Linguistic Analysis of Sinhala YouTube Comments on Sinhala Music Videos: A Dataset Study

W. M. Yomal De Mel and **Nisansa de Silva**
Department of Computer Science & Engineering,
University of Moratuwa, Moratuwa,
Sri Lanka
{mario.23,NisansaDdS}@cse.mrt.ac.lk

Table of contents

01. Introduction

This research examine emotional dimensions in Sinhala music through Music Information Retrieval (MIR) and Music Emotion Recognition (MER).

Limited exploration of Sri Lankan music within MIR; challenges due to scarcity of Sinhala NLP resources.

Sinhala songs represent a rich cultural heritage and diverse musical genres in Sri Lanka.

The objective of this research is to develop a comprehensive dataset of Sinhala YouTube comments to analyze linguistic patterns and emotional expressions.

Integrate computational linguistics and NLP to identify frequent words and word pairs, comparing with general Sinhala corpora.

01. Introduction Cont.

Utilization of Sinhala Wikipedia, newspaper articles, and previous NLP research to complement YouTube data.

The research highlights the lexical diversity and unique linguistic traits in Sinhala music comments, enhancing understanding of listener emotions.

This contribute to provide insights into how social media comments can serve as authentic sources for emotional analysis in underrepresented languages and music traditions.

02. Related Work

2.1 Development of Sinhala Corpora

This research examine emotional dimensions in Sinhala music through Music Information Retrieval (MIR) and Music Emotion Recognition (MER) [1].

Early contributions by Upeksha et al. [2] and De Silva et al. [3] focused on compiling data from online sources such as news, academic content, and gazettes.

A significant advancement came from Wijeratne et al. [4], who released a large dataset of Sinhala Facebook posts, making it publicly available.

More recently, Hettiarachchi et al. [5] introduced a corpus of 506,932 news articles, significantly expanding the available data for Sinhala NLP.

[1] N. de Silva, “Survey on publicly available sinhala natural language processing tools and research,” arXivpreprint arXiv:1906.02358, 2019.

[2] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. De Silva, and G. Dias, “Implementing a corpus for sinhala language,” in Symposium on Language Technology for South Asia, vol. 2015, 2015, p. 3.

[3] N. de Silva, “Sinhala text classification: observations from the perspective of a resource poor language,” ResearchGate, 2015.

[4] Y. Wijeratne and N. de Silva, “Sinhala language corpora and stopwords from a decade of sri lankan facebook,” arXiv preprint arXiv:2007.07884, 2020.

[5] H. Hettiarachchi, D. Premasiri, L. Uyangodage, and T. Ranasinghe, “Nsina: A news corpus for sinhala,” arXiv preprint arXiv:2403.16571, 2024.

02. Related Work (Cnt.)

2.2 Music Information Retrieval (MIR)

MIR strategies enable efficient access to music collections and benefit various groups, including industry professionals, researchers, and end-users [6].

Techniques such as pitch histograms, Spiral Arrays, and graph structures are used to analyze and represent music data [7].

2.3 Music Emotion Recognition (MER)

MER focuses on analyzing the emotional content in music using computational methods, with applications in personalized music recommendations and mood-based playlists [8,9].

Despite challenges like subjective emotional perception, the field is growing and has potential implications for psychology and cognitive science [10,11].

[6] M. A. Casey, R. Veltkamp, M. Goto, M. Leman,C. Rhodes, and M. Slaney, “Content-based music information retrieval: Current directions and future challenges,” Proceedings of the IEEE, vol. 96, no. 4,pp. 668–696, 2008.

[7] F. Simonetta, S. Ntalampiras, and F. Avanzini, “Multimodal music information processing and retrieval:Survey and future challenges,” in 2019 international workshop on multilayer music representation and processing (MMRP). IEEE, 2019, pp. 10–18.

[8] S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short termmemory deep neural networks,” Engineering Science and Technology, an International Journal, vol. 24,no. 3, pp. 760–767, 2021.

[9] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton,P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, “Music emotion recognition: A state of the art review,” in Proc. ismir, vol. 86, 2010, pp. 937–952.

