FROM WORDS TO SENTENCES: ADVANCING TURKISH EMOTION ANALYSIS THROUGH EMOTION ENRICHMENT

HANDE AKA UYMAZ

Thesis for Ph.D. Program in Computer Engineering

Graduate School

İzmir University of Economics

İzmir

2023

FROM WORDS TO SENTENCES: ADVANCING TURKISH EMOTION ANALYSIS THROUGH EMOTION ENRICHMENT

HANDE AKA UYMAZ

THESIS ADVISOR: ASSOC. PROF. DR. SENEM KUMOVA METİN

A Ph.D Thesis

Submitted to

the Graduate School of İzmir University of Economics

the Department of Engineering

İzmir

2023

ETHICAL DECLARATION

I hereby declare that I am the sole author of this thesis and that I have conducted my work in accordance with academic rules and ethical behaviour at every stage from the planning of the thesis to its defence. I confirm that I have cited all ideas, information and findings that are not specific to my study, as required by the code of ethical behaviour, and that all statements not cited are my own.

Name, Surname:
Hande Aka Uymaz

Date:
26.12.2023

ABSTRACT

FROM WORDS TO SENTENCES: ADVANCING TURKISH EMOTION ANALYSIS THROUGH EMOTION ENRICHMENT

Aka Uymaz, Hande

Ph.D. Program in Computer Engineering

Advisor: Assoc. Prof. Dr. Senem KUMOVA METİN

December, 2023

The comprehension of language by machines in natural language processing studies poses challenges due to the need for an accurate understanding of language, capturing the true meaning within the data source, and distinguishing emotional nuances. When representing textual data, current word vectorization models are successful in extracting semantic information. However, these models represent words that are often used together as similar to each other in vector space. Thus, words with opposite emotions may have similar vector representations because of their frequent co-occurrence. To overcome such deficiencies in emotion detection, current research focuses on enriching vectors by adding emotional information. In vector enrichment, the fundamental goal is to reproject the vector space to increase the

proximity of words with similar semantic and emotional meanings. This study applies three emotion enrichment models to Turkish words and sentences, using two semantic (Word2Vec and GloVe) and two contextual (BERT and DistilBERT) vectorization methods. Turkish, an agglutinative language by structure, is expected to produce different results than other languages frequently studied in this context. The results demonstrate promising outcomes of enrichment at both the word and sentence levels. Enriched sentence representation was proposed for the first time in the literature in both English and Turkish languages. Moreover, an optimization method involving filtering the emotion lexicons and reducing the dimensionality of the high-dimensional vectors to discern parts containing emotional information is proposed which can be applied to any language and vector model. Experimental results indicate that emotionally enriched vector representations yield better results than original models.

Keywords: Emotion, Sentiment, Emotion detection, Word Embedding, Emotion enriched vectors, Vector space models.

ÖZET

KELİMELERDEN CÜMLELERE: DUYGU ZENGİNLEŞTİRME İLE TÜRKÇE DUYGU ANALİZİNİ GELİŞTİRME

Aka Uymaz, Hande

Bilgisayar Mühendisliği Doktora Programı

Tez Danışmanı: Doç. Dr. Senem KUMOVA METİN

Aralık, 2023

Doğal dil işleme çalışmalarında dilin makineler tarafından anlaşılması, dilin doğru algılanması, veri kaynağındaki gerçek anlamın yakalanması ve duygusal nüansların ayırt edilmesi ihtiyacı nedenleriyle zorluklar içermektedir. Metinsel verileri temsil ederken mevcut kelime vektörleştirme modelleri anlamsal bilgilerin çıkarılmasında başarılıdır. Ancak bu modeller sıklıkla bir arada kullanılan kelimeleri vektör uzayında birbirine benzer şekilde temsil etmektedir. Bu nedenle, zıt duygulara sahip kelimeler, sık sık bir arada bulunmaları nedeniyle benzer vektör temsillerine sahip olabilir. Duygu tespitindeki bu tür eksikliklerin üstesinden gelmek için mevcut araştırmalar, duygusal bilgiler ekleyerek vektörleri zenginleştirmeye odaklanmaktadır. Vektör zenginleştirmede temel amaç, benzer semantik ve duygusal anlamlara sahip

kelimelerin yakınlığını artırmak için vektör uzayını yeniden projekte etmektir. Bu çalışmada, iki semantik (Word2Vec ve GloVe) ve iki bağlamsal (BERT ve DistilBERT) vektörleştirme yöntemi kullanarak üç duygu zenginleştirme modeli Türkçe kelime ve cümlelere uygulanmıştır. Yapı itibariyle eklemeli bir dil olan Türkçenin bu bağlamda sıklıkla çalışılan diğer dillerden farklı sonuçlar üretmesi beklenmektedir. Sonuçlar, hem kelime hem de cümle düzeyinde zenginleştirmenin umut verici sonuçlarını göstermektedir. Zenginleştirilmiş cümle gösterimi literatürde ilk kez hem İngilizce hem de Türkçe dillerinde önerilmiştir. Ayrıca, herhangi bir dil ve vektör modeline uygulanabilen, duygu sözlüklerini filtreleme ve yüksek boyutlu vektörlerin boyutunu azaltarak duygusal bilgi içeren bölümleri belirleme amacını taşıyan bir optimizasyon yöntemi önerilmiştir. Deneysel sonuçlar, duygusal açıdan zenginleştirilmiş vektör temsillerinin orijinal modellerden daha iyi sonuçlar verdiğini göstermektedir.

Anahtar Kelimeler: Duygu, Duygu tespiti, Kelime temsili, Duygu zenginleştirilmiş vektörler, Vektör uzay modelleri.

This thesis is dedicated to my all, Erdem and my precious son Ege.

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my advisor, Assoc. Prof. Dr. Senem KUMOVA METİN, for being a constant source of inspiration, for her guidance, and for illuminating my path throughout my doctoral thesis journey. I am truly grateful for her continuous encouragement, attentive listening, and invaluable assistance, which have played a pivotal role in shaping my studies.

I would like to express my gratitude for their contributions to my thesis monitoring processes to Assoc. Prof. Dr. Tarık KIŞLA and Assoc. Prof. Dr. Kaya OĞUZ, who were part of my jury, as well as Prof. Dr. Bahar KARAOĞLAN and Asst. Prof. Dr. İbrahim ZİNCİR, who also participated in my jury and made valuable contributions.

I am grateful to my father, Dr. Murat Tamer AKA, for teaching me to compare myself only with my own achievements and always believing in me; to my mother, Filiz AKA, for guiding me back to my strengths whenever I felt helpless, reminding me of what I can accomplish, and always supporting me; and to my sister, Gözde AKA BAL, for always believing in and understanding me.

Finally, I cannot thank enough to my dear husband Mehmet Erdem UYMAZ, who has always been with me on this journey. I am grateful for his constant support, attentive listening, calming responses to my endless questions, and even for believing in me more than I do. My dear son, the source of my joy, my precious Ege UYMAZ, I'm glad that this process proceeded with you because your presence has always been a motivation for me. Whenever I felt tired or overwhelmed, everything became more manageable with your father and you.

