explicit and implicit hate, and attack characterization. We constructed 59 initial questions using LLM-assisted drafting followed by manual cleanup, each paired with short answers (see App. A for prompt). Depending on the question, answers are binary, ordinal, or categorical. These questions form the basis for extracting structured signals from memes. App. B lists ten example questions.

Feature Extraction. We employ a `nf4`-quantized version of Hugging Face's `Qwen3` implementation (Yang et al., 2025) [2] to answer the predefined questions. For each meme-LF pair, the model is provided with (i) a system prompt describing the task (see App. C), and (ii) a user prompt containing the meme image, and the LF question. Based on the valid answers per question, responses are mapped to integers (see App. E). Values across all questions are aggregated to form a feature vector per meme for training a simple machine learning model.

Classification. Using `scikit-learn`, a Random Forest classifier was trained on the feature vectors for each language with 500 estimators, balanced class weights, and a 20% validation split. Random Forest is chosen for robustness to correlated LF features and for feature-importance estimates used in our analysis.

Refinement. We iteratively refine the pipeline to improve feature quality and performance. First, we analyze misclassified English validation memes to identify recurring patterns not captured by the existing LFs (e.g., narrator identity). When such patterns are identified, we introduce additional questions and re-run the feature extraction process only for the new LFs, expanding the feature space and coverage of previously unmodeled phenomena. This process, *AddLF*, adds 30 new questions for a total of 89 labeling functions.

[2] `https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct`

Second, we investigate two pruning approaches to select features most relevant to the classification. *F1Prune* greedily removes features one at a time, expanding the removal set whenever validation macro-F1 improves. *ImpPrune* removes the top k least important features from the Random Forest model, where k is chosen based on validation performance gain.

4 Results and Discussion

Table 1 shows the classification performance of the Random Forest variants presented above, along with direct `Qwen3-VLM` classification as single-shot and reasoning baselines (see App. D), and an aggregated *All* setting trained jointly across all languages.

Method	English	Chinese	Hindi	All
Qwen3-VL-30B	0.77	0.32	0.21	0.13
Qwen3-VL-30B(with reas.)	0.67	0.10	0.08	0.10
Base model	**0.85**	0.66	0.64	0.47
AddLF	**0.85**	**0.72**	0.66	0.48
AddLF + F1Prune	0.83	0.69	0.64	**0.49**
AddLF + ImpPrune	**0.85**	**0.72**	**0.67**	0.44

Table 1: Macro-F1 comparison of direct Qwen-VL classification and the trained Random Forest models

All proposed variants outperform the direct VLM baseline. The gap likely reflects the brittleness of single-shot end-to-end prompting for memes; unconstrained 'reasoning' further degrades performance through inconsistent decision paths. Gains are particularly pronounced for Chinese and Hindi, and are also reflected in the aggregated *All* column, explicitly answering RQ1: prompted weak supervision improves LGBTQ+ hate speech detection in memes. Our system ranks 1st for English, 2nd for Chinese, and 3rd for Hindi in the shared task.[3] Importantly, `AddLF` never impaired performance and improved results for Chinese and Hindi.

Pruning Success. *F1Prune* improves validation performance, but generalizes poorly, likely due to its local optimization strategy. However, it achieves the highest overall Macro-F1 (0.49) in the *All* setting, suggesting improved cross-lingual balance performance despite weaker per-language scores.

ImpPrune achieves the best per-language performance: English and Chinese remain unchanged

[3] Link to `LT-EDI` shared task ranking: `https://drive.google.com/file/d/18JdEfXCDfPQBrNqCi7S7-QfZL7Ln-WXU/view`

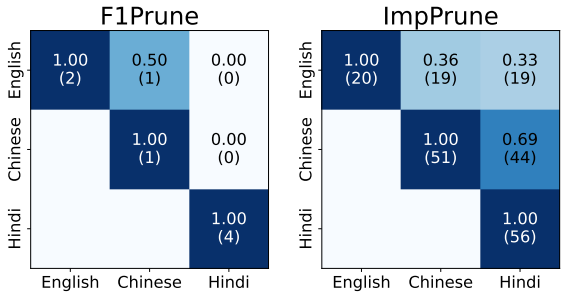

Figure 2: Jaccard similarity among features selected for removal. Values in parentheses indicate the number of shared removed features.

relative to *AddLF*, Hindi improves despite a substantial feature reduction from 89 to 33, suggesting that many LFs could be noisy or redundant, highlighting the importance of effective pruning.

