Ontology-Augmented Transformer Architecture For Advanced Sentiment Analysis in Sindhi

Presentation For the Initial Seminar

QUAID.E.AWAM UNIVERSITY OF ENGINEERING SCIENCE AND TECHNOLOGY ,NAWABSHAH

Department : Information Technology

Anisha Azam Ali

24-MS(I.T)-04

Supervisor: Prof Dr. Akhter Hussain Jalbani

Content

Introduction

Applications

Literature Review

Problem Statement

Aims and Objectives

Proposed Framework

Tools and Techniques

Expected Outcomes

Timeline

References

Introduction

The goal of artificial intelligence (AI), a revolutionary area of computer science, is to mimic human intelligence in machines. Machine learning, neural networks, computer vision, robotics, and natural language processing (NLP) are among the fundamental fields of artificial intelligence.

Sentiment analysis, or the act of recognizing and classifying feelings or sentiments represented in text, is the main focus of this proposal, which is centered on natural language processing.

Because there are few linguistic tools and annotated datasets available, sentiment analysis in low-resource languages like Sindhi is still mostly unexplored.

In order to enhance sentiment classification in Sindhi text, this proposal presents the concept of linking BERT (Bidirectional Encoder Representations from Transformers) with an ontology-driven framework.

What is Ontology?

Ontology in computer science and artificial intelligence refers to a formal, structured representation of knowledge ? it defines concepts, their properties, and the relationships among them within a specific domain.

Ontology in NLP and Sentiment Analysis:

Semantic Understanding: Ontologies allow machines to understand words beyond surface-level meaning, considering context, polarity, and related terms.

Disambiguation: Helps distinguish between multiple meanings of a word using concept hierarchies and relations (e.g., "bank" as a financial institution vs. riverbank).

Sentiment Assignment: Words or concepts can be assigned positive, negative, or neutral sentiment scores, improving accuracy over keyword-based or purely statistical models.

Low-Resource Language Support: In languages like Sindhi, where pre-trained resources and datasets are limited, ontologies provide handcrafted semantic knowledge to fill the gap.

Applications

1 Social Media Monitoring ? Analyze public sentiment on platforms like Twitter and Facebook in the Sindhi language.

2 News and Media Analytics ? Classify emotional tone in Sindhi news for media bias detection and audience insight.

3 E-Governance and Feedback Analysis? Understand citizen feedback written in Sindhi to improve government services

4 Business Intelligence? Evaluate Sindhi customer opinions to refine marketing and product strategies.

5 Academic and Linguistic Research? Support future studies on low-resource languages through datasets and ontology

6 Healthcare Opinion Mining? Extract sentiment from Sindhi health-related reviews to improve patient care.

7 Cultural and Political Studies ? Track regional sociopolitical trends through sentiment analysis of public texts.

Department : Information Technology

(Slide 10) No readable text found.

(Slide 11) No readable text found.

(Slide 12) No readable text found.

(Slide 13) No readable text found.

(Slide 14) No readable text found.

(Slide 15) No readable text found.

(Slide 16) No readable text found.

(Slide 17) No readable text found.

Literature Review(Summary)