TABLE OF CONTENTS

ABSTRACT iv

ÖZET vi

DEDICATION viii

ACKNOWLEDGEMENT ix

TABLE OF CONTENTS x

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xv

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: EMOTION AND SENTIMENT ANALYSIS 10

2.1. Emotion Models 11

CHAPTER 3: LITERATURE REVIEW 15

3.1. Text-Based Emotion and Sentiment Analysis and its Challenges 15

3.1.1. Categorization of Emotion/Sentiment Detection Methods 17

3.1.2. Datasets and Lexicons 21

3.1.2.1 Datasets 21

3.1.2.2 Lexicons 27

3.1.3. Emotion/Sentiment Detection Evaluation Metrics 30

3.2. Vector Space Models 32

3.2.1. Emotion Enriched Vectors 43

3.2.2. Sentimental Vectors 50

CHAPTER 4: ENRICHING VECTORS WITH EMOTIONAL CONTENT 57

4.1. Enriching Word and Sentence Vectors 58

CHAPTER 5: EMOTION ENRICHMENT EXPERIMENTS 60

5.1. Experimental Setup 60

5.1.1. Embedding Models 60

5.1.2. Datasources 62

5.1.3. Emotion enrichment models 64

5.2. Experimental Results 70
5.2.1. Similarity Measurements and Classificaton 70
5.2.2. Word-Level Emotion Enrichment Experiments 71
5.2.2.1 Word Level Cosine Similarity Measurements 72
5.2.2.2 Classification 77
5.2.3. Sentence-Level Emotion Enrichment Experiments 82
5.2.3.1 Sentence Level Cosine Similarity Measurements . . 83
5.2.3.2 Classification 86
5.2.3.3 Comparison of Sentence-level and Word-level Classification 89
CHAPTER 6: OPTIMIZING EMOTION ENRICHMENT: DIMENSIONALITY REDUCTION AND LEXICON FILTERING 91
6.1. Lexicon Filtering 92
6.2. Dimensionality Reduction 98
6.2.1. Experimental Results 99
CHAPTER 7: CONCLUSION 107
REFERENCES 131
APPENDICES 131
Appendix A - Pairwise cosine similarity histograms (Word2Vec and EEA1_Word2Vec) 132
Appendix B - Pairwise cosine similarity histograms (GloVe and EEA2_GloVe) 133
Appendix C - Pairwise cosine similarity histograms (Word2Vec and EEA2_Word2Vec) 134
Appendix D - Pairwise cosine similarity histograms (GloVe and EEA3_GloVe) 135
Appendix E - Pairwise cosine similarity histograms (Word2Vec and EEA3_Word2Vec) 136
CURRICULUM VITAE 137

LIST OF TABLES

Table 1. Datasets from Literature. 22
Table 2. Lexicons from Literature. 28
Table 3. Word vectors for example vocabulary *V*. 33
Table 4. Emotion Enriched Vectors. 44
Table 5. Sentimental Vectors..... 50
Table 6. The statistics of TT-NRC. 62
Table 7. Utilized datasets. 63
Table 8. The mean variation in similarity scores between word pairs within four emotion categories when comparing the original word embeddings with their emotionally enhanced counterparts. 73
Table 9. Average of in-category and opposite-category similarity scores. 77
Table 10. Accuracy and F1-scores for a weighted average across four emotions (anger, fear, joy, sadness) are presented, with the top accuracy scores highlighted in bold and the highest F1-scores underlined for each model. 78
Table 11. F1-scores of *sadness* emotion. 79
Table 12. F1-scores of *joy* emotion. 80
Table 13. In-category similarity scores - Enriching sentences with emotional sentences. 83
Table 14. In-category similarity scores - Enriching sentences with original and emotion-enriched lexicon words. 85
Table 15. Classification F1 scores - Enriching sentences with emotion sentences. 87
Table 16. Classification F1 scores - Enriching sentences with original and emotion-enriched lexicon words. 88
Table 17. The outcomes of the 5-fold cross-validated paired t-test comparing the top two performing configurations in experiments involving emotion sentences and emotion lexicon words. 89
Table 18. F1 scores for classification at the word-Level experiments 90

Table 19. Changes in the number of words in four emotion categories and the amounts of increase in average cosine similarity (CS) values within each category based on different threshold values for English Lexicon Words. 96
Table 20. Changes in the number of words in four emotion categories and the amounts of increase in average cosine similarity (CS) values within each category based on different threshold values for Turkish Lexicon Words. 97
Table 21. Change in English data set size and average in-category cosine similarity scores for all emotion categories. 98
Table 22. Pairwise in-category and opposite-category cosine similarity results of *English words* while using only one window. 100
Table 23. Pairwise in-category and opposite-category cosine similarity results of *Turkish words* while using only one window. 100
Table 24. Pairwise in-category and opposite-category cosine similarity results of *Filtered English words* while using only one window. 101
Table 25. Pairwise in-category and opposite-category cosine similarity results of *Filtered Turkish words* while using only one window. 101
Table 26. % increase in cosine similarity per window after filtering English lexicon words. 102
Table 27. % increase in cosine similarity per window after filtering Turkish lexicon words. 102
Table 28. English sentence embeddings enrichment with several combinations. (The best results are shown in bold.) 105
Table 29. Turkish sentence embeddings enrichment with several combinations. (The best results are shown in bold.) 105
Table 30. Turkish and English sentence embeddings enrichment of best-performing combinations. 110

LIST OF FIGURES

Figure 1. Plutchiks wheel of emotions (Acheampong et al., 2020), (Plutchik, 1980) 13
Figure 2. Russell's circumplex model. (Acheampong et al., 2020), (Russell, 1980) 13
Figure 3. Word2Vec model architectures (Mikolov et al., 2013). 37
Figure 4. Framework for the *word-level emotion enrichment* experimental study. 68
Figure 5. Framework for the *sentence-level emotion enrichment* experimental study. 69
Figure 6. Pairwise cosine similarity histograms (GloVe and EEA1_GloVe). 72
Figure 7. Pairwise similarity histograms (BERT - EEA1_BERT vectors) 74
Figure 8. Pairwise similarity histograms (BERT - EEA2_BERT vectors). 75
Figure 9. Pairwise similarity histograms (BERT - EEA3_BERT vectors). 75
Figure 10. The framework of lexicon filtering procedures. 94
Figure 11. Framework for vector partitioning with sliding window technique. 99
Figure 12. Framework for extracting sub-vectors. 104
Figure 13. The framework of the best-performing sentence level emotion enrichment with optimization. 109
Figure 14. Pairwise CS histograms (Word2Vec and EEA2_Word2Vec). 132
Figure 15. Pairwise CS histograms (GloVe and EEA2_GloVe). 133
Figure 16. Pairwise CS histograms (Word2Vec and EEA2_Word2Vec). 134
Figure 17. Pairwise CS histograms (GloVe and EEA3_GloVe). 135
Figure 18. Pairwise CS histograms (Word2Vec and EEA3_Word2Vec). 136