[10] Y.-H. Yang and H. H. Chen, “Machine recognition of music emotion: A review,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3,no. 3, pp. 1–30, 2012.

[11] S. Koelsch, “Investigating emotion with music: neuroscientific approaches,” Annals of the New York Academy of Sciences, vol. 1060, no. 1, pp. 412–418,2005.

02. Related Work (Cnt.)

2.4 Word Embedding

Word embeddings convert words into dense vectors that capture semantic and syntactic meanings, derived from large, unlabeled corpora. This helps in understanding word relationships effectively [12].

Word2Vec, developed by Mikolov et al. [13], introduced two models: Continuous Bag of Words (CBOW) and Skip-Gram. These models predict word relationships by learning from the context, which made them influential in the NLP field.

Word embeddings have significantly advanced modern NLP tasks, including sentiment analysis, machine translation, and information retrieval by enabling systems to understand and process text more accurately [14].

[12] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, “Refining word embeddings using intensity scores for sentiment analysis,” IEEE/ACM Transactions on Audio,Speech, and Language Processing, vol. 26, no. 3, pp.671–681, 2017.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean,“Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[14] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,”in Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1, 2015.

03. Data Preparation

3.1 Data Collection and Initial Processing

Data Source: Comments were collected from 27 YouTube videos featuring 20 Sinhala songs, covering various eras of Sinhala music.

Extraction Method: Google App Script was used to extract a total of 93,116 comments using YouTube video IDs.

Initial Data Cleaning: Preprocessing involved removing emojis, URLs, numbers, symbols, and non-Sinhala content to ensure data quality.

Filtered Dataset: After cleaning, 75,656 comments were identified for further analysis, forming the basis for linguistic evaluation.

03. Data Preparation (Cnt.)

3.2 Linguistic Segmentation and Transliteration

Language Segregation: The dataset was compared against the NLTK English corpus to separate purely English comments.

Focus on Sinhala Content: 35,428 comments were identified as containing Sinhala characters and incorporated into the final dataset.

Transliteration Process: Google Transliterator API was used to convert 30,633 English-scripted comments to Sinhala script, successfully transliterating 28,043.

Final Dataset: Post-transliteration, the dataset consisted of 63,471 Sinhala comments, representing 68% of the initial data.

