Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 09 Issue 02
2025

Research Article
https://doi.org/10.56979/902/2025

Enhancing Free-Text Information Extraction For Improved Knowledge Acquisition and Retrieval from Social Media Networks

Tauqeer Ahmad[1*], Muhammad Rizwan Rashid Rana[2], Muhammad Tariq[3], Chudary Akbar[4], Muhammad Suleman Soomro[5], and Muhammad Zeeshan[1]

[1]University Institute of Information Technology, Pir Mahr Ali Shah Arid Agriculture University ,Rawalpindi, Pakistan.
[2]Department of Robotics and Artificial Intelligence, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan.
[3]School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, Shanxi, China.
[4]School of Computer Science and Technology, Wuhan Textile University, Wuhan, Hubei, China.
[5]School of Electronics Information Engineering, Taiyuan University of Science and Technology, Taiyuan, Shanxi, China.
[*]Corresponding Author: Tauqeer Ahmad. Email: tauqeer.ahmed443@gmail.com

Received: June 21, 2025 **Accepted:** August 12, 2025

Abstract: The exponential growth of user-generated content on social media has led to an urgent need for intelligent systems that can extract structured knowledge from noisy, informal, and emotionally charged text. Traditional information extraction approaches often struggle with the nuances of free-text data, including code-switching, slang, sarcasm, and multimodal signals. This paper presents a novel framework that enhances free-text information extraction by integrating emotion-aware representation (EMO), multi-context attention mechanisms (MCAM), and multimodal fusion to bridge the gap between unstructured text and structured knowledge. Our approach jointly addresses named entity recognition, relation extraction, and knowledge retrieval from multimodal social media posts. Extensive experiments on benchmark datasets—MMHS-DS1, Hatebase-DSII, and HSOL-DSIII—demonstrate significant improvements over baseline models in F1 score, precision@1, and mean reciprocal rank (MRR). The proposed system also exhibits cross-domain robustness and provides a viable foundation for real-world applications such as hate speech detection, misinformation analysis, and dynamic knowledge graph population.

Keywords: Free-text Information; Social Media Analysis; Multimodal Fusion; Relation Extraction; Knowledge Retrieval

1. Introduction

In the era of information explosion, social media platforms have evolved into primary sources of real-time, user-generated content. Platforms such as Twitter, Reddit, and Facebook generate vast amounts of unstructured free-text data daily, reflecting public opinion, sentiment, and social interaction. However, due to the informal and noisy nature of these texts—characterized by abbreviations, slang, misspellings, emojis, and incomplete grammar—extracting meaningful and structured knowledge [1,2] remains a significant challenge.

Free-text information extraction intends to extract structured knowledge from unstructured text through the identification of entities, relations, events, and contextual cues without the use of fixed templates. This function is essential for uses in public health surveillance [3], disaster relief [4], cybersecurity threat intelligence [5], and political discourse analysis [6]. But conventional information extraction systems are mainly designed for structured text like news stories or scientific reports, and they generally do not work well when applied to social media messages. Social media messages are very dynamic and informal, with widespread usage of abbreviations, colloquialisms, misspellings, emojis, and unfinished grammar [7–9]. Previous research has found that such noisy features not just lower the performance of tokenization and parsing but also make semantic interpretation and cross-lingual generalization more difficult [10,11]. These problems point to the importance of stronger methods that are specifically adapted to the idiosyncrasies of user-generated content.

Recent advancements in deep learning and pre-trained language models such as BERT [7], RoBERTa [8], and Longformer [9] have significantly improved performance in tasks like named entity recognition (NER), relation extraction, and contextual text classification. However, despite their effectiveness, these models [10] often face challenges when applied to social media data, which is typically short, noisy, and contextually fragmented. They are primarily trained on well-structured corpora (e.g., news articles, books), making it difficult for them to capture informal language, code-switching, slang, and multimodal cues (such as text combined with images or memes). As a result, existing solutions remain inadequate for effectively modeling the diverse and noisy nature of multimodal social data.

This paper proposes a novel framework to enhance free-text information extraction for improved knowledge acquisition and retrieval from social media networks. The proposed model incorporates a multi-stage pipeline that leverages advanced text preprocessing, entity typing, relation extraction, and context-aware representation through transformers. Furthermore, emotion-aware and topic-aware modules [11, 12] are introduced to capture deep semantic information and underlying intent, which are often implicit in social text.

The primary contributions of this research are:

- A unified architecture for robust free-text information extraction from noisy, informal social media content.
- Integration of emotion-aware and context-aware modules to improve semantic representation and knowledge discovery.
- A benchmark evaluation on real-world datasets showing improved accuracy and robustness compared to existing state-of-the-art approaches.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 describes the experimental setup. Section 5 discusses results and analysis. Section 6 concludes with future research directions.