Over the years, integrating ontologies and machine learning models has significantly enhanced sentiment analysis across multiple languages, including low-resource ones like Sindhi and Urdu. Rattar et al.[1] Created a bilingual Sindhi-English ontology that classified verbs and adjectives to represent semantic relationships, achieving up to 100% accuracy through manual translation. Similarly, Khabour et al.[2] proposed an Arabic ontology-based sentiment analysis method that achieved 79.20% accuracy, while Zulf and Jamil [3] built an emotion ontology for Roman Urdu using WordNet-Affect, reporting 92.87% precision. In the domain of religious texts, Muhammad et al.[4] emphasized semantic enrichment in Hadith classification using ontologies, attaining a 94.5% F1 score. Rajput [5] developed an ontology-based framework for Urdu web content annotation, achieving perfect accuracy on attributes like make and model in ads. Barakzai et al.[6] utilized BERT for Sindhi sentiment analysis, obtaining 67.2% accuracy, indicating a need to embed Sindhi vocabulary into models. Khan et al. [7] employed multilingual BERT (mBERT) for Urdu sentiment analysis and surpassed conventional models with an F1 score of 81.49%. Tao and Liu [8] introduced DEMLOnto, an ontology-based model handling multilingual emotional semantics. Zhao and Lee [9] combined ontology-derived features with deep learning for fine-grained sentiment analysis, while Sharma et al.[10] used domain ontologies in OMASA for multi-aspect sentiment classification. Chowdary and Yadav [11] improved polarity detection using word sense disambiguation in an ontology-driven sentiment framework. Furthering this, Ali et al. [12]

Literature Review(Summary)

Developed sentiment resources for Sindhi, while Soomro et al.[13] conducted a systematic review highlighting limitations in Sindhi NLP tools. Singh et al. [14] tested hybrid deep learning on Urdu reviews, and Chandio et al. \[15] proposed an attention-based RU-BiLSTM model for Roman Urdu. Li et al. [16] implemented transfer learning for Roman Urdu sentiment tasks. Rehman and Bajwa [17] used a lexicon-based model for Urdu with promising results. Khan et al. [18] reviewed Urdu and Roman Urdu sentiment research in a multilingual context. Koto et al. [19] proposed zero-shot sentiment models using multilingual sentiment lexicons, and Asgari et al. [20] built UniSent?sentiment lexica for 1,000+ languages. Malinga et al. [21] explored large language models for multilingual lexicon building in low-resource settings. Bello et al. [22] focused on lexicon-based sentiment analysis in underrepresented languages. Muhammad et al. [23] developed NaijaSenti, a multilingual sentiment dataset for African languages. Salahudeen et al. [24] adapted Hausa tweets for sentiment classification using African tweet corpora, and Le et al. [25] tackled sentiment analysis on informal Indonesian tweets to support low-resource languages. Collectively, these studies emphasize the vital role of ontologies, multilingual resources, and deep learning in bridging the gap for under-resourced languages like Sindhi, paving the way for context-aware, semantically enriched sentiment analysis systems.

Problem Statement

Languages like Sindhi pose unique challenges for sentiment analysis due to their complex semantic structures and limited annotated resources. Traditional transformer-based models such as BERT struggle to capture context-specific sentiment and deep semantic relationships in such low-resource languages. The lack of comprehensive ontologies and domain-specific features further hinders performance. Previous studies (Sharma et al., 2024; Zhao & Lee, 2024) show that combining syntactic and semantic knowledge with machine learning can improve accuracy. Therefore, this study proposes an ontology-enhanced sentiment analysis framework that integrates semantic features with BERT to enhance performance for Sindhi text.

Aim and Objectives

Aim:

The goal is to create a sentiment analysis framework for Sindhi text by combining a BERT-based model with ontology-based semantic enrichment.

Objectives:

1.To Construct a Labelled dataset of particular Sindhi sentences and associated word-level sentiments.

2. To build a sentiment Ontology that captures domain-specific semantics of Sindhi text.

3. To train and compare two Bert-based sentiment classifiers: one on traditional dataset, and another using the ontology-enhanced datasets.

4. To evaluate and demonstrate improvements in model performance when using structured semantic knowledge.

Proposed

Frame Work

Proposed Frame Work

This research follows a two-phase methodology to evaluate the impact of ontology integration on sentiment analysis using a BERT-based model for the Sindhi language.

Phase 1: Without Ontology (Baseline Pipeline)Data Collection: A labeled Sindhi dataset with sentence-level and word-level sentiment annotations is collected.

Data Preprocessing: Text cleaning (removal of special characters, normalization) Tokenization, Encoding the data for BERT-compatible input.

