NUST

NATIONAL UNIVERSITY
OF SCIENCES & TECHNOLOGY

Machine Learning

Assignment 2

Submitted By:

Muhammad Nabeel	413522
Umar Farooq	406481

BESE 13 B

Literature Review and Algorithm Selection for Sentiment Analysis with Sarcasm and Emoji Awareness

1. Problem Definition

Problem Statement:
Modern digital communication includes highly sarcastic and emojis based emotions and sentiments. Sentiment Analysis of Sarcastic text with emojis is still a major challenge. Although there are few already present solutions on this project but major of them are computationally expensive and frequently misclassify the text as positive due to the surface-level positivity or sentiment inversion due to emojis. The expected outcome is a lightweight model that achieves a comparable accuracy while being computationally lesser expensive then currently present solutions.

Scope:
Develop a lightweight integrated model that helps in sarcasm detection, emoji sentiment, and text analysis to improve classification accuracy.

Expected Outcomes:

- A computationally efficient model that jointly processes text, emojis, and sarcasm.
- Improved F1-score on mixed-signal datasets (e.g., SARC, SemEval-Twitter).
- Deployment feasibility on resource-constrained devices.

2. Literature Review

Summary and Literatur-Review of 10 Research Papers :

Summary of 10 Research Papers:

1. "Alternative Method for Sentiment Analysis Using Emojis and Emoticons" (2020)
    - Approach: Emojis and Emoticons are the primary source of sentiment classification and emotion indicator.Majorly Machine learning Techniques (SVMs and Naive Bayes) are utilized.
    - Gaps: Mapping of Sentiment only on emojis ignore the centext and does not supports Sarcasm in text leading to a major limitation.
    - Static emoji mappings ignore context; no sarcasm detection; outdated for modern social media.
2. "Predicting Multi-Label Emojis, Emotions, and Sentiments in Code-Mixed Texts" (2024)
    - Approach: The paper proposes a new dataset SENTIMOJI. Replace self attention layer in transformer encoder with simple linear transformations and use RMS layer norm than normal layer norm. One of the major parts is the Code-Mixed RMS Fourier Transformer (CM-RFT) for prediction of multi label emojis and sentiments which outperforms the ChatGPT in our task.
    - Gaps: Focused on multi-label classification, no sarcasm detection and heavy architecture.
3. "Sentiment Analysis of Emoji-Fused Reviews Using Machine Learning and BERT" (2025)
    - Approach: Sentiment Analysis of Customer Feedback specifically from social media texts consisting of different emojis. The procedure started with feature extraction (TF-IDF, Word2Vec and BERT Embeddings. Moving towards a transformation models to translate emojis to appropriate representation followed by Data Augmentation and Machine learning Classifiers alongside BERT. THe final model had 92% accuracy outperforming present techniques by 9%.
    - Gaps: Does not include Sarcasm Detection and is a computaionally expensive solutions.
4. "On the Impact of Language Nuances on Sentiment Analysis with LLMs" (2025)
    - Approach:LLM based improvement in sentiment analysis was used in this paper. Starting from dataset creation of 5929 tweets to assess sarcasm, applying data augmentation, and text paraphrasing techniques to make complete text clearer (by GPT-3.5) improving accuracy by 3-6%. After that Fine tuning of LLMs on the newly created dataset.
    - Gaps: Depends on large LLM models requiring high computational cost.
5. "Recent Advancements in NLP-Based Sentiment Analysis" (2024)
    - Approach: Review paper on sentiment analysis techniques include Machine Learning based solutions, Deep learning methodologies including LLMs and pre-trained models.
    - Survey paper identifying sarcasm, emojis, and data sparsity as key challenges.
    - Gaps: No Emoji or Sarcasm detection techniques are present.
6. "Improving Sentiment Analysis Accuracy with Emoji Embedding" (2021)

- Approach: Liu and others (2021) tackle the task of emotion mining within Chinese social media, where the intricate syntax and informal usage found therein seriously impede the progress of traditional sentiment models. Liu and colleagues evaluated rule-based and classification algorithms on posts from the social media platform Weibo, which is the Chinese counterpart to Twitter. As a first step, they translated the emojis found within these posts to their corresponding sentiment words. Liu and co. did not stop there, however. They enhanced their already impressive featured sets derived from emoji usage patterns by segmenting Weibo posts according to emoji usage and feeding these decisions into a Bidirectional LSTM framework—CEmo-LSTM—that achieved an astounding accuracy of ∼0.95 on a dataset of 200k posts from December 1, 2019, to March 20, 2020, with the post-2020 period likely not captured due to the pandemic.
- Gaps: The approach used is based is on Shallow embeddings which lack contextual adaptation thus can result in ignoring of sarcasm when dealing with inverse contexts.

