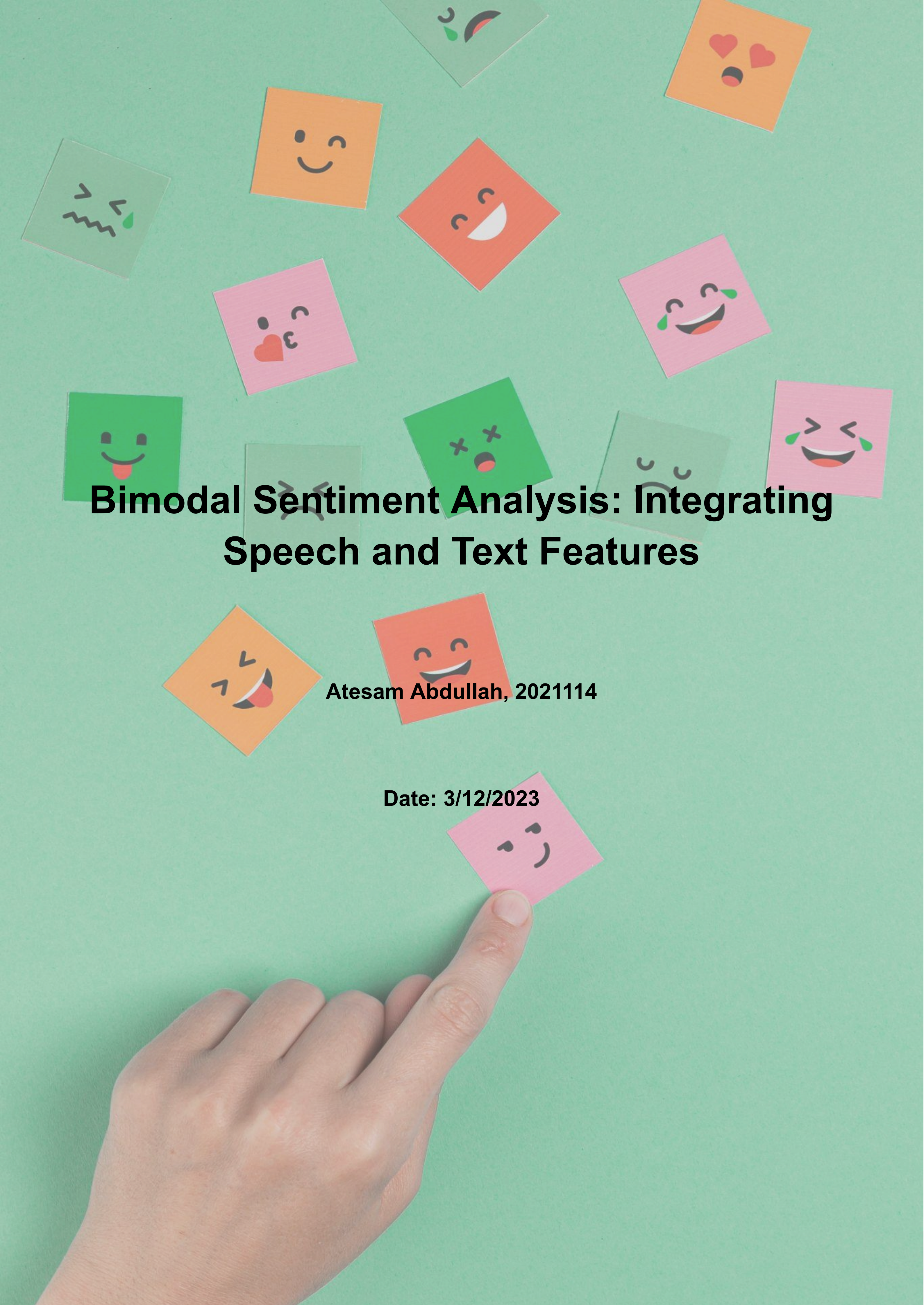

Bimodal Sentiment Analysis: Integrating Speech and Text Features

Atesam Abdullah, 2021114

Date: 3/12/2023

I. Introduction

The concept of bimodal sentiment analysis, which combines speech and text analysis to gauge human emotion, emerged in the early 2010s. Since then, the field has witnessed rapid growth, driven by the increasing availability of multimodal data and advancements in machine learning techniques. Today, the global bimodal sentiment analysis market is estimated to reach a staggering $1.5 billion by 2028, with an average annual growth rate of 23.5%. This growth is fueled by the increasing adoption of bimodal sentiment analysis across diverse industries, including customer service, healthcare, marketing, and finance.