LIST OF ABBREVIATIONS

ANEW	: Affective Norms of English Words
BERT	: Bidirectional Encoder Representations from Transformers
Bi-LSTM	: Bi-directional Long Short-Term Memory
CF	: Configuration
CL	: Classifier
CMA-ES	: Covariance Matrix Adaptation Evolution Strategy
CNN	: Convolutional Neural Network
CS	: Cosine Similarity
CVAW	: Chinese Valence Arousal Words
DNN-SM	: Deep Neural Network Softmax Dense Layer
DUTIR	: Dalian University of Technology Information Retrieval
E-ANEW	: Extended Affective Norms of English Words
EEA	: Emotion Enrichment Approach
ELMo	: Embeddings from Language Models
EmoLex	: NRC Word-Emotion Association Lexicon
FN	: False Negatives
GloVe	: Global Vectors for Word Representation
IDF	: Inverse Document Frequency
ISEAR	: International Survey on Emotion Antecedents and Reactions
LLR	: Linear Logistic Regression
MLP	: Multilayer Perceptron
NLP	: Natural Language Processing
NLPCC	: Natural Language Processing and Chinese Computing
OOV	: Out-of-Vocabulary
PCA	: Principal Component Analysis

RQ : Research Question

SEL : Spanish Emotion Lexicon

SEND : Stanford Emotional Narratives Dataset

SMO : Sequential Minimal Optimization

SO-CAL : The Semantic Orientation Calculator

TEC : Twitter Emotion Corpus

TEI : Tweet Emotion Intensity

TEL : Turkish Emotion Lexicon

TF : Term Frequency

TP : True Positives

TS-GRU : Tree-Structured Long Short-Term Memory

TT-NRC : Turkish Translated NRC Emotion Lexicon

ULMFit : Universal Language Model Fine Tuning

VADER : Valence Aware Dictionary for sEntiment Reasoning

WEKA : Waikato Environment for Knowledge Analysis

CHAPTER 1: INTRODUCTION

Emotion is a concept that represents the feelings of individuals experiencing various events. While people can have similar emotional responses to similar situations, they can also feel completely independent emotions. In addition, the expression of emotions can vary depending on factors such as culture, ethnic group, personality, gender, or geographical location. For example, a joke that might amuse people from a certain culture or living in a particular region may not be funny to someone from a different culture. In some cultures, showing positive emotions like laughter and cheerfulness may be considered embarrassing or improper, while the excessive expression of negative emotions (e.g., sorrow, lament, etc.) may be more acceptable. Due to these reasons, understanding emotions among people is often a complex and challenging process. On the other hand, with the advancements in artificial intelligence and natural language processing, significant strides have been made in enabling machines to better understand human emotions.

While there are approximately 7,000 languages in the world, the majority of studies focus on the most spoken language, English. However, considering that language is a communication tool reflecting the cultural values and norms, history, and lifestyle of a community, the understanding of each language by a machine may involve different stages, and language dependent studies may produce more effective results.

In daily life, we can analyze human emotions in many ways, including gestures and facial expressions, voice, images, and texts. Internally, when we consider human feelings, we can actually break down emotions into many sub-categories (e.g., shyness, enthusiasm, obsession). Looking back in history, emotion has been a subject of research by scientists from various fields such as neuroscience, sociology, and psychology, starting with Charles Darwin's book "The Expressions of the Emotions in Man and Animals", which explores the relationship between emotions and evolution (Darwin, 1872). It is challenging to categorize the concept that makes us human into limited categories in a field where researchers from various fields express their ideas.

Therefore, emotion detection studies generally rely on classifying emotions according to various emotion theories rather than a single definition. In short, researchers in the field attempt to categorize shared emotions, such as happiness, sadness, fear, anger, and surprise, by examining data sources within the frameworks of emotion theories provided by scientists like (Ekman, 1992) and (Plutchik, 1980).

Natural language processing (NLP) is a subfield of artificial intelligence that deals with the understanding, interpretation, and even response to human language by computers. In other words, NLP is a branch of science that focuses on better understanding complex human language, extracting hidden meanings, analyzing sources such as speech and text, and transferring this information and skills to computers. These applications can be used in various fields such as text mining, language understanding, speech recognition, text and speech synthesis, language-based user interfaces, translation, emotional analysis, and many more. Algorithms are developed by examining massive datasets in the literature to teach machines the structure and uses of native languages. Sentiment/emotion analysis, a subfield of natural language processing, works to analyze hidden emotions in expressions. Although closely related concepts, emotion analysis entails recognizing and comprehending the diverse emotions conveyed in data sources to ascertain the emotional state of individuals or groups. On the other hand, sentiment analysis is centered on evaluating the general sentiment whether positive, negative, or neutral in a given text. Emotion analysis methods have a wide range of applications, from customer relations to human resources management, and education to the health sector. This technology allows companies to assess customer satisfaction, improve human resources processes, effectively monitor student performance in education, and understand and personalize the emotional states of patients in the health sector. Additionally, emotion recognition technologies can be used across a broad spectrum, making interactions with technology more sensitive and personalized, from virtual assistants to automation systems.

Text, which has been one of the most important communication tools for centuries, is a very important data source for sentiment analysis studies. Additionally, with the rapid development of social media, access to, storage, and analysis of textual data in a computer environment have become even easier. Text data can be analyzed for various

reasons.

For instance, companies can leverage emotion analysis to examine product reviews and user feedback, obtaining valuable insights. Additionally, by evaluating customer service interactions, such as email, chat, or call records, companies can gain a nuanced understanding of feedback, contributing to enhanced customer satisfaction and improved service quality. This approach is a valuable tool for companies to discern positive or negative emotional responses, fostering deeper insights into customer experiences and fortifying customer relationships. Similarly, media organizations can utilize emotion analysis to assess general reader reactions to news articles or social media posts, providing critical insights into the public perception of news and its broader impact.

There are various techniques to distinguish emotion or sentiment in a text, including machine learning and lexicon-based methods. However, before applying these methods, text data belonging to natural languages with complex rules or vocabulary need to be transformed into numerical forms, a process called "vectorization".

In the field of NLP, various models exist for projecting texts into a vector space. These models can be based on frequency (e.g., tf-idf), semantic information (e.g., GloVe (Pennington et al., 2014)), or contextual information (e.g., BERT (Devlin et al., 2018)). This concept, called vector representation, involves using a vector of real numbers of fixed length to represent a text unit, such as a word or a sentence, in NLP subfields like machine translation and text classification. Here, a "text unit" is the smallest language unit that can be analyzed in a text and is usually a word, a phrase, or a sentence.

Classical or semantic vectorization techniques like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) capture semantic information by representing individual words as distinct entities in the vector space. In this vectorial representation, words that are closely positioned to each other are considered to exhibit semantic closeness. In contrast, contextual vectorization methods like BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018) take into account the context of a word or a text unit. These contextual embeddings create vector representations by considering polysemy, where a word may acquire different meanings in different sentences. This

perspective is in line with Ludwig Wittgenstein's (1889–1951) idea that "the meaning of a word is its use".

While vector space models rely on the assumption that words with similar contexts have similar meanings, they may face challenges in the field of emotion or sentiment detection. For example, words like "happy" and "sad", despite representing opposite emotions, may have high similarity scores in the vector space since they are frequently found in similar contexts. Recent studies in the field of emotion/sentiment detection propose improvements to original vector representations that include emotional and/or sentimental information along with semantic or contextual information. These representations are called emotion-enriched or sentimental representations.