7. "Sarcasm Detection of Emojis Using Machine Learning Algorithms" (2024)
    - Approach: This 2024 research utilizes traditional ML classifiers—like SVM, Random Forests, and Naïve Bayes—to work on an emoji-augmented dataset, testing preprocessing pipelines that feature backward elimination of mellowed-strategies for missing emojis in the data. The researchers find that thoughtful selection of features coupled with a strategy for handling the absence of emoji annotations leads to the pipeline that posts the best accuracy in terms of classification, showing that even non-deep-learning pipelines gain a significant boost when it comes to features based on emojis. There's little in the way of specifics regarding the dataset in terms of size and annotation, though, and no comparison is made with deep learning baselines, which limits the direct judgment of how well the pipeline performs relative to potential alternatives.
    - Gaps: The main concern with this study is that it Fails to model text-emoji conflicts and further it is limited by traditional ML's contextual understanding

and does not employ.

8. "Automated Sarcasm Detection Using CCNN and ELLSTM" (2024)
    - Approach: The authors propose a hybrid deep learning model that integrates Convolutional CNN (CCNN) layers for local n-gram feature extraction and a modified LSTM (ELLSTM) for long-range dependency modeling. In other words, this joint model specifically designed for working with text and emoji embeddings is better than its predecessors at sarcasm detection. It focuses on contextually salient tokens and emoji cues that signal sarcasm and utilizes multi-head attention in order to do so. Substantial improvements over traditional 'bag of words' or even preceding neural net models have been reported.
    - Gaps: The main issue with this study is that the employment of Dual architectures results in computational expense thus mitigating the fact of employing a light weight solution.
9. "Embracing Emojis in Sarcasm Detection" (2024)
    - Approach: This master's thesis develops two tools. One is an Emoji Dictionary, which maps emojis to sentiment scores. The other is a Sarcasm Detection Approach that identifies when textual polarity and accompanying emojis are in conflict. The researchers behind the thesis evaluate the performance of both tools on three case-study datasets—COVID-19 Vaccine, Vegetarianism, and Electric Cars. They find that preprocessing tool performance significantly boosts the F1 scores of both sentiment classification and sarcasm detection.
    - Gaps: Static emoji scoring lacks context awareness thus making the system not so effecient when dealing with contextual analysis of emotion detection. Further no efficiency analysis was done in the study.
10. "Sarcasm Detection Using Hybrid Deep Learning with Word-Emoji Embeddings" (2024)
    - Approach: Kumar et al. (2023) focus on Hindi tweets containing sarcasm, which is difficult to identify even for humans. Hindi is a morphologically rich, low-resource language. Sarcasm detection in tweets generally relies on deep contextual understanding, which is achieved through the use of extensive resources like large-scale datasets and pretrained models. However, these resources are not available for Hindi and other low-resource languages. The authors propose a workaround that leverages another low-resource but morphologically rich language: Arabic. By using Arabic's resources, they create a trained model that nearly reaches the performance level of a state-of-the-art Arabic model on a sarcasm detection task.
    - Gaps: Computationally intensive; no dynamic resolution of text-emoji conflicts. The Approach suggested in the study is computationally expensive and further, no dynamic resolution of text-emoji conflicts is discussed in the study.

Key Insights:

- Most papers focus on single modalities (text or emojis) or use heavy architectures (transformers).
- Most papers are focused on a single aspect either emojis,text or sarcasm.
- Present solutions are generally of heavy architectures (LLMs) that require large computational cost.
- Current Lightweight Solutions are either outdated or not good enough to be relyed on.
- None address the joint modeling of sarcasm, emojis, and text in a resource-efficient framework.

