A Vision towards Offensive Language Identification: Trends, Insights, and Prospects

Paper ID: 17

Abhisek Sharma, Sarika Jain, Bharat Bhargava

Outline

- Motivation
- State-of-the-art
- Research Gaps
- Contributions
- Research Questions
- Solutions and Research Directions (Thematic Clusters)
- Proposed Framework
- Conclusion & Future Directions

References

Motivation

Offensive discourse: racism, sexism, cyberbullying, radicalization, religious hate.

Growing problem on online platforms: harms individuals & communities.

Current models: lack of context, dataset bias, limited multimodality.

State-of-the-Art

- **Feature Extraction: Earlier with machine-learning models** (TF-IDF, n-grams), **with deep learning** (Word2Vec, contextual embeddings (BERT, mBERT, XLM-R)).
- **Models:** SVM, RF, CNN, LSTM, transformer-based approaches (e.g., BERT).
- **Trend:** shift to pre-trained transformers + hybrid models.

Research Gaps

- Narrow scope: misses subtlety, implicit bias
- Dataset issues: outdated, biased, English-only, Static
- Lack of multimodal dataset and approaches
- Lacks culturally sensitive approaches
- Low explainability → black-box transformers

Contributions of the Paper

- 13 structured research questions (RQ1–RQ13)
- Solutions in 4 thematic clusters
- Proposed Knowledge-infused framework: cultural knowledge graphs + explainability

Research Questions (RQ1–RQ13)

RQ1: What is the scope of automatic offensive language identification?

RQ2: What proportion of work is done per subcategory of offensive language detection?

RQ3: What are the challenges while handling multi-modal data?

RQ4: What is the state of offensive language identification work based on language?

RQ5: What can be introduced in the approaches to make models understand variations in language and terms used while conversing?

RQ6: What is the state of non-English in offensive language identification?

RQ7: What is the state of the datasets that are available in the domain of offensive language identification?

RQ8: How can annotation bias be avoided?

RQ9: What is the state of machine/deep learning models for offensive language identification?

RQ10: How are current systems performing on various datasets?

RQ11: How do knowledge representations like knowledge graphs help in the improvement of offensive language identification?

RQ12: How can the models be made so that they can perform in a contextually and culturally rich manner?

RQ13: What are the venues that evaluate contributions in the domain of offensive language identification?

Solutions and Research Directions (Thematic Clusters)

- Scope & Subcategories (RQ1–RQ2)
- Datasets, Multimodality, Multilinguality (RQ3–RQ8)
- Models & Approaches (RQ9–RQ10)
- Cultural Knowledge & Evaluation (RQ11–RQ13)

Scope & Subcategories (RQ1–RQ2)

Problem:

Existing definitions of offensive language are narrow (e.g., focusing on explicit hate or abuse), missing subtleties such as implicit bias, microaggressions, and culturally specific interpretations.

Solutions:

- Broader taxonomies of offensive language, covering categories like sexism, religious hate, radicalization, cyberbullying, and multimodal toxicity (text + memes, voice).
- Integration of ontology-based approaches to formalize definitions and subcategories.