Emotion enrichment aims to incorporate emotional content into the representation of the text units by utilizing samples labeled with emotions. The primary step in this enrichment process entails the comparison between the vector representation of a text unit without an emotion label with the representations of labeled samples, such as emotion lexicon words. This comparison is conducted to identify the closest and thus most similar labeled samples to the unlabeled ones. Alternative approaches to this procedure often vary in the number of closest samples utilized for emotion enrichment, incorporate additional information such as emotional intensity scores, use different distance metrics (e.g. Cosine distance, Euclidean distance) for comparison, and adopt different methods for conducting the enrichment operation (e.g., Alshahrani et al., 2019; Mao et al., 2019; Agrawal et al., 2018; Wang and Meng, 2018; Su et al., 2018).

In this thesis, our emphasis was on words and sentences as text units, intending to generate vectors encompassing emotional information for utilization in emotion detection studies. Our study is based on Plutchik's emotion model which has 8 different emotion categories which are anger, fear, sadness, joy, surprise, trust, disgust, and anticipation. Our specific focus was on the Turkish language, ranked among the top 20 most spoken languages globally by various sources. Despite its prominence, Turkish has received comparatively less attention in the field of emotion analysis, with limited available data resources.

Turkish language, which belongs to the Altaic language family, has its unique features and challenges. Being an agglutinative language, Turkish involves the addition

of affixes to root words for conveying grammatical information. This complexity can pose challenges in tokenization and morphological analysis, especially when compared to languages with simpler structures. It is common in Turkish to construct lengthy words that might correspond to a sentence with multiple words in English. Take, for example, the term "gelmeliymişsinizcesine", which translates to "as if you should have come" in English. Given the lack of constraints on the number of suffixes that can be added, the language allows for the formation of words of variable lengths, adapting to the desired meaning. Moreover, Turkish utilizes a Subject-Object-Verb word order in sentences, which differs from the Subject-Verb-Object order observed in languages such as English. This distinction in word order may influence the efficacy of models trained in other languages when applied to Turkish.

The main research questions of our study, in which we investigated the effects of emotion enrichment on the Turkish language, are as follows:

RQ1 - What is the most efficient original word/sentence embedding method for enhancing the detection of emotions in Turkish texts, thereby improving the performance of emotion detection studies?

To address the initial research question, we employed GloVe and Word2Vec for semantic embeddings, and BERT for contextual embedding in word-level experiments. While GloVe and Word2Vec generate static embeddings that capture the semantic relationships between words, BERT offers dynamic embeddings capable of distinguishing polysemy. For sentence-level word representation, we utilized BERT and DistilBERT contextual embeddings.

RQ2 - Can enhanced representations of words and sentences outperform their original counterparts?

The objective of creating emotion-enriched vectors is to improve the effectiveness of emotion analysis studies. The primary goal is to cluster the vectors closely in the vector space, considering emotion categories, while also encompassing the meanings and/or contexts of the target text units. This research question investigates whether the performance of emotion-enriched vectors will exhibit enhancement in the Turkish dataset as opposed to the original vectors.

RQ3- Is the efficacy of original and enhanced representations subject to variation

based on emotion categories?

While polysemy is a common linguistic phenomenon, some words possess a single dominant meaning. Likewise, certain words may predominantly express specific emotions or be frequently employed for that purpose. The concentration of such words in conveying a particular emotion can enhance the identification of that emotion in a given text, making it more noticeable than others which may influence the success of emotion enrichment processes.

RQ4 - Which emotion enrichment methods give better results on word-level and sentence-level emotion detection?

Firstly, we conducted comparative analyses of three emotion enrichment methods at both the word and sentence levels, evaluating their cosine similarity scores and performance in classification experiments. The study systematically explored the differences between these methods, providing insights into their distinct characteristics and effectiveness.

When reviewing studies in the literature, we observe that emotion/sentiment enrichment has been primarily conducted at the word level, with various methods proposed. In this study, we suggest extending the enriched vector approach to the sentence representation dimension by representing sentences not through their constituting words but as a whole, followed by the application of the enrichment process. Furthermore, we aim to analyze whether there is a difference in the methods utilized between word-level and sentence-level emotion enrichment in terms of enrichment method selection or parameters.

RQ5 - In the context of representing emotionally enriched sentence vectors, how can we improve precision and effectiveness by optimizing the computational efficiency of vectors and refining emotion lexicons, while taking into account linguistic nuances in both Turkish and English languages?

We aimed to address the computational demands posed by 768-dimensional BERT vectors, especially when dealing with extensive datasets. Implementing a sliding window technique, we systematically analyzed consistent patterns within these vectors to improve computational efficiency. This is not only for reducing cost but also for having insights into the nuanced integration of emotions within language

representations.

Concurrently, we recognized the challenges within emotion lexicons, such as word categorization per emotions and the inclusion of terms with ambiguous emotional associations. Acknowledging the impact of cultural nuances, we conducted experiments to filter emotion lexicon words more accurately. Emphasizing the importance of contextual usage, we aimed to capture connections between words' general contexts and their emotional usage in diverse datasets.

Our research, extending across both Turkish and English languages, focused on reducing BERT vectors' dimensionality and refining emotion lexicons to achieve more precise and effective results in representing emotionally enriched sentence vectors. By integrating computational efficiency with nuanced linguistic considerations, our study contributes to a comprehensive understanding of emotion representation in language.

In this study, the effectiveness of the emotion-enrichment process is evaluated through similarity and classification experiments. The first experiment involves measuring the average change in cosine distance within emotion categories using pairwise cosine similarity scores. The expected outcome is a decrease in distance between emotionally enriched representations, indicating improved similarity within the same emotional category. In the second approach, emotion identification is treated as a classification task, categorizing text units (word or sentence) enriched by various embedding methods into predefined emotion categories. The study compares the effectiveness of semantic embeddings (GloVe and Word2Vec) with BERT as contextual embeddings in classifying words and BERT and DistilBERT embeddings for sentences.

This thesis, conducted through detailed investigations and experiments addressing the research questions, makes contributions to the field that can be summarized as follows:

- The research categorizes and summarizes frequently employed datasets in text-based emotion detection, compares lexicons formed or utilized for studies focused on sentiment and emotion detection using lexicon-based approaches, and offers a summary of methods proposed in the literature to enrich vector representation based on emotional and sentiment information. These elements

are thoroughly examined in Aka Uymaz and Kumova Metin (2022), and details can be located within Chapter 3 of the thesis.

- Three emotion enrichment methods, the details of which were explained in Chapters 4 and 5, were applied to a Turkish dataset, and their success was measured and compared. As far as we know, emotion-enriched vectors were applied for the first time in Turkish in Aka Uymaz and Kumova Metin (2023a) and Aka Uymaz and Kumova Metin (2023b).

- While word-level emotion enrichment has always been applied in the literature, with this study (in Chapter 5.2.3), sentence-level emotion enrichment is proposed using different parameters to both Turkish and English sentence vectors.

- The effectiveness of enriching sentences with emotion-lexicon words has been the focus, emotion enrichment at the sentence level is aimed to optimize in Chapter 6. Calculation demands of 768-dimensional BERT vectors are addressed, and hidden emotional cues of specific dimensions are explored. By using the sliding window technique, our approach aims to enhance computational efficiency and provide new perspectives on emotion representation.