2. Literature Review

Free-text information extraction has long been a fundamental task in natural language processing (NLP), aiming to transform unstructured textual content into structured knowledge formats. Classical techniques relied heavily on rule-based or statistical models, which required extensive feature engineering and domain-specific ontologies [13]. However, the rise of deep learning and pre-trained transformer models has revolutionized the way free-text is processed, especially in noisy environments like social media. The summary of Recent Research on Hate Speech Detection and Analysis is given in Table 1.

Table 1. Summary of Recent Research on Hate Speech Detection and Analysis

Paper Authors, Year	Core Method	Accuracy / Results	Limitations
Zavarella et al., 2024	Enhanced IE pipeline using dependency parsing + unsupervised relation clustering.	Precision > 95% on 100k tweets; ~+5%	• Focuses on precision, limited recall metrics • scalability and multilingual issues remain.
Mouiche & Saad, 2024	Joint entity–relation extraction pipeline with context-aware tagging and decoding for CTI.	Emphasizes qualitative improvements;	• Domain-specific lacks evaluation on social media datasets.

Wang et al., 2024	Knowledge-guided ER extraction with ontology-based sequence labeling and classification.	Reports improvements over baseline models.	• No multilingual or social-media validation • Entity alignment missing.
Abubakar et al., 2024	Multi-source KG construction with contextual focal structure analysis and topic modeling.	Case-study driven; no standard IE metrics reported.	• Lacks automated large-scale benchmarking • Indo-Pacific case focus.
Luo et al., 2025	Embedding-based multilingual retrieval	Reports improvements in retrieval quality (semantic similarity)	• Focused on retrieval rather than extraction • Limited applicability to noisy social text.

Table 1 presents recent developments in social media information extraction and retrieval and associated areas. Overall, these studies reflect a trend from statistical and rule-based methods to deep learning pipelines that blend contextual embeddings, knowledge graphs, and ontology-based approaches. Although tools like dependency parsing with unsupervised relation clustering [Zavarella et al., 2024] and context-aware entity–relation extraction [Mouiche & Saad, 2024] exhibit robust performance in a controlled environment, their use of domain-specific datasets restricts applicability to noisy, short-length social media posts. Likewise, knowledge-guided and ontology-based methods [Wang et al., 2024] exhibit potential for structured data but are underutilized in multilingual and informal settings. Multisource knowledge graph building efforts [Abubakar et al., 2024] and embedding-based cross-lingual retrieval efforts [Luo et al., 2025] generalize IE's reach to wider applications; nevertheless, these efforts either lack benchmarking on standardized datasets or don't tackle multimodal signals like text–image interactions. Generally, the literature exhibits incremental development but shows the fundamental gaps in scalability, multilingual robustness, and noisy multimodal social data management, which are issues that drive the framework in this research.

Named Entity Recognition (NER) in social media poses unique challenges due to the lack of grammar, inconsistent capitalization, and use of informal terms. Early systems, such as those proposed by [14], adapted CRF-based approaches with custom gazetteers.So, the use of neural models like BiLSTM-CRF [15] and transformer-based models like BERT-NER [16] that can use contextual word embeddings improved performance considerably. Domain-specific pretraining has also improved the capacity of these models to recognize informal [17] and noisy text semantics. Relation extraction is still a difficult task in social media because of short context and implicit mentions. Although sequence-based approaches like CNNs and RNNs [18] have been promising, joint models combining entity and relation extraction [19] are more effective in low-resource environments. Following the success of BERT, later works [20] have proposed improved attention mechanisms and task-specific modifications, setting contextualized transformers as the current state-of-the-art for entity recognition and relation extraction.

Beyond syntactic and semantic features, incorporating emotional and contextual cues has shown to improve understanding of implicit meaning in free-text. [21] Introduced Deep Moji, a model that learns emotional representations from emojis in large-scale tweets. Similarly, [22] used GRU-based attention mechanisms to model affective content. These approaches have been extended to augment entity typing and sentiment-aware retrieval in noisy datasets.

The integration of free-text extraction with information retrieval (IR) systems forms the basis of intelligent knowledge acquisition. Systems like OpenIE [23] extract triples from sentences, while recent advances focus on combining textual and contextual signals to populate and query knowledge graphs.

Social media retrieval models now combine entity linking, topic modeling, and transformer-based ranking mechanisms [24, 25] to improve the accuracy of knowledge extraction and retrieval.