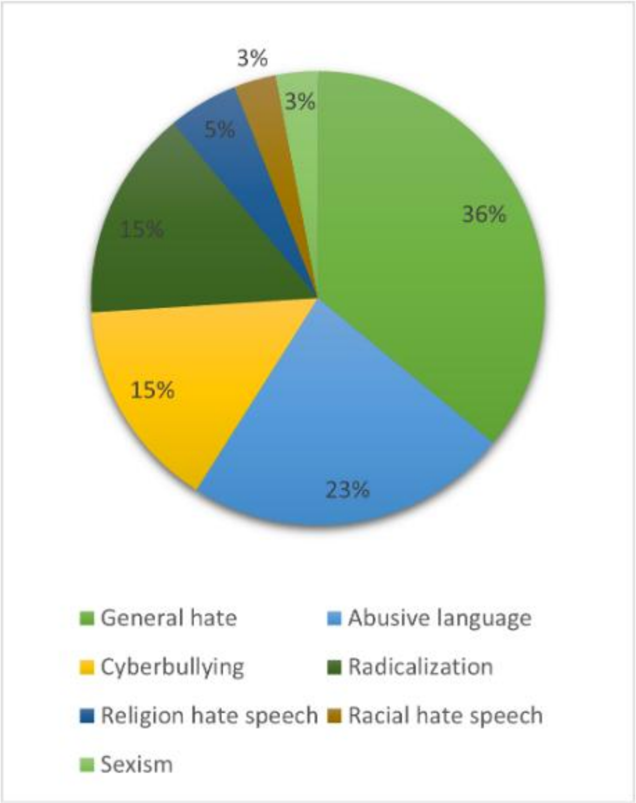


Figure 1: Distinct Sub-tasks under Offensive Language Identification Domain

Datasets, Multimodality, Multilinguality (RQ3–RQ8)

Problem: The learning models are required to understand all sorts of input data apart from the text for bridging the gap between the computer's understanding and human-level understanding of the context. Along with the benefits that the multi-modal datasets bring, there comes a fair share of challenges that pose barriers in development. Rahate, A. et al. [2] have categorized these challenges into six categories as listed here:

- Available multi-modal representations are domain-specific, which limits their use across different tasks.
- Datasets are small in size, contain bias, and are unbalanced.
- Limited multimodal data (text + image/audio/video).
- The current datasets are either missing data about real-life conditions or have noisy representations.
- The missing and noisy dataset sources and the contextually unaware models used impact the interpretability, explainability, and fairness.
- English dominates (≈51\% of research), while low-resource languages (Hindi, Arabic, Tamil, Bengali, etc.) remain underrepresented.

Solutions:

- Dynamic Dataset Updates: Maintain living datasets that adapt to evolving slang and cultural terms.
- Crowdsourced Multilingual Annotation: Leverage diverse annotators to reduce cultural bias.
- Multimodal Corpora: Curate datasets combining text, memes, voice (intonation), and video.
- Bias-Aware Annotation: Consensus-based methods and calibration with cultural guidelines [3].

Models & Approaches (RQ9–RQ10)

Problems:

- Over-reliance on supervised methods (≈73%).
- Limited exploration of semi-supervised and unsupervised approaches.
- Transformer models dominate but remain black boxes (low explainability).

Solutions:

- Hybrid Neuro-Symbolic Architectures: Integrating BERT/transformers with knowledge graphs for context preservation.
- Semi-supervised & self-supervised learning: Reduce dependency on labeled data.
- Explainable AI (XAI): Use attention visualization and post-hoc knowledge graph reasoning to explain classifications.

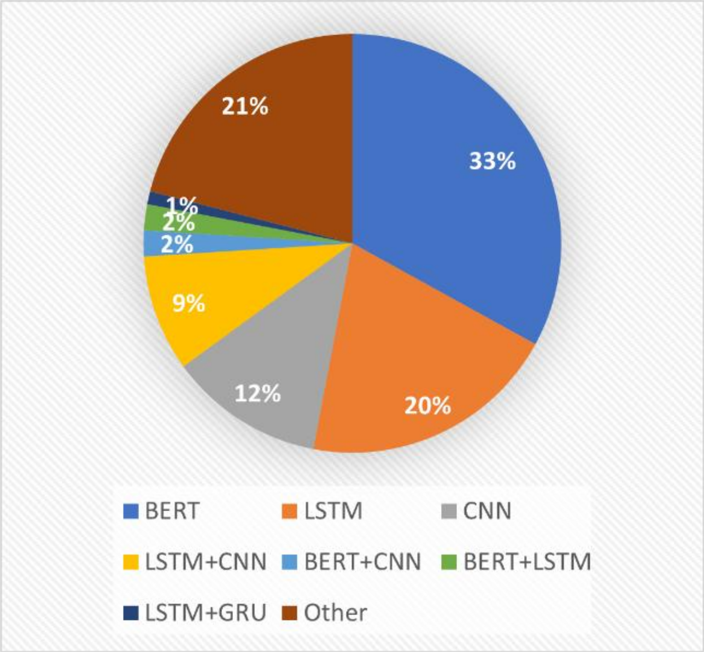


Figure 2: Popular Deep-learning Models in Offensive Language Detection

Cultural Knowledge & Evaluation (RQ11–RQ13)