- In response to potential issues within emotion lexicons, a method to create a more refined and accurately categorized emotion lexicon is proposed in Chapter 6. By filtering emotion lexicon words based on contextual differences, we aimed to improve the accuracy of emotion lexicons.

This thesis proceeds with the following chapters: Chapter 2 presents the background, encompassing definitions related to emotion/sentiment analysis. Chapter 3 provides detailed information about the utilized data sources, vector space models, emotion/sentiment enrichment methods, and evaluation metrics utilized in the previous work. Chapter 4 defines word-level and sentence-level emotion enrichment. In Chapter 5, we present the details of the enrichment procedures starting from the utilized embedding and enrichment models and data sources to experimental results in two phases of experiments: cosine similarity measurements and classification for both.

Chapter 6 provides the dimensionality reduction of BERT vectors and lexicon filtering. Finally, Chapter 7 presents the conclusion.

CHAPTER 2: EMOTION AND SENTIMENT ANALYSIS

Individuals can manifest their emotional responses to events in a variety of ways, encompassing facial expressions, body language, written communication, and spoken words. The process of analyzing emotions involves the identification of these emotions through diverse means, including audio recordings, videos, textual content, images, and EEG signals (Baali and Ghneim (2019), Lech et al. (2020), Calvo et al. (2020), García-Martínez et al. (2021)). Nevertheless, discerning the precise human emotion being conveyed poses a formidable challenge, both for individuals and computer systems. This challenge arises from the fact that people can convey the same emotion in varying ways or may simultaneously express multiple emotions (Sailunaz et al., 2018). Furthermore, the expression of emotions can vary depending on factors such as culture, ethnicity, personality, gender, or geographical location (Sailunaz et al., 2018). For example, previous research has indicated that Asian cultures tend to exhibit lower levels of life satisfaction and a greater prevalence of negative emotions in comparison to North American culture (Scollon et al., 2004).

While the terms *emotion* and *sentiment* are often used interchangeably as synonyms, they carry distinct connotations within the field of natural language processing. In this context, *emotion* denotes more precise and intense emotional states such as *love* and *frustrated*, whereas *sentiment* typically encompasses a narrower spectrum, categorizing emotions into generally three primary groups: *positive, negative*, and *neutral*. To illustrate, the emotional category of *love* is associated with a *positive* sentiment, while *frustrated* is linked to a *negative* sentiment. Consequently, emotion detection primarily involves identifying specific emotions within a given data source, relying on various emotion models. Conversely, sentiment analysis aims to capture the overall emotional tone of a data source, encompassing polarity information.

The widespread use of the internet and social media has opened up new avenues for individuals to share their emotions and opinions about various events, products, or services. In today's interconnected world, people can express their feelings through

text, emojis, images, and videos across digital platforms. Researchers in the field of emotion and sentiment analysis are actively exploring this expansive digital space, recognizing that text serves as one of the primary means of communication on the internet. They aim to understand how individuals convey their emotions in the online sphere and to uncover valuable insights into public sentiment, consumer preferences, and societal trends. As online communication methods continue to evolve, the research on emotion/sentiment detection remains highly pertinent, offering a unique view into the collective emotional responses of the global community. In order to provide the necessary background, the upcoming subsection briefly presents the emotion models.

2.1. Emotion Models

Emotion has been a subject of study across various academic disciplines, including psychology, neuroscience, and sociology, with its exploration tracing back to Charles Darwin (Darwin, 1872). This broad spectrum of fields and diverse approaches has made it challenging to arrive at a universally accepted definition. Emotion researchers tend to approach the study of emotions from various angles, leading to the emergence of multiple approaches rooted in a classification framework instead of a singular definition. This classification approach, which enables the differentiation or comparison of emotions based on various emotion theories, is commonly known as an emotion model. Existing emotion models can be broadly categorized into two groups, aligning with different emotion theories: categorical and dimensional models (Sailunaz et al., 2018).

In categorical models, emotions are classified into distinct categories, such as *happiness*, *sadness*, and *anger*. For instance, Shaver et al. (1987) established a framework consisting of six emotional categories (*sadness*, *love*, *joy*, *anger*, *surprise*, and *fear*) to differentiate emotions in everyday life contexts. Similarly, Ekman (1992) delineated six basic emotions, namely *fear*, *anger*, *joy*, *disgust*, *sadness*, and *surprise*. When describing these basic emotions, Ekman (1992) introduced specific associated characteristics, including, distinctive physiology, distinctive universals in antecedent events and distinctive universal signals. Based on these criteria, some emotions do not meet the criteria to be considered basic emotions within Ekman's framework, as they

lack certain distinguishing characteristics that set them apart from other mental states (Ekman and Cordaro, 2011). For instance, although *love* is included in Shaver's model, it is not included in Ekman's list of basic emotions due to the absence of a consistently associated facial expression (Sabini and Silver, 2005). In the study of Sabini and Silver (2005), this distinction is illustrated through two different expressions of the same emotion, namely *parental love*. For instance, a parent conveys their affection through a smile when reciprocating their child's smile, but also manifests a sense of concern through a look of distress when the child is facing difficulty.

Many emotion datasets are constructed primarily based on categorical models, with Ekman's well-known six-category model being the most frequently adopted due to its practicality in gathering training data (Tahon et al., 2018). In contrast, Plutchik's model extends the emotional spectrum to include eight distinct emotion categories, incorporating two additional dimensions, *trust* and *anticipation*, each characterized by varying degrees of intensity, effectively augmenting Ekman's emotional framework (Plutchik, 1980). Plutchik's definition introduces eight emotions, featuring pairs of opposing emotions such as *trust* versus *disgust* and *sadness* versus *joy*. Despite the variability in the number of emotion categories across different models, the representation of emotions as discrete categories is argued to be more comprehensible for individuals (Alshahrani, 2020). As a result, categorical models remain widely utilized in the field of emotion research.

In a subset of emotion detection studies (e.g., Agrawal et al. (2018)) that employ categorical emotion models, a given text may receive multiple emotion labels. From this perspective, categorical emotion models can be categorized into two distinct groups. The first category is known as the *single label emotion model*, where each text unit is associated with a single emotion label. In contrast, the second category, referred to as the *multi-emotion model*, allows for the assignment of multiple emotion labels to a given text unit, implying that constituent words or terms of a sentence may simultaneously exhibit more than one emotion. Within the *multi-emotion model*, each word or term is represented by an emotion vector, where index i of the vector indicates the degree of association with an emotion, typically derived from a lexicon.