Significance

This enhanced understanding of sentiment holds immense potential for various applications, including:

1. Customer Service Enhancement: Businesses can proactively identify and address customer dissatisfaction during interactions, improving customer satisfaction and retention rates.
2. Marketing Strategy Optimization: Marketers can refine their strategies based on customer reactions to new products, services, and advertising campaigns.
3. Improved Healthcare Outcomes: Healthcare providers can assess patients' emotional well-being and identify potential mental health concerns, facilitating timely interventions.
4. Safeguarding Online Environments: Social media platforms can detect and flag potentially harmful content, promoting a safer and more inclusive online environment.

II. Problem Statement

Gap Analysis

- Data Availability and Annotation: Limited access to large-scale, well-annotated multimodal datasets hinders BSA model development and evaluation.
- Model Accuracy and Robustness: Inconsistent and inaccurate sentiment classification, particularly in noisy and complex environments, remains a challenge.
- Integration with Real-World Applications: Computational complexity, privacy concerns, and integration hurdles impede BSA adoption in real-time scenarios.
- Explainability and Interpretability: Opaque model decision-making processes hinder trust and adoption in sensitive applications.
- Cross-Cultural Adaptation: BSA models often struggle to generalize across different cultures due to linguistic and emotional expression variations.

Scope

1. Study existing sentiment analysis methods using speech or text data independently.
2. Develop a robust and reproducible prototype merging speech and text for sentiment analysis.
3. Perform thorough testing to ensure system accuracy and reliability.
4. Write a detailed research paper on methodology, implementation, and findings.

III. Literature Review

Title	Year	Methodology	Results	Key Contributions
Chen et al [1]	2018	CNN, CNN-BiLSTM	Acc: 71.5,% 77.4%	Dataset creation. Human Labeling. Baseline models.
Poria et al [2]	2018	textCNN, bcLSTM, dRNN	F1 score: 64.25, 66.68 67.57	Dataset creation. Added context and dialogue linkage. Baseline models.
Ghosal et al [3]	2020	COSMIC	F1 score: 73.20	Commonsense-guided framework. GRU-cell based modelling.

Wu et al [4]	2021	Bimodal Multi-Head Attention (BMHA)	F1 score: 82.72	Multi-Head Attention approach. Relation b/t pair-wise modalities. Extensive testing (4 datasets, 12 models)
Shah et al [5]	2023	RoBERTa + Wave2Vec, ANN ensemble	Acc: 68.04%, 74.4%	ANN based ensemble model.
Li and Okada [6]	2023	BERT, RoBERTa	F1 score: 84.62, 85.65	Extraction of facial data features. Use of GPT for context generation.
Hu et al [7]	2022	LSTM + T5	Acc: 65.09%	Multimodal sentiment knowledge-sharing framework (UniMSE). Combination of MSA and ERC.
Hazarik et al [8]	2018	GRU cell based model	Acc: 64.0% F1 score: 63.5	Interactive COnversational memory Network (ICON). Self and inter-speaker dialogue influence.

IV. Proposed Solution

Dataset

The MELD dataset [9] is a valuable benchmark for evaluating and advancing bimodal sentiment analysis techniques. This comprehensive collection of multimodal data encompasses over 1400 conversations, encompassing 13,000 utterances, categorized into seven emotions: neutral, surprise, fear, sadness, joy, disgust, and anger. The utterances are also categorized into three labels (positive, negative, and neutral), providing a rich foundation for developing and testing bimodal sentiment analysis models.

Proposed Methodology

The model integrates text and sound modalities using separate encoders to capture contextual representations. These representations are fused before guiding the decoder to generate the output sequence. Training employs cross-entropy loss, and evaluation uses multiple metrics to assess performance.

- Text Encoder: CNN - BiLSTM
- Audio Encoder: CNN - BiLSTM
- Fusion Layer: Fully Connected Layer
- Decoder: LSTM

Innovation

- It uses a BiLSTM encoder, which is a more powerful architecture than the CNN and RNN encoders common in previous models.
- It uses an attention layer, which allows the model to focus on the most relevant parts of the input sequence.
- It is trained on a large corpus of conversational data, which allows the model to learn more generalizable patterns of behavior.
- Our model, on the other hand, takes into account both the sentiment of the individual words and phrases in the input sequence, as well as the context of the conversation.

V. References

[1] https://doi.org/10.48550/arXiv.1802.08379
[2] https://doi.org/10.48550/arXiv.1810.02508
[3] https://arxiv.org/pdf/2010.02795.pdf
[4] https://arxiv.org/pdf/2103.02362.pdf
[5] https://doi.org/10.3390/bdcc7020085
[6] https://arxiv.org/pdf/2305.06162v3.pdf
[7] https://arxiv.org/pdf/2211.11256v1.pdf
[8] https://aclanthology.org/D18-1280.pdf
[9] https://affective-meld.github.io/