Problems:

- Offensive language meaning varies by culture and context (e.g., the "N-word" in US vs Africa).
- Lack of structured cultural knowledge graphs.
- Evaluation mostly focuses on datasets/competitions; little emphasis on cultural/contextual performance.

Solutions:

- Cultural Knowledge Graphs: Encode cultural norms and contextual markers.
- Neuro-Symbolic Integration: Train models on cultural knowledge + domain knowledge for contextual awareness.

Cultural Knowledge & Evaluation (RQ11–RQ13) Cont.

Table 2: Summary of benchmark datasets and competitions for offensive language detection.

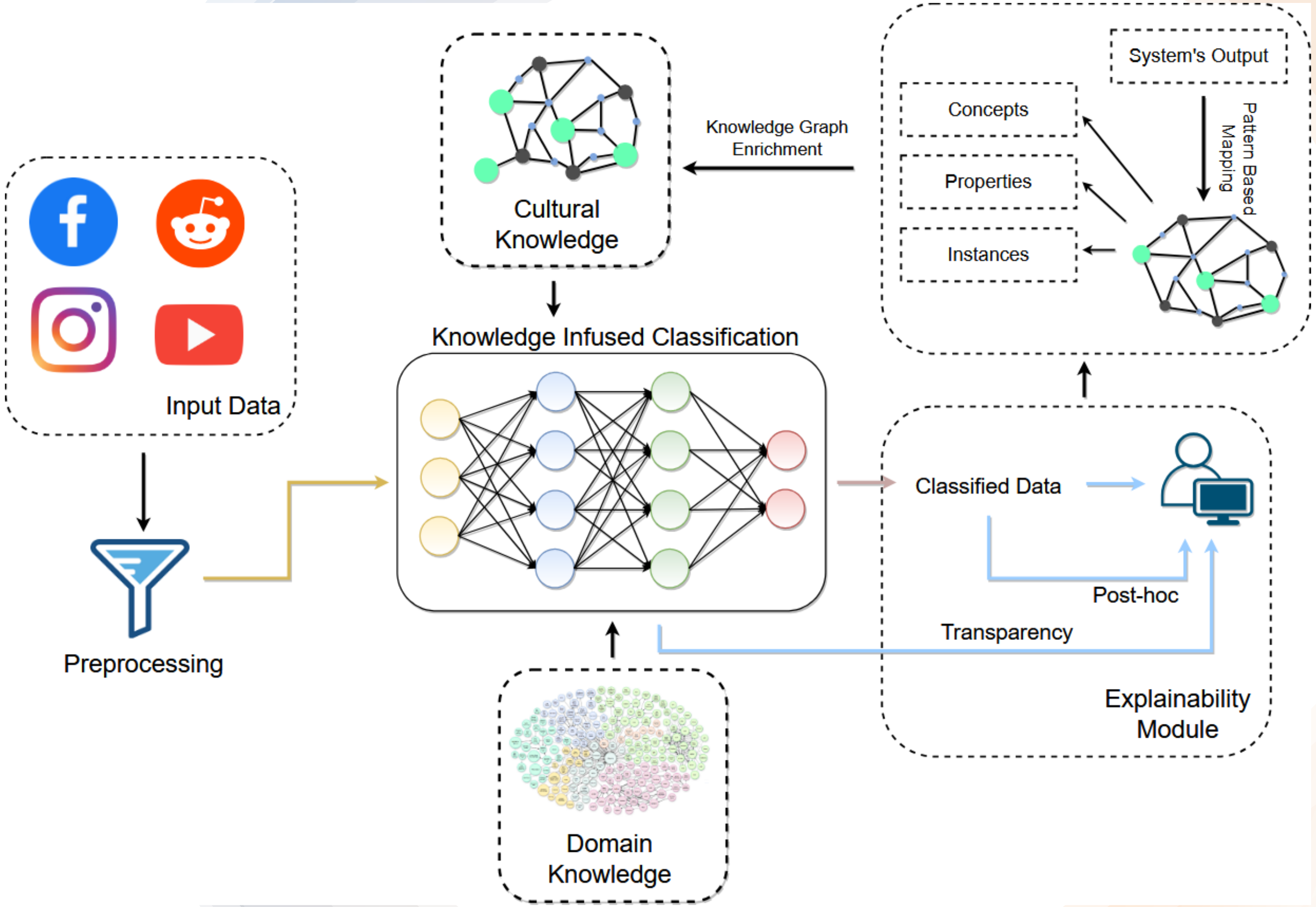
Dataset / Competition	Language(s)	Size	Task / Categories	Limitations
OLID/OffensEval (SemEval-2019 Task 6) (2019-2020) [4]	English	14k tweets	3-level: Offensive/Not, Targeted/Untargeted, Target (Ind./Group/Other)	Small size; Twitter-only; evolving slang not captured
SOLID SemEval-2020 Task 12 [5]	5 languages	∼10M comments from social media (Reddit, YouTube, etc.)	Offensive language detection, hate speech classification, toxicity identification	Highly imbalanced across languages; inconsistent annotation quality across platforms and languages; limited context beyond comment-level
HatEval (SemEval-2019 Task 5) [6]	English, Spanish	19.6k tweets	Hate against women and immigrants	Narrow categories; limited generalizability
HASOC (2019–2024) [7][8]	English, Hindi, German, Tamil, Malayalam	5k–10k per language	Subtask 1: Hate vs Not Offensive; Subtask 2: Hate / Offensive / Profane	Small datasets; annotation bias; imbalance in class distribution
EXIST (2021–present) [9]	Multilingual (English, Spanish)	6k+ social media posts	Broad sexism: explicit and implicit	Focused on sexism only; limited multimodality
GermEval (2019–present) [10]	German	5k–10k per edition	Offensive/Abusive language in German text	Language-specific; not multilingual
Davidson et al. (2017) [11]	English	247k tweets	Hate speech, Offensive, Neither	Crowd-sourced lexicon; overlap across classes; outdated terminology