In contrast to categorical models of emotions, which classify emotions into distinct

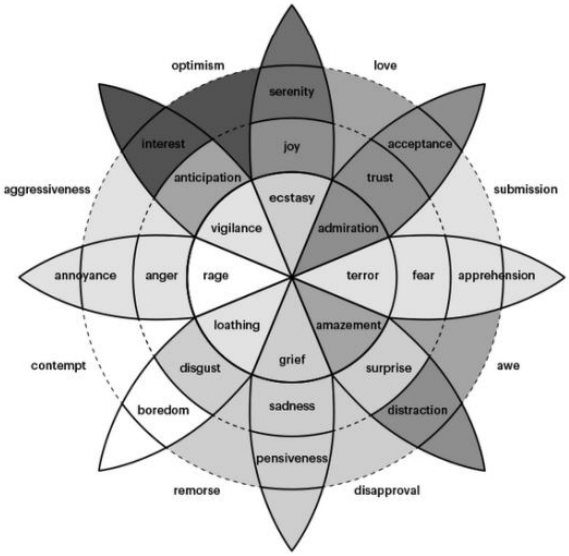


Figure 1. Plutchiks wheel of emotions (Acheampong et al., 2020), (Plutchik, 1980)

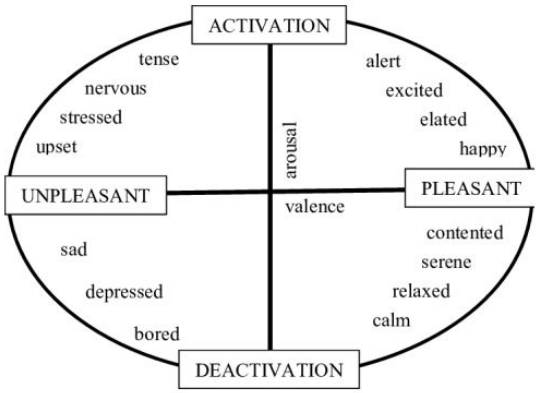


Figure 2. Russell's circumplex model. (Acheampong et al., 2020), (Russell, 1980)

categories, dimensional emotion models suggest that emotions are more appropriately depicted along a spectrum within a dimensional space, typically utilizing two or three key dimensions. These commonly employed dimensions include *valence*, *arousal*, and *dominance*. *Valence* represents whether an emotion is positive or negative. *Arousal* measures the intensity or excitement associated with an emotion. *Dominance* indicates the level of control or influence a person has over their emotion (Calvo and Kim, 2012), (Sreeja and Mahalakshmi, 2017). These dimensional models aim to capture the nuanced relationships and interplay between different emotions by positioning them along these defined dimensions within the emotional space.

An example of a well-known dimensional model is Plutchik's Wheel of Emotions,

which illustrates emotions within a two-dimensional space defined by their valence and arousal levels (Plutchik, 1980; Acheampong et al., 2020). In the presented model (as illustrated in Figure 1), the expression of basic emotions involves a hierarchical arrangement within nested circles, signifying different degrees of emotional intensity. Within this model, each emotion is characterized by three distinct intensity levels. For example, *rage*, situated at the center of the wheel, exhibits a higher level of intensity when compared to *anger*, which, in turn, displays greater intensity than *annoyance*, positioned in the outermost ring of the model. The circumplex emotion model of affect is another example of the dimensional emotional model that is used to classify emotions by considering two core dimensions: *valence* (ranging from pleasant to unpleasant) and *arousal* (spanning from activation to deactivation) as can be seen in Figure 2. It organizes emotions within a circular diagram, thus offering a structured framework for the analysis of emotional states concerning these dimensions, as demonstrated in various studies (e.g., Perikos and Hatzilygeroudis (2016), Hasan et al. (2014)). Russell and Mehrabian (1977) introduced the Pleasure-Arousal-Dominance (PAD) model as an alternative to earlier models, which focused solely on pleasure and arousal. This model includes three dimensions: *Pleasure*, *Arousal* and *Dominance*, with *dominance* being the newly added third dimension. Researchers have applied the PAD model in various studies (e.g., Stojanovska et al. (2018) Gao et al. (2016)).

CHAPTER 3: LITERATURE REVIEW

In this chapter, we thoroughly explore the existing research that inspires our study. Our primary focus is placed on two critical aspects while examining the previous work: emotion and sentiment analysis from text, as well as the use of vector space models. Our aim in reviewing this literature is to establish a strong basis for our research and acquire a clear understanding of the key concepts and methodologies.

Starting with Section 3.1, firstly text-based emotion detection is defined, addressing the associated challenges. We then move on to examine the various methods employed by researchers for the detection and analysis of emotions (Section 3.1.1). Additionally, the diverse sources of data are explored that have been used in these analyses (Section 3.1.2). This examination enables us to gain a comprehensive understanding of the wide range of techniques and data sources applied by researchers in this field.

In Section 3.2, vector space model applications in representing textual data are investigated. Specifically, in Sections 3.2.1 and 3.2.2, we focus on emotion-enriched and sentimentally enriched vectors.

3.1. Text-Based Emotion and Sentiment Analysis and its Challenges

Writing serves as a primary means of conveying our thoughts and emotions, a practice that has expanded significantly with the advent of social media. On platforms like Twitter, blogs, and in comments on a product or a service, individuals express their emotional and sentimental states. This trend has sparked interest among researchers who have developed various methods for discerning sentiments and emotions in social media texts (e.g., Sarsam et al. (2021), Zhang et al. (2019), Gaind et al. (2019), Zimbra et al. (2018)). However, determining the emotional content of text poses challenges. Texts can contain multiple emotions, and certain words may have multiple meanings that correspond to different emotions. Identifying words or phrases that carry emotion or sentiment, particularly in texts sourced from online platforms, is complicated by grammatical errors, misspellings, sarcasm, and abbreviations. Moreover, natural

languages abound with metaphors, making it more challenging to capture the intended meaning behind the text. Another obstacle is detecting emotion or sentiment in texts featuring idioms or proverbs. For example, the English idiom "cry over spilled milk" conveys the idea of "worrying or being upset about something that has already happened and cannot be changed", but its individual word meanings do not directly suggest this interpretation. Consequently, several studies in the literature have aimed to identify the true emotions conveyed by idioms and proverbs, often employing lexicons and keyword lists (e.g., Williams et al. (2015), Ibrahim et al. (2015), Klebanov et al. (2013), Shastri et al. (2010)).

As previously mentioned, culture plays a pivotal role in shaping emotional expressions and shares a strong connection with language. It significantly shapes how people convey their feelings. Typically, individuals within the same cultural group share a common language, which contributes to the preservation of their culture. Furthermore, these communities often exhibit similarities in how they express emotions. Hence, some investigations center on disparities in emotional responses within various cultural contexts, such as studies like Scollon et al. (2004) and Lim (2016). Given the presence of both structural differences among languages and variations in how languages are used within different cultures, it is essential to recognize that emotion detection models or methods tailored for one language may not yield equivalent results when applied to other languages. Thus, the impact of these factors should be thoroughly investigated.

In this study, our primary focus was on the Turkish language, which is considered to be a language with fewer available resources. Turkish belongs to the Turkic language family, and some linguists have proposed that the Turkic language family is a potential component of the broader Altaic language family. The language exhibits distinct features that present complexities for Natural Language Processing. One of the most distinctive aspects of Turkish is its agglutinative nature, where morphemes are attached to a root word using a grammar different from that of many other languages commonly studied in the field. This frequently results in the formation of long words in Turkish, some of which can convey entire sentences' worth of meaning in English. For example,

the Turkish word “okuyacaklarmışçasına” can be translated into English as “as if they were about to read”, illustrating how Turkish can encapsulate various concepts within a single word by stringing together numerous morphemes. Since there’s no limit to the number of suffixes that can be appended, it’s possible to create words of variable lengths as needed.