Labeling and Encoding: Sentences and associated words are labeled with sentiment classes (positive, negative, neutral) and converted into numerical formats.

Model Training: A baseline BERT model is trained on the preprocessed Sindhi data without ontology support. The model learns from purely lexical and contextual cues.

Model Evaluation: Performance is measured using standard metrics (accuracy, F1-score) and serves as a benchmark for the ontology-enhanced model.

Cont?d

Phase 2: With Ontology (Ontology-Enhanced Pipeline)

Ontology Integration: A domain-specific sentiment ontology is developed from the dataset, representing semantic relationships between words, sentiments, and domains.

2. Sentiment and Token Mapping:

Words and tokens in the dataset are linked to corresponding ontology concepts.

Sentiment values from the ontology are mapped back to tokens to enrich input features.

3. Feature Integration: Ontology-based sentiment and semantic information are injected into the input embeddings alongside BERT token embeddings.

4. Model Retraining and Comparison:

An ontology-enhanced version of the BERT model is trained using the enriched dataset.

This model is evaluated and compared against the baseline to determine performance gains.

Proposed FrameWork Cont?d

Final Steps:

Comparison of Models: The performance of the ontology-enhanced model is compared against the baseline BERT using accuracy, F1-score, and confusion matrix.

Visualization and Insights: Graphs and tables are used to visualize improvements and analyze model behavior with and without ontology support.

Tools And Techniques

Tools:

? HuggingFaceTransformers for Bert Model Fine Tuning.

? Jupyter NoteBook for experimentation and training environment.

? Libraries: seaborn, matplotlib, sklearn, rdflib and transformers

? Editor of Ontology: Protégé

Techniques:

? Python is used as a core development language.

? Ontology Parsing: Using rdflib to extract sentiment annotations and semantic relationships.

? Tokenization: For Sindhi text, use the BERT tokenizer.

? Refinement of BERT for baseline and ontology-enhanced models in sentiment classification.

? Evaluation: Models are evaluated using confusion matrices, F1-score, recall, accuracy, and precision

Expected Outcomes:

A rich and Diverse Sindhi sentiment dataset.

A reusable OWL-based sentiment ontology.

Improved Sindhi text sentiment classification accuracy.

The efficiency of ontology-based semantic enrichment in natural language processing is demonstrated.

Transformer models for other low-resource languages can be combined with ontologies using this reusable framework.

Timeline

References

[1] M. K. Rattar, M. A. Rattar, M. M. Rind, M. Hyder, and W. Khan, ?Sindhi English Bilingual Parallel Ontological Dictionary,? Sindh University Research Journal (Science Series), doi: 10.26692/sujo/2018.06.0046.

[2] S. M. Khabour, Q. A. Al-Radaideh, and D. Mustafa, ?A New Ontology-Based Method for Arabic Sentiment Analysis,? Big Data and Cognitive Computing, vol. 6, no. 2, Jun. 2022, doi: 10.3390/bdcc6020048.

[3] G. Zulf and N. Jamil, ?Generating an emotion ontology for Roman Urdu text,? International Journal of Computational Linguistics Research, vol. 7, no. 3, pp. 83?91, 2016.

[4] M. A. Muhammed et al., ?Arabic Ontology for Hadith Texts ? A Survey,? The Egyptian Journal of Language Engineering, vol. 11, no. 1, pp. 1?14, 2024.

[5] Q. Rajput, ?Ontology-based semantic annotation of Urdu language web documents,? Procedia Computer Science, vol. 35, pp. 662?670, 2014.

[6] F. Barakzai, S. Bhatti, and S. Saddar, ?Sentiment Analysis of Sindhi News Articles using Deep Learning,? presented at the IEEE International Conference on Computer Science and Information Technology, 2022, doi: 10.1109/CSIT56902.2022.10000519.

.