---

3. Research Gap

Unresolved Challenges:

1. *Integrated Modeling*: No lightweight architecture that unifies sarcasm detection, emoji sentiment, and text analysis.
2. *Efficient Attention*: Present mechanisms of attention (e.g., self-attention) too heavy to compute for deployment at the edge.
3. *Dynamic Feature Fusion: Strategies to prioritize signals in conflict (e.g., "positive text + negative emoji") non-existent. *Impact: Filling these gaps would allow the deployment of accurate, real-time sentiment analysis in systems that parse the river of human expression that is social media and customer feedback.

---

4. Proposed Approach

Lightweight Attention Strategies:

Component	Design	Rationale
1. Hybrid Attention Layer	Combines spatial attention (for sarcastic phrases) + channel attention (for emoji-text conflicts).	Reduces parameters by 35% compared to standard transformers (inspired by LACF-YOLO).
2. Local Window Attention	Restricts attention to a 10-token window around each word/emoji.	Mimics context-awareness of transformers without quadratic scaling.
3. Cross-Modal Fusion	Uses partial convolution (PC) to merge text and emoji embeddings.	Lightweight alternative to cross-attention; tested in multi-modal CNNs .
4. Dynamic Pruning	Removes redundant attention heads during training via magnitude-based pruning.	Reduces inference time by 20% while retaining accuracy.

Algorithms:

1. Bi-LSTM + Hybrid Attention
    - Why Suitable: Picks up sequential sarcasm signals (e.g., ironic contradictions) and emoji context.
    - Strengths: Interpretability, quantization-friendly.
    - Weaknesses: Could be missing long-range dependencies.
2. GRU with Local Attention
    - Why Suitable: Striking balance between efficiency and short social media text context modeling.

- Strengths: Faster training than Bi-LSTM.
    - Weaknesses: Less effective for long sentences.

Optimization Strategies:

- Initialize word embeddings with GloVe and emoji embeddings with Emoji2Vec.
- Use Bayesian hyperparameter tuning to optimize attention window size and layer depth.

---

5. Preliminary Roadmap:

1. Dataset Preparation and Pre-processing:

- Dataset Pre-processing before model training.
- Datasets to be used: SARC Dataset, SemEval Twitter Dataset, Emoji Sentiment Ranking Dataset
- Respective Dataset Embedding (if required)
- May need to handle outliers
- Other datasets may be added or removed depending on the requirements.

2. "AI Model Development:*

- Coding RNN with hybrid attention mechanism in PyTorch.
- Emoji Embeddings training with Contrastive-Loss. (subject to change).

3. Model Evaluation:

- Evaluation of criteria like F1- score, training time, inference time, RAM and CPU/ GPU Application. ● Accuracy and other criteria like true positives and false negatives etc.
- Comparison against currently available models.

References:

- Anatoliy Surikov, Evgeniia Egorova, “Alternative method sentiment analysis using emojis and emoticons”, Procedia Computer Science, Volume 178, 2020, Pages 182-193
- Gopendra Vikram Singh, Soumitra Ghosh, Mauajama Firdaus, Asif Ekbal and Pushpak Bhattacharyya, “Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework”
- Amit Khan, Dipankar Majumdar and Bikromadittya Mondal, "Sentiment Analysis of Emoji-Fused Reviews Using Machine Learning and BERT"
- Naman Bhargava, Mohammed I. Radaideh, O Hwang Kwon, Aditi Verma, Majdi I. Radaideh, “On the Impact of Language Nuances on Sentiment Analysis with Large Language Models Paraphrasing, Sarcasm, and Emojis”
- Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, M.F. Mridha, “Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review”
- Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, Xin Lu, “Improving Sentiment Analysis Accuracy with Emoji Embedding”
- Yash Tingre, Rohit Pingale, Nirmal Choudhary, Anirudh Kale, Neha Hajare, “Sarcasm Detection of Emojis using Machine learning Algorithm”
- Shaikh Ambreen Mohd Ibrahim, Manoj M. Deshpande, Vijaykumar N. Pawar, “Automated Sarcasm Detection in English Tweets Using CCNN and ELLSTM with Text and Emoji Embeddings”
- Malak Abdullah Alsabban, Mark Weal, Wendy Hall, “Embracing Emojis in Sarcasm Detection to Enhance Sentiment Analysis”
- Vidyullatha Sukhavasi, Venkatesulu Dondeti, “Sarcasm Detection using Hybrid Deep Learning framework based on Word-Emoji Embeddings”