Furthermore, only a few languages, such as frequently spoken English, have more resources compared to many other languages, as pointed out by Aka Uymaz and Kumova Metin (2022). Consequently, a larger body of research is dedicated to these more commonly used languages in the field of emotion detection and other NLP tasks. For example, there’s a notable scarcity of sufficiently large Turkish datasets publicly available that categorize emotions according to Plutchik’s theory. As a result, due to its complex and different grammatical structure and limited data sources, the Turkish language presents challenges in the realm of NLP and sentiment analysis.

In the subsequent sections, given the challenges inherent in researching emotion in text, we will outline the categorization of detection methods, datasets, lexicons, and evaluation metrics employed in text-based emotion detection.

3.1.1. Categorization of Emotion/Sentiment Detection Methods

Numerous studies have been conducted with the objective of identifying emotions and sentiments within text resources. Examining the previous research, studies can be categorized in various ways, but the majority of them tend to fall into one of two primary categorization approaches.

As a first approach, we can classify text-based emotion detection methods into two distinct groups: *lexicon-based* and *machine learning* methods (Canales and Martinez-Barco, 2014). To briefly elaborate, in the *lexicon-based* approach, the methodology relies on the presence of a lexicon or a predefined list of keywords to recognize emotions within the provided text. Conversely, in the *machine learning* group, supervised or unsupervised methods are employed without the necessity of additional external resources.

In the second approach, methods can be categorized into four distinct groups as defined by Sailunaz et al. (2018): *keyword-based, lexicon-based, machine learning*

and *hybrid*.

Keyword-based emotion detection involves a straightforward process of matching the words within a sentence with predetermined emotional keywords that represent specific emotion categories. In this method, a predefined list of keywords is typically associated with each emotion, and the task of emotion detection revolves around identifying words that align with these emotional keywords. To illustrate the concept of keyword-based detection, we can examine the approach presented by Ema et al. (2018). Their method follows a series of sequential steps. First, it involves proverb matching, where a list of proverbs and their associated emotions, based on their meanings, is used to check whether the given sentences contain any of these proverbs. For instance, an example proverb like "shaking like a leaf" is linked to the emotion of *fear*. The process continues with keyword matching, where a set of 25 emotion categories and 460 related keywords are compared to the text. If a match is found with any of these keywords, the method searches for negation words within the sentence. The list of negation keywords comprises words such as "not", "nor", "rarely", "aren't", and "never". Additionally, the method involves comparing a list of emoticons and abbreviated expressions commonly used in social media with the tokens present in the sentence. This list contains pairs such as ";-D", associated with the emotion *happiness*, and "g4u" (short for "good for you"), linked to the emotion *advice*.

Lexicon-based methods are one of the commonly utilized approaches in emotion analysis. Instead of depending on a predefined list of emotional keywords, researchers make use of a lexicon, which is essentially a knowledge base containing words associated with specific emotional categories or dimensions. When examining a text, this approach assigns weights to individual terms by referencing the lexicon and determining their associated scores. To calculate the emotional score of a given text, the method adds up the weights of each word within the text. There are multiple techniques available for calculating the overall score of a text.

Strapparava and Mihalcea introduced a lexicon-based approach to address the SemEval emotion annotation task in 2007, as described in their publication (Strapparava and Mihalcea, 2008). They employed a dataset consisting of news headlines and the WordNet Affect lexicon. They devised an algorithm that operates by examining the

presence of lexicon words within the news headlines and subsequently calculates a score based on the frequency of these lexicon words in the text.

Another approach, as outlined by Chaumartin, involves the use of lexicon-based methods to label emotions and valence in news headlines (Chaumartin, 2007). In Chaumartin's research, SentiWordNet and a specific subset of the WordNet-Affect lexicons are utilized. To calculate the sentiment score of a given text, a dependency graph is employed. The sentiment scores of all individual words in the text are summed up using the lexicon. However, Chaumartin takes a distinct approach by selecting the root word from the dependency graph and then multiplies its valence and emotion scores, as derived from the lexicon, by a factor of 6.

Azizan et al. (2019) utilized a concept-level sentiment analysis method that combined lexicon-based and learning-based approaches. In essence, this method assesses the sentiment of documents by computing an overall sentiment score based on a lexicon containing both positive and negative words. The words in the documents were tokenized and then compared to the positive and negative words in a lexicon consisting of 2195 positive words and 4972 negative words. The results of their study indicate that this straightforward and cost-effective lexicon-based approach produced promising outcomes, achieving an accuracy rate of 52%.

In the context of employing machine learning for emotion detection in text data, both supervised and unsupervised methods can be applied. Early research frequently relied on techniques such as Naive Bayes, decision trees, and support vector machines, as demonstrated in studies by An et al. (2017), Hasan et al. (2014), Grover and Verma (2016). Furthermore, more sophisticated approaches, including (deep) neural networks demanding substantial computational resources, have emerged, as exemplified by Baali and Ghneim (2019), Hamdi et al. (2019), Kratzwald et al. (2018), Jianqiang et al. (2018).

Douiji et al. (2016) employed an unsupervised machine learning method, which involved identifying the emotions of individual words within YouTube comments based on Ekman's six fundamental emotions. The likelihood of a word belonging to a particular emotional category is computed using the normalized pointwise mutual information. The overall probability for the comment is determined by averaging

the probabilities of its constituent words, resulting in a reported precision rate of 92.75% for the unsupervised approach. As a neural network-based approach, the study conducted by Kuta et al. (2017) can be examined. In this research, a Tree-Structured Gated Recurrent Unit (TS-GRU) algorithm is proposed to discern text sentiment and is compared with another neural network model, the Tree-Structured Long Short-Term Memory (TS-LSTM). An alternative to supervised methods is presented by Wu, Wu, Wu, Yuan, Liu and Huang (2019), who explore a semi-supervised approach based on a variational autoencoder model. This approach takes advantage of unlabeled data and focuses on sentiment analysis within a dimensional model, aiming to assign sentiment scores along valence and arousal dimensions.

Ahmad et al. (2020) conducted a study in which they demonstrated the application of machine learning techniques in the realm of emotion classification for Hindi sentences. Their approach involved representing Hindi sentences using pre-trained word embeddings for both monolingual and cross-lingual contexts. The training data included annotations of Plutchik's basic emotions at the sentence level. However, to address the issue of limited training data, their model incorporated transfer learning by leveraging larger emotion detection datasets in English. To enable the translation of Hindi to English embeddings, alignment matrices were employed, and emotion detection was carried out using a deep learning model that was based on Bi-directional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN). The model underwent training using various strategies, including All Unfreeze, Single Bottom-Up Unfreeze, Single Top-Down Unfreeze, and Gradual Unfreezing, to enhance its performance on a small dataset.

Perikos and Hatzilygeroudis (2016) utilized a combination of classifiers to automatically detect emotions in textual content. Their ensemble of classifiers consisted of a knowledge-based mechanism that applied a keyword strategy, as well as two statistical machine learning techniques: the Naive Bayes classifier and the maximum entropy learner. They collected their data from the International Survey on Emotion Antecedents and Reactions (ISEAR) and affective text datasets. They employed the Stanford Parser for sentence-level text analysis and conducted preprocessing by removing stop words and applying lemmatization. Subsequently, they represented

text features using a Bag-of-Words model. This processed input was then fed to a combination of classifiers to identify the presence of emotion in a sentence and its polarity. Their research findings indicated that the ensemble approach yielded promising results in terms of performance.