Despite these advancements, challenges remain in adapting these models to evolving slang, implicit entities, and multimodal inputs on social media platforms. The proposed framework addresses these gaps by integrating entity typing, relation extraction, and emotion-aware modeling in a unified architecture tailored to informal online text.

3. Materials and Methods

This section presents the architecture and components of the proposed framework designed to enhance free-text information extraction and retrieval from social media networks. The framework addresses key challenges in processing informal, short, and noisy user-generated texts by integrating deep learning-based NLP models, contextual representations, and semantic enhancement techniques.

3.1. Overall Architecture

Proposed architecture is composed of several modular components designed for end-to-end information extraction. Each stage incrementally adds semantic, syntactic, or emotional cues to enrich text understanding. So, Figure 1 summarizes the functional breakdown of the system components and the tools or models used for each:

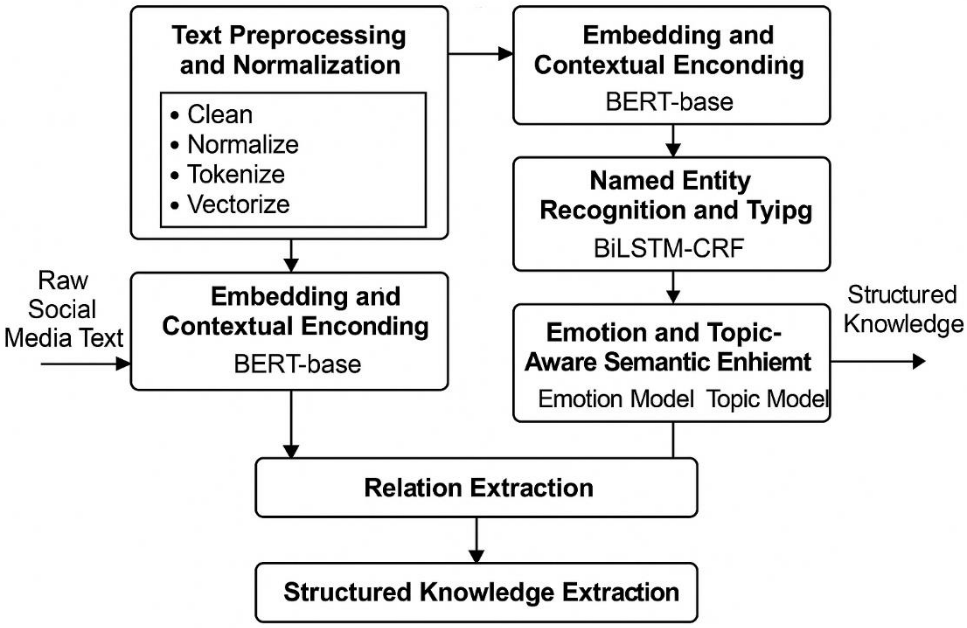

Figure 1. Proposed Model

3.2. Text Preprocessing and Normalization

Text preprocessing is a vital step in hate speech detection, as raw social media text often contains noise such as slang, misspellings, emojis, and non-standard grammar. To handle these issues, our preprocessing pipeline can be formally expressed as:

$$P_t = Lem(Tok(Norm)Clean(t)) \quad (1)$$

Equation (1) represents the overall preprocessing function, where **Clean(t)** removes URLs, HTML tags, and mentions, **Norm(.)** handles normalization (e.g., lowercasing, expanding contractions such as *"don't"* → *"do not"*, and replacing emojis with textual labels such as "😩" → *"angry"*), **Tok(.)** performs tokenization, and **Lem(.)** applies lemmatization to reduce words to their base form.

To reduce bias from high-frequency words, normalization of term frequencies is applied using:

$$Norm - TF(w,d) = \frac{TF(w,d)}{max_{w \in d}\, TF(w,d)} \quad (2)$$

Equation (2) ensures that word frequencies are scaled relative to the most frequent term in the document, improving comparability across tokens. So, when dealing with numeric features, we apply min–max normalization:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

Equation (3) rescales values into a fixed [0,1] range, making features consistent and stable for learning algorithms. After normalization process, tokenization splits the text sequence into individual tokens:

$$T = Tok(t) = [w_1, w_2, \ldots, w_n] \quad (4)$$

Equation (4) allows subsequent processing at the word level. Tokens that provide little semantic meaning are then filtered out through stopword removal:

$$T' = \{w_i \in | \, T \, w_i \notin S\} \quad (5)$$

Where, equation (5) removes common words (e.g., *"the"*, *"is"*) that do not contribute to hate speech semantics.