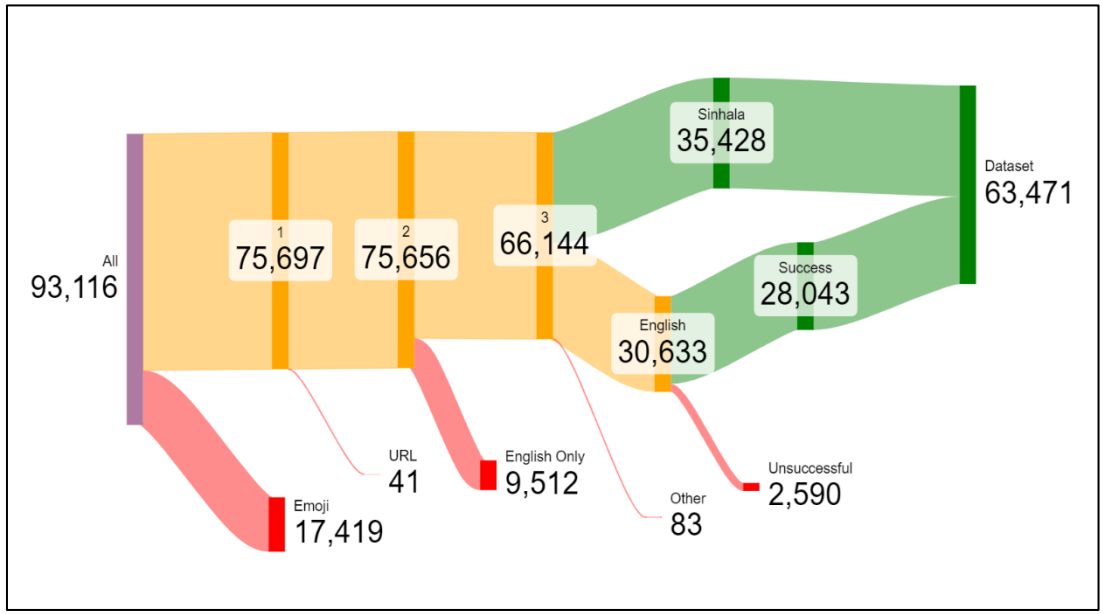

Fig. 1: Stages of the Data Preparation Process

03. Data Preparation (Cnt.)

3.3 Data Composition and Methodological Significance

Dataset Composition: The dataset comprises 63,471 Sinhala comments, with varying engagement across songs.

Notable Insights: High engagement was observed for songs like "අන් සතු ඔබ" (14,100 comments), while others like "පියණුනී" had fewer interactions (130 comments).

Importance of Rigorous Processing: Systematic extraction and thorough data cleaning were crucial in maintaining dataset integrity.

Multilingual Data Handling: The use of the Google Transliterator API enhanced the dataset's linguistic diversity, vital for comprehensive analysis.

Artist	Song	Comment Count
Thisara Weerasinghe	අන් සතු ඔබ	14100
Sarith and Surith	සල්ලි සල්ලි	12254
Dinesh Gamage	දැනෙනා තුරු මා	5378
Methun SK	කාරි නෑ සඳ	3935
Saman Lenin	අමුඬේ	3757
BnS	උන්මාද ප්‍රේම ගීය	3292
Danith Sri	එහෙම දේවල් නෑ හිතේ	2825
Sanuka Wick	පෙරවදනක්	2634
Raini Charuka	කළුවරට හිත බය	2596
Ridma	සොබනා	2058
Prageeth Perera	කෝමලියා	1871
Abisheka and Mihdu	දන්නවාද ආදරේ නීතිය	1722
Sanuka Wick	සරාගයේ	1506
Sandeep Jayalath	නුරාවී	1372
Sanuka	මෝහිනී	1319
Saman Lenin	අම්බරුවෝ	809
Victor Rathnayake	තනිවෙන්නට මගේ ලොවේ	791
Nanda Malini	මා සඳට කැමති බව	735
Victor Rathnayake	අපෙ හැඟුම් වලට	387
Ashanthi	පියාණනී	130

Tab. 1: Comment Counts by Artist and Song

04. Data Analysis

4.1 Overview of Data Analysis Process

Comprehensive analysis of 67,959 unique words in Sinhala YouTube comments.

Rigorous filtration refined dataset to 54,834 unique Sinhala words.

Examined word frequencies and word pairs to identify key linguistic patterns.

Highlighted the importance of understanding domain-specific language use.

04. Data Analysis (Cnt.)

4.2 Insights from Word Frequencies and Word Pairs

Analysis of 328,922 distinct word pairs reveals linguistic diversity.

High-frequency words indicate prevalent themes, e.g., "ලස්සනයි," "සුපිරි."

Top word pairs suggest emotional and aesthetic expressions.

Analysis aids in understanding cultural and emotional nuances in comments.

04. Data Analysis (Cnt.)

4.3 Derivation of Stop Words

Identified 964 frequent words using standardized Z-scores.

Words with z-scores > 3.0 marked as potential stop words.

Compared against existing Sinhala and English stop word lists.

182 words matched English stop words; 53 matched Sinhala stop words.

Process enhances the precision of sentiment analysis in Sinhala.