Founta et al. (2018) [12]	English	80k tweets	Offensive, Abusive, Hateful, Aggressive, Cyberbullying, Spam, Normal	Multi-class, but noisy labels; subjectivity in annotation

Table 2: Mapping of Research Questions (RQs) to Solutions and Future Prospects.

RQ	**Solutions**	**Future Prospects**
RQ1–RQ2 (Scope, Subcategories)	Broader definitions of offensive language; taxonomies covering hate, abuse, cyber-bullying, radicalization, sexism, religious hate.	Comprehensive ontology of offensive language; integration into multilingual benchmarks.
RQ3–RQ4 (Multimodal Challenges, Language Coverage)	Development of multimodal datasets (text, image, audio, video); dynamic dataset updates.	Unified multimodal benchmarks; real-world datasets reflecting cultural and contextual variation.
RQ5, RQ12 (Language Variation, Cultural Awareness)	Use of knowledge graphs to encode cultural/linguistic differences; contextual embeddings.	Large-scale **cultural knowledge graphs**; culturally adaptive NLP models for offensive language.
RQ6–RQ7 (Non-English & Dataset State)	Curated multilingual datasets (HASOC, GermEval, EXIST); crowd-sourced annotation.	Expansion to low-resource languages; automatic cross-lingual offensive language detection.
RQ8 (Annotation Bias)	Multiple annotators, consensus-based labeling; transparent annotation guidelines.	Fair and bias-aware annotation protocols with cultural calibration.
RQ9–RQ10 (Models & Performance)	Adoption of deep learning (BERT, RoBERTa, XLM-RoBERTa); hybrid neuro-symbolic models.	Explainable and robust architectures integrating symbolic (KG) + neural methods.
RQ11 (Knowledge Representation)	Knowledge graphs for contextual understanding and culture-aware classification.	Neuro-symbolic AI mainstream adoption; knowledge-infused transformers.
RQ13 (Venues & Evaluation)	Competitions (SemEval, HASOC, GermEval, EXIST) as testbeds.	Standardized global benchmark tasks including cultural/multimodal subtasks.