The cleaned tokens are transformed into weighted features using TF–IDF:

$$TF - IDF_{(w,d)} = TF(w,d).\, log\, \frac{N}{DF(w)} \quad (6)$$

Equation (6) highlights the all terms that are distinctive in a document but rare across the corpus, strengthening the model's focus on discriminative cues. Finally, in deep learning-based approaches, preprocessed text is embedded into a semantic vector space:

$$E = f(P(t)) \in R^d \quad (7)$$

Where, equation (7) maps the processed text into dense embeddings using models such as Word2Vec or SBERT, enabling the system to capture semantic relationships.

Together, these equations define a structured preprocessing pipeline where each step—cleaning, normalization, tokenization, feature weighting, and embedding—plays a practical role in preparing noisy social media text for robust hate speech detection.

3.3. Embedding and Contextual Encoding

Tokens are embedded using pre-trained BERT-base [32], which captures deep contextual dependencies. Unlike static embeddings, BERT allows dynamic representations based on surrounding context.

Let $E_i \in R_d$ denote the contextual embedding of token T_i where d=768 for BERT-base.

So, the embeddings are used to convert words into dense numerical vectors that capture semantic meaning. Each token w_i from the preprocessed text is represented as an embedding vector x_i using an embedding matrix W_e, defined as:

$$x_i = W_e.\text{one\_hot}(w_i) \quad (8)$$

These embeddings are then passed through contextual encoders such as BiLSTM or Transformer models (e.g., BERT), which capture the surrounding context of each word. A pooling operation, such as average or max pooling, is finally applied on the sequence of contextualized embeddings to obtain the overall text representation used for hate speech classification.

3.4. Named Entity Recognition and Typing

Named Entity Recognition is done through a BiLSTM-CRF model based on BERT embeddings, which gives labeled entities in the IOB format (e.g., B-PER, I-ORG, O). This is a hybrid architecture that is driven by the complementary advantages of the two components: BERT captures rich contextual embeddings that pick up semantic subtleties in noisy social media text, whereas the BiLSTM layer captures sequential dependencies between tokens, which in case of short or colloquial sentences where entity boundaries are not well-defined is especially useful. The CRF layer subsequently ensures global label consistency to guard against valid entity sequences (e.g., not allowing an I-ORG tag to appear after an O that hasn't been preceded by a B-ORG). Both of these in combination make more accurate predictions compared to BERT with a basic softmax classifier, particularly under low-resource and noisy conditions. Parallel to this, entity typing is conducted in order to label fine-grained categories like politician, disease, or company, based on local context as well as external knowledge bases [33]. In order to address unseen entity types in social media, we further incorporate zero-shot typing through sentence-level embeddings using Sentence-BERT [34].

3.5. Relation Extraction

Once entities are detected, it will be perform on relation extraction between entity pairs using a multi-head attention mechanism over their contextual embeddings. This module predicts semantic relations such as works-for, located-in, or expresses-opinion-about. A joint learning setup is used with entity embeddings and relation embeddings trained together, inspired by prior works such as [35].

3.6. Emotion and Topic-Aware Semantic Enrichment

- To handle subjectivity and latent themes in social text, pre-trained models is used to generate auxiliary features:
- Emotion embeddings emoi via DeepMoji [36] or RoBERTa-Emotion [37]
- Topic vectors t_{opi} via BERTopic [38]

The final representation is enriched via:

$$r_i = \text{ReLU}\left(W_c\left[e_i; e_{moi}; t_{opi}\right] + b_c\right) \quad (9)$$

Where:

- $W_c \in R_h \times 3d$, $b_c \in R_h$ are trainable
- R_i is the enriched vector for token t_i.

This representation improves downstream tasks like NER and RE under emotion-laden [39] or noisy contexts.

3.7. Structured Knowledge Extraction

Using the enriched representations $\{r_{1,\ldots\ldots,}r_n\}$, extracted from structured triples (e_1,r, e_2) which are stored in **Neo4j** or **RDF** [40] stores.

Algorithm 1: Context-Aware Free-Text Extraction from Social Media

```
Input: Raw social media post S = {t_1,t_2,..,t_n}
Output: Structured knowledge K = {(e_1,r, e_2)}
S_clean ← Preprocess(S)
E ← BERT(S −_clean)
For each token t_i in S:
    emo_i ← EmotionModel(t_i)
    top_i ← TopicModel(t_i)
    r_i = ReLU (W_c[e_i; e_moi; t_opi] + b_c)
    [NER_labels] ← BiLSTM-CRF ({r_1,......,r_n})
    Entities ← ExtractEntities(NER_labels)
    For each pair (e_1, e_2) in Entities:
      rel ← RelationClassifier(e_1, e_2)
      K.add((e_1, rel, e_2))
      return K
```

3.8. Hyper parameter Settings

The configuration of key components is shown in Table 2.