To analyze cross-lingual behavior, Figure 2 reports the Jaccard similarity between removed feature sets. *F1Prune* removes only a few, language-specific features, resulting in near-zero Jaccard similarity across languages. This indicates that the features differ substantially across languages. In contrast, *ImpPrune* removes larger numbers of features, leading to more informative overlap patterns. Nearly all features pruned for English are also removed for Chinese and Hindi, and Chinese and Hindi share 44 pruned features, corresponding to a Jaccard similarity of 0.689, suggesting the presence of language-agnostic uninformative or misleading signals. These findings directly address RQ2, highlighting both language-specific effects and the presence of language-agnostic, weak signals. One possible explanation is that the English LFs may reflect a predominantly Western perspective on homophobia, transphobia, queer language, and memes, limiting their effectiveness across languages.

Selective Propagation of Biased Patterns. While both the LFs and the VLM may introduce biases, the classifier operates purely on numerical vectors and is therefore blind towards the intent captured in both the LFs and their textual answers, incorporating features based on predictive utility. Although some upstream biases are not automatically mitigated, their informative patterns can be repurposed regardless of previous intent. For example, if a model overly sensitive to a LF probing for "attacks" at any reference to homosexuality, the RF might still use the response to distinguish homosexual from transgender memes.

Feature Signal Analysis. Figure 3 visualizes the UMAP projection of Hindi training features colored by RF importance. A small subset of LFs carries most of the weight, including both highly similar patterns (e.g., 78 ("*Is the topic sexual orientation rather than gender identity?*") and 29 ("*Is the joke about sexual orientation rather than gender identity?*") distinguishing the topic between homosexuality/transgender identity), indicating useful redundancy, while opposing signals (e.g., feature 89 ("*Is this meme neutral or unrelated to sexuality or gender?*") on neutral stance) provide complementary information. Similar patterns can be seen for English and Chinese (App. G). For English and Hindi, feature 78 is dominant, while Chinese is led by feature 33 ("*Does it question someone's gender category?*") on the identification of transphobia, reflecting the imbalance in the Chinese dataset. Clusters further reflect semantic themes: the upper-left cluster probes trans identities and stereotypes, whereas the bottom of the figure relates to positive portrayals of LGBTQ+ people. These findings answer RQ3: feature importance and clustering reveal both redundancy and complementarity among LFs, as well as shared and language-specific patterns that shape model behavior.

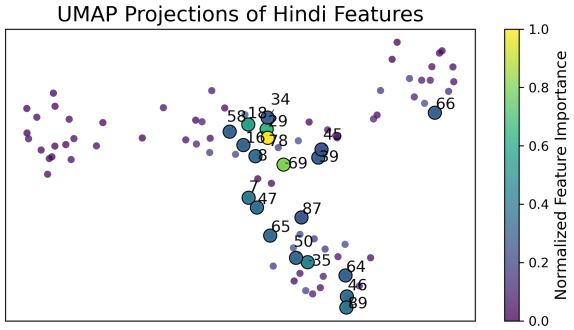

Figure 3: UMAP projection of Hindi question-level features colored by Random Forest importance. Features in the top 20 highest importances are highlighted.

5 Conclusion

In this work, we presented a prompted weak supervision approach for hate speech detection in multilingual memes, leveraging question-based LFs and a VLM for feature extraction. This approach consistently outperforms direct VLM classification, improves interpretability and enables effective feature refinement through pruning.

Limitations

One limitation of our work is that the labeling functions were primarily developed from a Western perspective of what hate speech toward LGBTQ+ people looks like. This may overlook culturally specific markers of homophobia or transphobia present in the two non-Western languages considered in our study.

Furthermore, the VLM was tasked with simultaneously interpreting both the visual content of the meme and the text embedded within the image. Without a dedicated OCR stage, this joint processing may introduce additional errors and potentially affect the model's ability to correctly interpret and label the memes.

Future work could explore culturally adaptive prompting strategies, incorporate explicit OCR pipelines to improve text extraction, and involve more diverse perspectives in the design of labeling functions to improve robustness across languages and contexts. Such directions could enable more accurate and accessible queerphobia detection systems that generalize to additional languages without requiring computationally expensive model retraining.