Department : Information Technology

[7] L. Khan et al., ?Multi-class sentiment analysis of Urdu text using multilingual BERT,? Scientific Reports, vol. 12, no. 1, pp. 5436?5446, 2022.

[8] W. Tao and T. Liu, ?Building ontology for different emotional contexts and multilingual environment in opinion mining,? Intelligent Automation & Soft Computing, pp. 1?7, 2017.

[9] L. Zhao and S.-W. Lee, ?Integrating Ontology-Based Approaches with Deep Learning Models for Fine-Grained Sentiment Analysis,? Computers, Materials & Continua, vol. 81, no. 1, pp. 1?14, 2024.

[10] S. Sharma, M. Saraswat, and A. K. Dubey, ?Multi-aspect sentiment analysis using domain ontologies,? in Iberoamerican Knowledge Graphs and Semantic Web Conference, Springer, 2022.

[11] N. Yadav and C. R. Chowdary, ?Feature-based sentiment analysis using a domain ontology,? presented at the 13th International Conference on Natural Language Processing, 2016.

[12] W. Ali, N. Ali, Y. Dai, J. Kumar, S. Tumrani, and Z. Xu, ?Creating and Evaluating Resources for Sentiment Analysis in the Low-Resource Language: Sindhi,? in Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2021, pp. 188?194.

References

[13] S. Soomro, A. Memon, and M. A. Memon, ?A Systematic Review on Sentiment Analysis for Sindhi Text,? Baghdad Science Journal, vol. 18, no. 2, pp. 1?10, 2021.

[14] N. Singh and U. C. Jaiswal, ?Sentiment Analysis Based on Urdu Reviews Using Hybrid Deep Learning Models,? Applied Computer Systems, vol. 28, no. 1, pp. 26?34, 2023

[15] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, ?Attention-Based RU-BiLSTM Sentiment Analysis Model for Roman Urdu,? Applied Sciences, vol. 12, no. 7, p. 3641, 2022. [16] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, ?Roman Urdu Sentiment Analysis Using Transfer Learning,? Applied Sciences, vol. 12, no. 20, p. 10344, 2022.

[17] Z. U. Rehman and I. Bajwa, ?Lexicon-Based Sentiment Analysis for Urdu Language,? in 2016 International Conference on Innovative Computing Technology (INTECH), 2016

[18] L. Khan, A. Amjad, and N. Ashraf, ?A Review of Urdu Sentiment Analysis with Multilingual Perspective: A Case of Urdu and Roman Urdu Language,? Computers, vol. 11, no. 1, p. 3, 2022

[19] F. Koto, T. Beck, Z. Talat, I. Gurevych, and T. Baldwin, ?Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon,? in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024), 2024

[20] E. Asgari, F. Braune, B. Roth, C. Ringlstetter, and M. Mofrad, ?UniSent: Universal Adaptable Sentiment Lexica for 1000+ Languages,? in Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), 2020

References

[21] M. Malinga, M. W. Nkongolo, and M. M. Malinga, ?A Multilingual Sentiment Lexicon for Low-Resource Language Translation and Sentiment Analysis,? arXiv preprint arXiv:2411.04316, 2024.

[22] B. S. Bello, S. S. Muhammad, and D. I. Adelani, ?Lexicon-Based Sentiment Analysis for Underrepresented Languages,? Journal of African Languages and Linguistics, vol. 45, no. 2, pp. 123?140, 2023

[23] S. H. Muhammad et al., ?NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis,? arXiv preprint arXiv:2201.08277, 2022

[24] S. S. Abdullahi, D. I. Adelani, and B. S. Bello, ?HausaNLP at SemEval 2023 Task 12: Leveraging African Low-Resource Tweet Data for Sentiment Analysis,? in Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)

[25] T. A. Le, D. Moeljadi, Y. Miura, and T. Ohkuma, ?Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets,? in Proceedings of the 12th Workshop on Asian Language Resources (ALR12), 2016