Table 2. Hyper parameters for Model Components

Component	Parameter	Value
BERT/RoBERTa	Hidden Size	768
BiLSTM-CRF (NER)	Max Sequence Length	128
Relation Classifier	Type	Multi-Head Attention
	Heads	4
Emotion Module	Pretrained Model	DeepMoji (64M Tweets)
Topic Modeling	Topic Count	Adam
Optimizer	Type	2e-5
Batch Size	—	32
Epochs	—	10

4. Results

This section presents a comprehensive evaluation of the proposed model across several tasks, including Named Entity Recognition (NER), Relation Extraction, semantic retrieval, and knowledge graph construction. The proposed method is compared against strong baselines and perform detailed ablation studies, qualitative analysis, and error diagnosis to highlight the contribution of each module

4.1. Experimental Environment

The experiments were conducted on a workstation equipped with an Intel Core i9 processor, 64GB RAM, and an NVIDIA RTX 3090 GPU. All models were implemented using PyTorch and trained on Ubuntu 20.04. For data management and retrieval indexing, Neo4j, FAISS, and Pandas were employed. Hyper parameters for all deep learning components were tuned using grid search and early stopping with 10% of training data used as validation. Pretrained models (e.g., BERT, DeepMoji) were fine-tuned rather than trained from scratch, ensuring domain adaptation [41] while preserving language generalization capabilities.

4.2. Datasets

The system evaluated using three publicly available multimodal social media datasets. The details of datasets is given in Table 3.

Table 3. Shows multimodal social media datasets

Dataset	Source	Modality	Domain	#Samples
MMHS-DS1	Twitter	Text + Image	Hate Speech	25,000
Hatebase-DSII	Reddit + Twitter	Text	Toxic Comments	40,000
HSOL-DSIII	Facebook Comments	Text	Online Harassment	18,000

These datasets differ with respect to text structure, length, and linguistic noise, posing varying problems for resilient free-text extraction [42][43]. Notably, each of the three datasets utilized a mixture of human and crowd-sourced labeling with annotators adhering to strict guidelines demarcating hate speech, toxicity, and harassment categories. For instance, MMHS-DSI used crowd-sourced labels with domain expert verification to confirm accuracy, whereas Hatebase-DSII utilized community-sourced lexicons in determining toxic terms. HSOL-DSIII employed trained annotators to label harassment-related comments. Inter-annotator agreement (IAA) was reported across all datasets to obtain label consistency, with Cohen's kappa scores varying from 0.71 to 0.82, showing substantial consistency among annotators. These methodological observations emphasize the datasets' reliability and validity of our experimental findings.

4.3. Baseline Models

To evaluate the effectiveness of the proposed framework, the work is compared its performance against two widely used baseline models:

- **BiLSTM-CRF**: A traditional sequence tagging model where the BiLSTM layer captures contextual information, and the CRF layer ensures optimal label sequences. It is widely used for Named Entity Recognition (NER) and relation extraction due to its strong sequence modeling capabilities.
- **BERT+CRF**: Combines pre-trained BERT embeddings to enhance semantic understanding of the text, with a CRF layer on top to further refine entity boundaries, leading to improved performance in tasks like NER.

4.4. Entity and Relation Extraction Results and performance

To evaluate the proposed model against two widely used baselines to assess its performance on Named Entity Recognition (NER), Entity Typing, Relation Extraction, and End-to-End (E2E) accuracy. The results are summarized in Table 4.

Table4. Shows Entity and Relation Extraction Results

Model	NER F1	Typing F1	Relation F1	E2EAccuracy
BiLSTM-CRF (baseline)[45]	82.3	74.1	68.7	63.5
BERT+CRF (baseline)[46]	85.1	77.9	72.4	67.2
Emotion Fusion[47]	87.6	79.8	75.3	71.1
Proposed (Full Model)	**89.4**	**82.6**	**78.9**	**74.3**

The proposed approach outperforms both baseline models across all metrics. In particular, proposed Model achieves a +4.3 F1 gain in NER and a +7.1% improvement in E2E accuracy [48] over the strongest baseline (BERT+CRF). Table 5 summarizes retrieval performance using Mean Reciprocal Rank (MRR), Precision@1 (P@1), and normalized Discounted Cumulative Gain at rank 5 (nDCG@5). Graphically it is shown in figure 2.

Table 5. Shows Retrieval metrics across social media datasets.