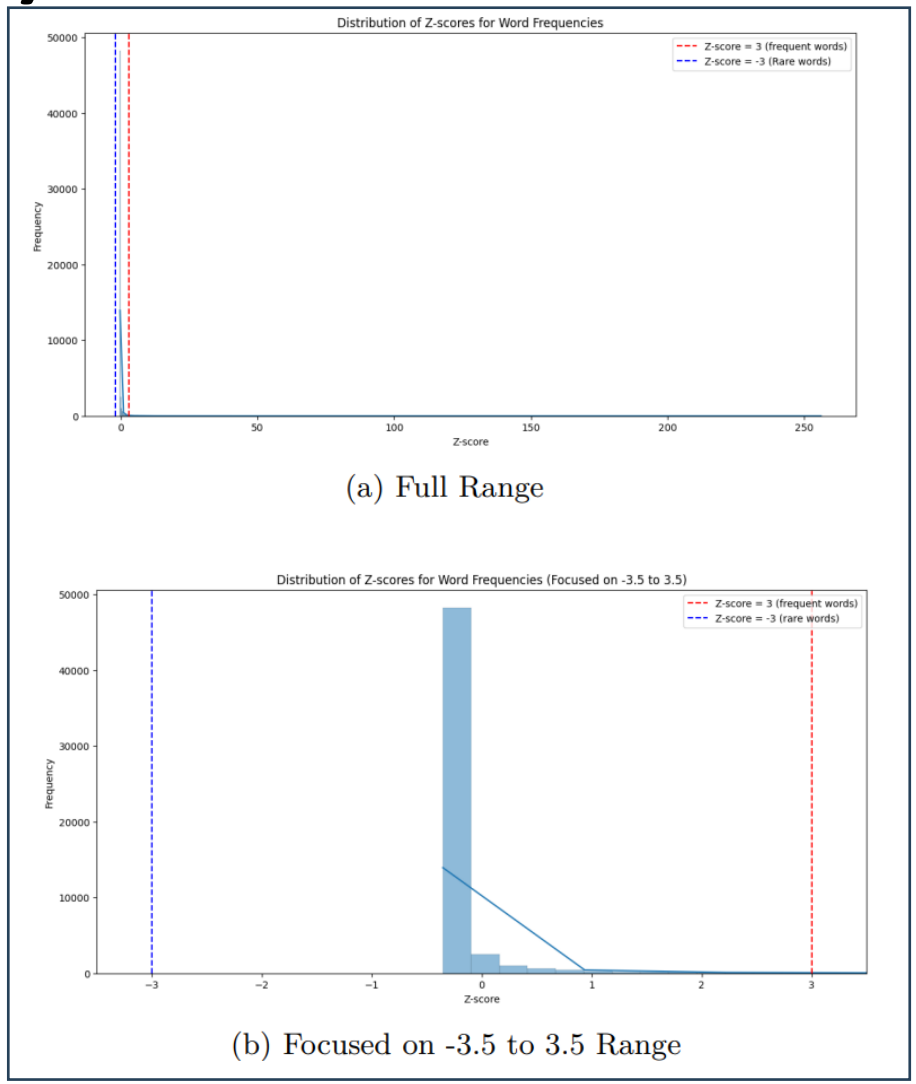

Fig. 2:Distribution of Z Value of Word Frequency

04. Data Analysis (Cnt.)

4.4 General Data Corpus Construction

Wikipedia articles on diverse topics enriched the general corpus.

Included data from Sinhala newspaper articles and government documents.

Comprehensive word corpus of 121,850 unique words was created.

Used a mix of One-Hot Encoding and Word2Vec embeddings for visualization.

Employed t-SNE for dimensionality reduction to observe clustering patterns.

Larger clusters formed by words from news articles and common words across sources.