Ethical Considerations

Our work addresses multilingual meme classification; however, performance disparities across languages highlight potential inequities. The use of English-written labeling functions and a multilingual VLM may disadvantage non-English content, particularly when cultural context is required to interpret meaning. This raises concerns about uneven system performance and the risk of misclassification in underrepresented languages.

AI Writing Assistance Disclosure

AI tools were used for grammar and phrasing assistance, as well as code snippets generation, with all content reviewed and edited by the authors.

References

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Fai Leui Chan, Duke Nguyen, and Aditya Joshi. 2024. "is hate lost in translation?": Evaluation of multilingual LGBTQIA+ hate speech detection. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 146–152, Canberra, Australia. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ryan Smith, Jason A. Fries, Braden Hancock, and Stephen H. Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM / IMS J. Data Sci.*, 1(2).

Jinyan Su, Peilin Yu, Jieyu Zhang, and Stephen H. Bach. 2023. Leveraging large language models for structure learning in prompted weak supervision. In *2023 IEEE International Conference on Big Data (BigData)*, pages 875–884.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *ArXiv*, abs/2012.12975.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A survey on programmatic weak supervision. *Preprint*, arXiv:2202.05433.

Eszter Zsisku, Arkaitz Zubiaga, and Haim Dubossarsky. 2024. Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination. In *Proceedings of the 16th ACM Web Science Conference*, pages 241–249.

A LF Creation Prompt

Figure 4 shows the initial prompt used for ChatGPT assistance on the creation of the questions that composed the labeling functions.

I have potentially homophobic or transphobic memes I want to classify into homophobic, transphobic or not.
I want to use weak prompted supervision.
Generate a list of questions I can use as labeling functions.
The memes are either in English, Hindi or Chinese.
Generate a jsonl dataframe file.
The columns should be: question, acceptable_answers.
In acceptable_answers include any acceptable answers from the models: for example, Yes, YES, yes, and yeah, are some of the acceptable ways to say yes

Figure 4: Prompt for LF creation.

B Labeling Function Examples

Below are ten questions which were part of our 89-questions labeling functions.

- Does this meme express hostility toward LGBTQ+ people?
- Is any group attacked or mocked because of sexual orientation or gender identity?
- Does this meme insult transgender people?
- Does it deny the validity of trans identities?
- Does the meme imply that a group's identity is fake, absurd, or laughable?
- Is the meme criticizing homophobia or transphobia?
- Is the LGBTQ+ person portrayed as the narrator or protagonist rather than the target of ridicule?
- Does the punchline flip the insult back at the speaker instead of attacking LGBTQ+ people?
- Could the content of the image potentially reverse the meaning of the text?
- Does the meme appeal to traditional values, family honor, or social duty to criticize LGBTQ people?

You are a labeling system.
You will be shown a meme image and then asked a single question about it.
Rules:
- Do NOT explain your reasoning.
- Do NOT add extra words, punctuation, or commentary.
- Output only the answer.
- Be concise and deterministic.
- If unsure, choose the closest valid answer.
- The meme may contain English, Hindi, or Chinese text.
- Focus on meaning rather than language.
- Watch for sarcasm or parody.
- If the meme criticizes hateful views, do NOT mark it as hateful.
- Never include explanations.
- Never include multiple answers.

Figure 5: System prompt for feature extraction.

C Feature Extraction System Prompt

Figure 5 shows the system prompt we used to explain the meme classification task during feature extraction.

D Baseline System Prompt

Figures 6 and 7 show the system prompt used run the baseline test without and with external reasoning allowed, respectively.

E Feature Extraction Answer-Integer Mapping

The model is instructed to produce a short, constrained answer (e.g., *yes*/*no* or a small set of categorical options). The generated output is then matched against a predefined list of valid answers. If the output does not match any valid option, the query is repeated, with a maximum of ten retries. In the rare case that no valid response is obtained after all retries, a default fallback answer, which is unused elsewhere in the label space, is assigned to ensure completeness. Table 2 shows the mapping between valid VLM generated answers and a representative integer.

F Number of Features After Refinement

Table 3 shows the total number of features after each step of refinements.

You are a labeling system.
You will be shown a meme image.