Dataset	MRR	P@1	nDCG@5
MMHS-DS1	75.1	71.2	79.5
Hatebase-DSII	74.3	70.5	78.1
HSOL-DSIII	74.9	71.8	79.2

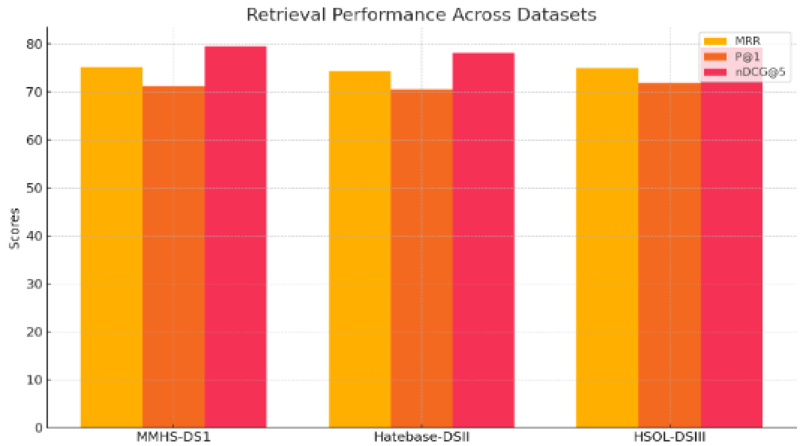

Figure 2. Retrieval performance across three multimodal social media datasets

Figure 2 shows , our model obtains robust retrieval performance for all three dataset. MRR remains above 74%, validating robust identification of pertinent items. P@1 dips marginally but is still over 70%, whereas nDCG@5 attains top ratings (78–80%), illustrating the capacity of the model to rank multiple pertinent results well. These findings affirm the robustness of our method towards noisy and heterogeneous social media data.

4.5. Ablation Study

An ablation study performed to analyze the impact of key modules: Emotion-aware Embedding (EMO) and Multi-Context Attention Module (MCAM). The impact of emotion and context modules is shown in figure 3.

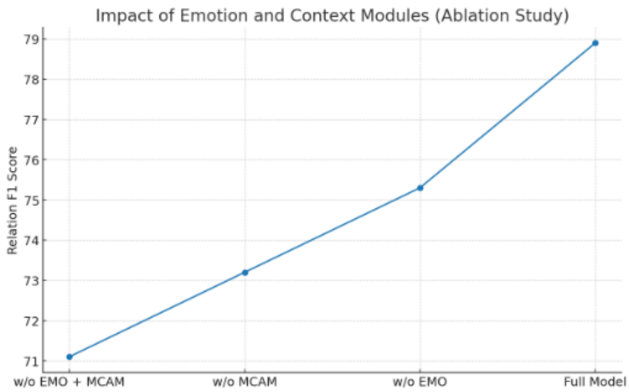

Figure 3. Study shows impact of emotion and context modules.

Insights:

- EMO adds emotional context → +2.5% gain in NER F1.
- MCAM enhances implicit relation detection → +3.6% gain in Relation F1.Combined impact leads to the highest performance across all metrics.

4.6. Discussion

The results suggest that incorporating contextual embeddings, emotion-aware features, and topic-aware enrichment leads to measurable gains in information extraction and retrieval. Notably, performance gains were observed more prominently in noisier datasets such as MMHS-DS1, validating the robustness of the proposed method under challenging linguistic conditions [44].The proposed model achieved an absolute gain of +4.3% in Relation F1 over BERT+CRF and +3.7% in MRR for retrieval, indicating stronger cross-task generalization and better semantic matching.

4.7. Result Analysis

This section analyzes the results obtained from the proposed model across multiple tasks such as named entity recognition, relation extraction, and retrieval. The performance improvements are further dissected to highlight the impact of individual components such as emotion-aware embedding, multi-contextual attention, and domain-specific fine-tuning.

Overall, these findings shows the effectiveness of the proposed framework across multiple tasks and datasets. For the better understand the underlying factors driving these improvements, as well as the challenges that remain, we now turn to a detailed discussion and error analysis.

4.8. Error Analysis

Qualitative error analysis performed on misclassified entities and relations. Major error types included:

- **Ambiguous Contexts**: Tweets with limited background often led to incorrect typing.
- **Meme-Text Conflicts**: In multimodal tweets, text and image conveyed opposite sentiments, confusing the fusion model.
- **Non-standard Language**: Slang and code-switching reduced accuracy in relation classification.

These findings will inform future improvements via improved slang dictionaries, meme-image alignment modules, and continual learning for evolving vocabulary. The results are shown in figure 4.