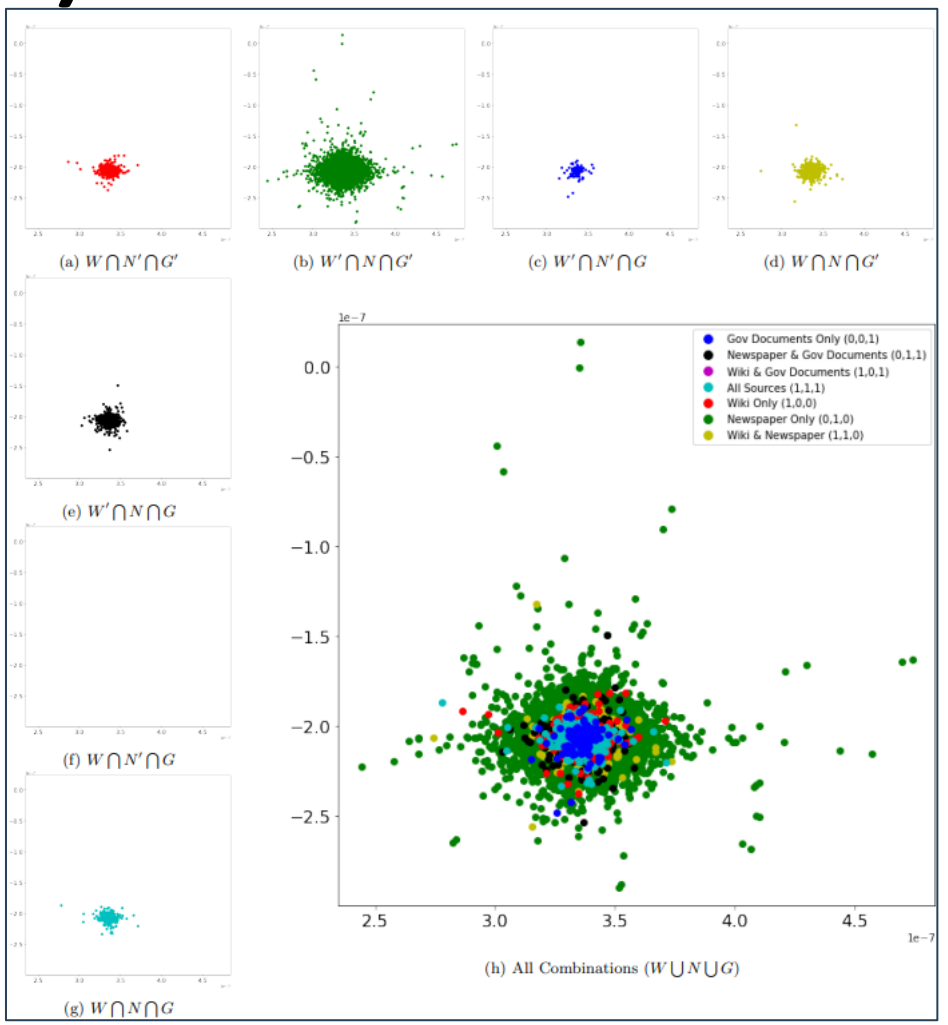

Fig. 3:Word Distribution among Corpora, where the set of words in each of the corpora are denoted as: W= Wikipedia Corpus, N = NEWS Corpus, G= Government Document Corpus

04. Data Analysis (Cnt.)

4.5 Comparison of YouTube Comments with General Domain

36% overlap of unique words between YouTube and general corpus.

YouTube domain has distinctive usage patterns.

Significant overlap suggests representativeness of general usage.

Identified words more frequent in YouTube domain: “සුපිරි,” “ලස්සනයි.”

General domain common words less frequent in YouTube: “සඳහා” “අතර.”

Z-score analysis revealed domain-specific word usage trends.