Proposed Framework

- **Inputs:** multimodal (text, memes, audio, video)
- **Infusion:** domain + cultural knowledge graphs
- **Classifier:** hybrid neural + symbolic embeddings
- **Explainability:** transparency + post-hoc reasoning
- Continuous enrichment of knowledge graphs



Conclusion & Future Directions

Insights:

- Offensive language is culturally and contextually dynamic
- Static models fail with evolving slang & cultural nuances
- Knowledge-infused learning bridges statistical + contextual reasoning
- Explainability builds trust for real-world deployment

Directions:

- Broader taxonomies and ontology-based definitions
- Dynamic dataset updates, multilingual/multimodal corpora, bias-aware annotation
- Hybrid neuro-symbolic models, semi/self-supervised learning
- Cultural knowledge graphs and context-aware embeddings

The proposed framework presents a high-level overview of a system that incorporates all the proposed solutions into consideration which can be further extended and modified based on problem requirements.

References

[1] Jahan, M.S. and Oussalah, M., 2021. A systematic review of hate speech automatic detection using natural language processing. arXiv preprint arXiv:2106.00742.

[2] Rahate, A., Walambe, R., Ramanna, S. and Kotecha, K., 2022. Multimodal co-learning: challenges, applications with datasets, recent advances and future directions. Information Fusion, 81, pp.203-239.

[3] Rooein, D., Zouhar, V., Nozza, D. and Hovy, D., 2025. Biased Tales: Cultural and Topic Bias in Generating Children's Stories. arXiv preprint arXiv:2509.07908.

[4] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R., 2019. Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.

[5] Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M. and Nakov, P., 2020. SOLID: A large-scale semi-supervised dataset for offensive language identification. arXiv preprint arXiv:2004.14454.

[6] Yang, X., Obadinma, S., Zhao, H., Zhang, Q., Matwin, S. and Zhu, X., 2020. SemEval-2020 task 5: Counter-factual recognition. arXiv preprint arXiv:2008.00563.

[7] Mandl, T., Modha, S., Kumar M, A. and Chakravarthi, B.R., 2020, December. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Forum for information retrieval evaluation (pp. 29-32).

[8] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C. and Patel, A., 2019, December. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th forum for information retrieval evaluation (pp. 14-17).

[9] Plaza, L., Carrillo-de-Albornoz, J., Amig´o, E., Gonzalo, J., Morante, R., Rosso, P., Spina, D., Chulvi, B., Maeso, A. and Ruiz, V., 2024, March. Exist 2024: sexism identification in social networks and memes. In European Conference on Information Retrieval (pp. 498-504). Cham: Springer Nature Switzerland

[10] Risch, J., Stoll, A., Wilms, L. and Wiegand, M., 2021, September. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact claiming comments. In Proceedings of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments (pp. 1-12).

[11] Davidson, T., Warmsley, D., Macy, M. and Weber, I., 2017, May. Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).

[12] Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis,I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M. and Kourtellis, N., 2018, June. Large scale crowd-sourcing and characterization of twitter abusive behavior. In Twelfth International AAAI Conference on Web and Social Media.

Thank you