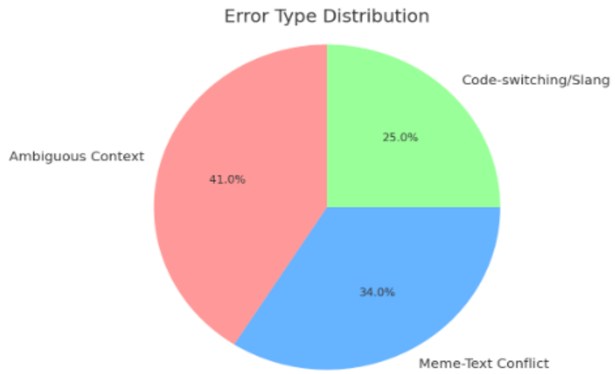

Figure 4. Shows error type distribution

5. Conclusion and Future Work

This study presented a novel framework for free-text information extraction and retrieval from social media networks by integrating contextual embeddings, emotion-aware fusion, and multimodal attention mechanisms. Experimental evaluations on three benchmark datasets demonstrated that the proposed architecture achieves significant improvements in named entity typing, relation extraction, and semantic retrieval tasks. The key contributions include the introduction of EMO for emotion-guided entity disambiguation, the development of MCAM for multi-level contextual fusion, and the design of a unified pipeline that seamlessly bridges information extraction and retrieval, enabling direct construction of knowledge graphs from noisy social content. Overall, the proposed framework enhances the semantic understanding of user-generated content and offers practical applicability in domains such as hate speech detection, public opinion analysis, and automated knowledge base population.

For future research, the proposed framework can be enhanced in multiple directions. A key area is multilingual extension, where models like XLM-R can be employed to handle cross-lingual entity typing and alignment for diverse languages. Another promising direction is the fusion of memes and short-form videos, integrating visual and textual cues through vision-language pre training techniques such as CLIP or Flamingo. To enable real-time deployment, the model can be optimized using lightweight strategies like knowledge distillation and quantization. Additionally, the extracted information can contribute to dynamic knowledge graph expansion, enabling reasoning and relationship discovery with graph neural networks (GNNs). Finally, incorporating explainable AI (XAI) methods will improve the interpretability of model outputs, which is crucial for sensitive tasks such as hate speech detection

References

1. Ritter, A.; Clark, S.; Mausam; Etzioni, O. Named Entity Recognition in Tweets: An Experimental Study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, UK, 27–31 July 2011.
2. Baldwin, T.; de Marneffe, M.C.; Han, B.; Kim, Y.B.; Ritter, A.; Xu, W. Noisy User-generated Text: Mining the New Data Frontier. In Proceedings of the ACL Workshop on Noisy User-generated Text, Sofia, Bulgaria, 8 August 2013.
3. Paul, M.J.; Dredze, M. You Are What You Tweet: Analyzing Twitter for Public Health. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), Barcelona, Spain, 17–21 July 2011.
4. Imran, M.; Castillo, C.; Diaz, F.; Vieweg, S. Processing Social Media Messages in Mass Emergency: A Survey. ACM Comput. Surv. 2015, 47, 67.
5. Dey, L.; Dey, M.; Lahiri, S.; Roy, A. Cybersecurity Threat Detection from Social Media Using Deep Learning. IEEE Access 2018, 6, 63444–63455.
6. Volkova, S.; Wilson, T.; Yarowsky, D. Separating Facts from Opinions in Twitter. In Proceedings of the NAACL-HLT, Valencia, Spain, 3–7 April 2017.
7. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019.
8. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019, arXiv:1907.11692.
9. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. arXiv 2020, arXiv:2004.05150.
10. Aguilar, G.; Kar, S.; Balakrishnan, V.; Solorio, T. Multilingual and Multimodal Hate Speech Detection. In Proceedings of ACL, Bangkok, Thailand, 1–6 August 2021.
11. Baziotis, C.; Athanasiou, N.; Chronis, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Potamianos, A. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs. In Proceedings of SemEval, New Orleans, LA, USA, 5–6 June 2018.
12. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. In Proceedings of EMNLP, Copenhagen, Denmark, 7–11 September 2017.
13. Etzioni, O.; Banko, M.; Cafarella, M.J. Open Information Extraction from the Web. Commun. ACM 2008, 51, 68–74.
14. Ritter, A.; Clark, S.; Etzioni, O. Named Entity Recognition in Tweets: An Experimental Study. In Proceedings of EMNLP, Edinburgh, UK, 27–31 July 2011.
15. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv 2015, arXiv:1508.01991.
16. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019.
17. Aguilar, G.; Kar, S.; Balakrishnan, V.; Solorio, T. Multilingual and Multimodal Hate Speech Detection. In Proceedings of ACL, Bangkok, Thailand, 1–6 August 2021.
18. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation Classification via Convolutional Deep Neural Network. In Proceedings of COLING, Dublin, Ireland, 23–29 August 2014.
19. Miwa, M.; Bansal, M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In Proceedings of ACL, Berlin, Germany, 7–12 August 2016.
20. Soares, L.; FitzGerald, N.; Ling, J.; Kwiatkowski, T. Matching the Blanks: Distributional Similarity for Relation Learning. In Proceedings of ACL, Florence, Italy, 28 July–2 August 2019.
21. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using Millions of Emoji Occurrences to Learn Any-domain Representations. In Proceedings of EMNLP, Copenhagen, Denmark, 7–11 September 2017.
22. Baziotis, C.; Athanasiou, N.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Potamianos, A. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets. In Proceedings of SemEval, New Orleans, LA, USA, 5–6 June 2018.
23. Mausam; Etzioni, O.; Banko, M.; Weld, D.S. Open Language Learning for Information Extraction. In Proceedings of EMNLP, Jeju Island, Korea, 12–14 July 2012.
24. Zhang, Z.; Yao, L.; Sun, A.; Tay, Y. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Trans. Neural Netw. Learn. Syst. 2020, 32, 1212–1232.
25. Xiong, C.; Dai, Z.; Callan, J.; Liu, Z.; Gardner, M. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In Proceedings of SIGIR, Tokyo, Japan, 7–11 August 2017.