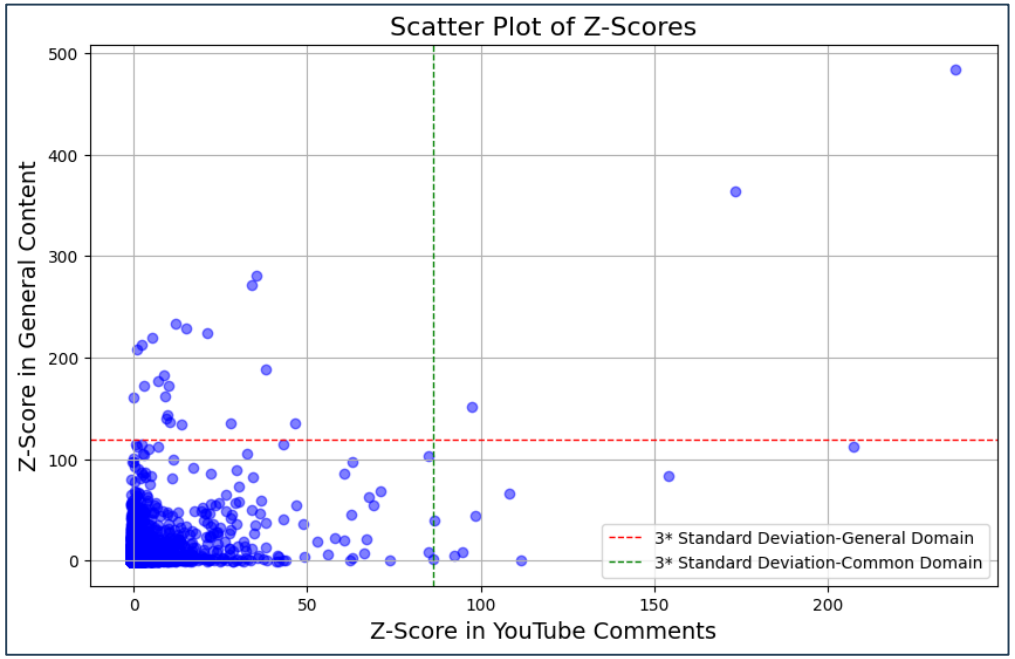

Fig. 4: Scatter Plot of Z-Scores

05. Conclusion

This research uniquely combines MIR, MER, and NLP to explore Sinhala music, filling a gap in the study of underrepresented languages.

Developed a robust dataset of 63,471 Sinhala YouTube comments, capturing diverse songs and eras, which serves as a resource for future studies.

Analysis of social media comments revealed distinct patterns of emotional expression, highlighting the audience's aesthetic appreciation and engagement.

Identified a 36% overlap in unique words between the YouTube comment domain and general Sinhala corpus, showcasing the broader relevance of linguistic patterns.

Comparative analysis of word frequencies revealed dynamic variations in language usage between YouTube comments and general Sinhala content.

The findings and methodologies pave the way for advanced NLP applications and further studies on the intersection of music, language, and emotion in Sinhala culture.

Thank You!

Q & A

References

1. N. de Silva, “Survey on publicly available sinhala natural language processing tools and research,” arXivpreprint arXiv:1906.02358, 2019.

2. D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. De Silva, and G. Dias, “Implementing a corpus for sinhala language,” in Symposium on Language Technology for South Asia, vol. 2015, 2015, p. 3.

3. N. de Silva, “Sinhala text classification: observations from the perspective of a resource poor language,” ResearchGate, 2015.

4. Y. Wijeratne and N. de Silva, “Sinhala language corpora and stopwords from a decade of sri lankan facebook,” arXiv preprint arXiv:2007.07884, 2020.

5. H. Hettiarachchi, D. Premasiri, L. Uyangodage, and T. Ranasinghe, “Nsina: A news corpus for sinhala,” arXiv preprint arXiv:2403.16571, 2024.

6. M. A. Casey, R. Veltkamp, M. Goto, M. Leman,C. Rhodes, and M. Slaney, “Content-based music information retrieval: Current directions and future challenges,” Proceedings of the IEEE, vol. 96, no. 4,pp. 668–696, 2008.

7. F. Simonetta, S. Ntalampiras, and F. Avanzini, “Multimodal music information processing and retrieval:Survey and future challenges,” in 2019 international workshop on multilayer music representation and processing (MMRP). IEEE, 2019, pp. 10–18.

8. S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short termmemory deep neural networks,” Engineering Science and Technology, an International Journal, vol. 24,no. 3, pp. 760–767, 2021.

References

9. Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton,P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, “Music emotion recognition: A state of the art review,” in Proc. ismir, vol. 86, 2010, pp. 937–952.

10. Y.-H. Yang and H. H. Chen, “Machine recognition of music emotion: A review,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3,no. 3, pp. 1–30, 2012.

11. S. Koelsch, “Investigating emotion with music: neuroscientific approaches,” Annals of the New York Academy of Sciences, vol. 1060, no. 1, pp. 412–418,2005.

12. L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, “Refining word embeddings using intensity scores for sentiment analysis,” IEEE/ACM Transactions on Audio,Speech, and Language Processing, vol. 26, no. 3, pp.671–681, 2017.

13. T. Mikolov, K. Chen, G. Corrado, and J. Dean,“Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

14. S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,”in Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1, 2015.