Your task is to classify the meme into exactly one of the following categories:
- Homophobia
- Transphobia
- Non_Anti_LGBT

Rules:
- Do NOT explain your reasoning.
- Do NOT add extra words, punctuation, or commentary.
- Output only one of the three labels exactly as written.
- Be concise and deterministic.
- If unsure, choose the closest valid label.
- The meme may contain English, Hindi, or Chinese text.
- Focus on meaning rather than language.
- Watch for sarcasm, irony, or parody.
- If the meme criticizes or mocks homophobia or transphobia, classify it as Non_Anti_LGBT.
- The label should reflect the target and intent of the meme, not just keywords.
- Never include explanations.
- Never include multiple labels.

Figure 6: System prompt baseline classification without allowing external reasoning.

You are a labeling system.
You will be shown a meme image.

Your task is to classify the meme into exactly one of the following categories:
- Homophobia
- Transphobia
- Non_Anti_LGBT

Instructions:
- Carefully analyze the meme step by step.
- Consider text, visuals, context, sarcasm, irony, and intent.
- Explicitly explain your reasoning before giving the final label.

Rules:
- The meme may contain English, Hindi, or Chinese text.
- Focus on meaning rather than language.
- Watch for sarcasm, irony, or parody.
- If the meme criticizes or mocks homophobia or transphobia, classify it as Non_Anti_LGBT.
- The label should reflect the target and intent of the meme, not just keywords.
- If unsure, choose the closest valid label.

Output format (strictly follow this format):
<reason>
Your step-by-step reasoning here.
</reason>
<output>
One label only: Homophobia, Transphobia, or Non_Anti_LGBT
</output>

- Do NOT put the label outside the <output> tags.
- Do NOT include anything outside these tags.
- Do NOT include multiple labels.
- Ensure the final answer appears only inside <output> tags."""

Figure 7: System prompt baseline classification with allowing external reasoning.

Answers	Integer
no, No, NO, nah, n, false, False	0
yes, Yes, YES, yeah, y, true, True	1
0, zero	0
1, one	1
2, two	2
3, three	3
4, four	4
5, five	5
A, a, homophobic, Homophobic, gay people	0
B, b, transphobic, Transphobic, transgender people	1
C, c, neither, Neither, neutral, none, no group	2
sexual orientation, orientation	0
gender identity, gender	1
neither, neutral, none, no target	2
INV (Default)	6

Table 2: Mapping from answer variants to integer representations.

Method	English	Chinese	Hindi	All
Base model	59	59	59	59
AddLF	89	89	89	89
AddLF + F1Prune	87	88	85	80
AddLF + ImpPrune	69	38	33	4

Table 3: Number of features considered after each refinement method.

G Feature Pattern UMAP Projections

Figures 8 and 9 show the UMAP Visualizations of the feature pattern of English and Chinese Memes from the training data respectively. Features in the top 20 highest importances are highlighted.

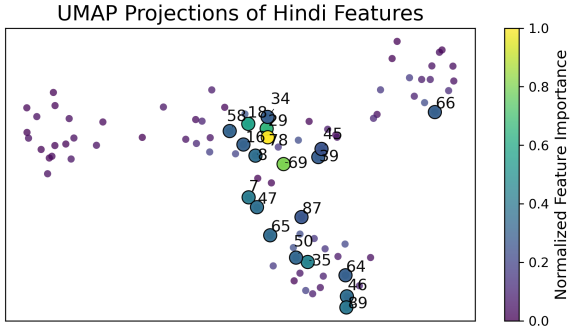

Figure 8: UMAP projection of English question-level features colored by Random Forest importance

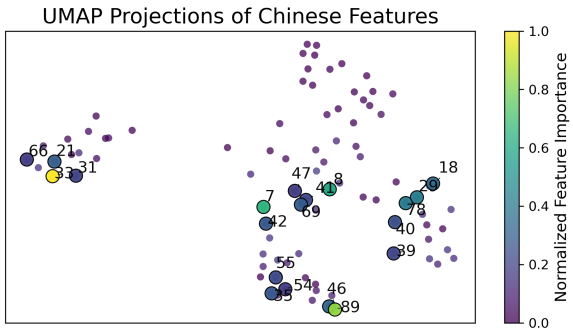

Figure 9: UMAP projection of Chinese question-level features colored by Random Forest importance