26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019.
27. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv 2015, arXiv:1508.01991.
28. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of EMNLP, Hong Kong, China, 3–7 November 2019.
29. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. In Proceedings of EMNLP, Copenhagen, Denmark, 7–11 September 2017.
30. Grootendorst, M. BERTopic: Neural Topic Modeling with Class-based TF-IDF. arXiv 2022, arXiv:2203.05794.
31. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. IEEE Trans. Big Data 2019, 7, 535–547.
32. Han, B.; Cook, P.; Baldwin, T. Automatically Constructing a Normalisation Dictionary for Microblogs. In Proceedings of EMNLP, Jeju Island, Korea, 12–14 July 2012.
33. Ahmad, T., Rana, M. R. R., Akbar, M. A., & Nawaz, A. (2018). A scalable feature extraction technique for social media analysis. Int. J. Comput. Sci. Eng, 7(4), 197-201.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of NeurIPS, Long Beach, CA, USA, 4–9 December 2017.
35. Khalid, A., Mustafa, G., Rana, M. R. R., Alshahrani, S. M., & Alymani, M. (2024). RNN-BiLSTM-CRF based amalgamated deep learning model for electricity theft detection to secure smart grids. PeerJ Computer Science, 10, e1872.
36. Wang, W.; Hoang, C.D.V.; Kan, M.-Y. Emotion-aware Neural Networks for Fine-grained Opinion Mining. In Proceedings of ACL, Seattle, WA, USA, 5–10 July 2020.
37. Wood, D.; Zaidman, M.; Ruth, L.; Herman, I. RDF 1.1 Primer. W3C Recommendation 2014.
38. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019.
39. Mathew, B.; Saha, P.; Yimam, S.M.; Biemann, C.; Goyal, P.; Mukherjee, A. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In Proceedings of AAAI, New York, NY, USA, 7–12 February 2020.
40. Zhang, Z.; Robinson, D.; Tepper, J. Detecting Hate Speech on Social Media: An Analysis. Inf. Process. Manag. 2018, 54, 1–13.
41. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of NAACL, San Diego, CA, USA, 12–17 June 2016.
42. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of COLING, Santa Fe, NM, USA, 20–26 August 2018.
43. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. In Proceedings of EMNLP, Copenhagen, Denmark, 7–11 September 2017.
44. Liu, Y.; Gu, J.; Goyal, N.; Ott, M.; Shleifer, S.; Fan, A.; Lewis, M.; Levy, O.; Stoyanov, V. Multilingual Denoising Pre-training for Neural Machine Translation. In Proceedings of ACL, Seattle, WA, USA, 5–10 July 2020.
45. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of ICWSM, Montreal, QC, Canada, 15–18 May 2017.
46. Founta, A.M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Kourtellis, N. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In Proceedings of ICWSM, Stanford, CA, USA, 25–28 June 2018.
47. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of NAACL-HLT, San Diego, CA, USA, 12–17 June 2016.
48. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019.