Finally, hybrid approaches are defined as combinations of the aforementioned categories. Recent studies like those by Grover and Verma (2016), Ghazi et al. (2014), and Tiwari et al. (2016) have incorporated both lexicon-based and machine-learning methods. Sailunaz et al. (2018) presented that the application of hybrid methods yielded superior accuracy scores in comparison to alternative approaches.

3.1.2. Datasets and Lexicons

In the field of emotion and sentiment analysis, numerous data sources are employed, including keyword lists, lexicons, and datasets that contain data annotated with sentiment or emotional information. The following section will extensively examine commonly used datasets and lexicons within this field.

3.1.2.1 Datasets

In the field of natural language processing, there is a noticeable disparity in the number of studies conducted on predominantly spoken languages when compared to other languages. This discrepancy results in an uneven distribution of resources among different languages. As expected, English, being one of the most widely spoken languages, has a relatively extensive pool of resources. Consequently, upon reviewing previous datasets, it becomes evident that the majority of emotion detection datasets have been developed primarily for the English language.

Table 1 presents a collection of datasets commonly utilized in text-based emotion detection. Subsequently, we will provide brief examples from widely-used English datasets, followed by examples from languages with fewer available resources, such as Turkish (an agglutinative language), Chinese (a monosyllabic language), and French and Spanish (fusional languages).

Table 1. Datasets from Literature.

Dataset	Data size	Balanced or not	Annotation	Emotion model	Language	Public Access
ISEAR (Scherer KR, 1994)	7666 sentences	✓	1096 annotators	Discrete	English	✓
EmoBank (Buechel and Hahn, 2017)	10000 sentences	-	5 raters double annotated the sentences from an average reader's and writer's perspective.	Dimensional	English	✓
TEC (Mohammad, 2012)	21051 Tweets	✗	Self-labeled via hashtags	Discrete	English	Available on request.
Tweet Emotion Intensity (TEI) (Mohammad and Bravo-Marquez, 2017)	7097 Tweets	✗	4 raters	Discrete	English	✓
EmotionLines (Chen et al., 2018)	2000 dialogues, 29245 utterances	✗	5 annotators	Discrete	English	✓
The Valence and Arousal dataset (Preoţiuc-Pietro et al., 2016)	2895 Facebook posts	-	2 annotators	Dimensional	English	✓
DailyDialog (Li et al., 2017)	13118 dialogues	✗	3 raters	Discrete	English	✓
GoEmotions (Demszky et al., 2020)	58k Reddit comments	✗	3 annotators	Discrete	English	✓
RECCON (Poria et al., 2020)	over 1000 dialogues & 10000 utterances	✗	2 annotators	Discrete	English	✓
(Bostan and Klinger, 2018)	over 100k	✗		Discrete	English	✓
DENS (Liu et al., 2019)	9710 passages	✗	3 annotators	Discrete	English	Available on request.
Demirci (Demirci, 2014)	6000 Tweets	✓	Self-tagged tweets via hashtags	Discrete	Turkish	✗
Boynukalın (Boynukalın, 2012)	4265 items from ISEAR, 1161 fairy tales	✗	3 raters	Discrete	Turkish	✗

Continued on next page

Dataset	Data size	Balanced or not	Annotation	Emotion model	Language	Public Access
TREMO (Tocoglu and Alpkocak, 2018)	27350 entries	✗	48 annotators	Discrete	Turkish	✓
Ren_CECps 1.0 (Quan and Ren, 2009)	35096 sentences, 11255 paragraphs and 1487 documents	✗	11 annotators	Discrete	Chinese	✓
(Lee and Wang, 2015)	2312 posts	✗	2 annotators	Discrete	Chinese	✗
(Blandin et al., 2021)	973 newsletters	✗	Using FEEL lexicon.	Discrete	French	✗
EmoEvent (del Arco et al., 2020)	8409 Spanish, 7303 English Tweets	✗	3 raters	Discrete	Spanish	✗

The International Survey on Emotion Antecedents and Reactions (ISEAR) dataset (Scherer KR, 1994), an English-oriented dataset, was created by merging responses from 3000 questionnaires across 37 different countries. This dataset comprises 7666 sentences, each labeled with one of seven discrete emotions: *fear, anger, guilt, joy, sadness, disgust* and *shame*. Researchers have utilized the ISEAR dataset in various studies, such as those by Calvo and Kim (2012) and Abdel Razek and Frasson (2017). Notably, this dataset is balanced, with nearly equal numbers of samples for each emotion.

Another well-known English dataset is EmoBank (Buechel and Hahn, 2017), which consists of 10000 sentences collected from diverse sources like news headlines, blogs, letters, newspapers, fiction, travel guides, and essays. EmoBank's annotations are based on a dimensional emotion model, encompassing the dimensions of valence, arousal, and dominance. Additionally, a subset of EmoBank is annotated according to Ekman's basic emotions.

Another dataset that considers the dimensional emotion model is the Valence and Arousal dataset (Preoţiuc-Pietro et al., 2016), which includes 2895 English language Facebook posts. These posts are annotated based on the Circumplex emotion model and are labeled with respect to valence and arousal dimensions.

The Twitter Emotion Corpus (TEC) (Mohammad, 2012) dataset stands out as one of the most extensive resources available for text-based emotion detection. Comprising 21051 tweet samples, this dataset incorporates hashtags related to Ekman's six emotion categories, such as *happy* and textitangry. Experiments have confirmed the consistency of self-labeled annotations using these hashtags.

The Tweet Emotion Intensity dataset (Mohammad and Bravo-Marquez, 2017), on the other hand, encompasses 7,097 tweets categorized into four distinct emotion types: *joy*, *fear*, *sadness* and *anger*. This dataset was meticulously assembled by selecting 50 to 100 terms associated with each emotion category, each representing different levels of emotional intensity. For example, for the *anger* emotion category, terms like *angry*, *annoyed* and *frustrated* were carefully chosen, and tweets containing these selected terms were then incorporated into the dataset.

The EmotionLines dataset introduces a novel approach by highlighting the significance of contextual emotion dynamics (Chen et al., 2018). It represents the pioneering instance of a dataset where dialogue utterances are meticulously labeled according to their emotional content. These texts are annotated with Ekman's six primary emotions and the *neutral* emotion category. This labeled dataset encompasses a total of 2000 dialogues and 29,245 utterances, sourced from Facebook messenger dialogues and scripts from the television series "Friends".

Moving on to the DailyDialog dataset, it offers a diverse collection of multi-turn dialogues spanning ten distinct topics, such as everyday life, health, and politics (Li et al., 2017). This dataset is compiled from various English dialogues available on websites, utilized by English language learners. It has been manually annotated with emotions, encompassing six different emotional categories.

The GoEmotions dataset is characterized as one of the most extensive manually annotated resources, featuring a vast collection of 58,000 Reddit comments (Demszky et al., 2020). Each comment within this dataset has undergone annotation by three raters, with an additional two raters stepping in when consensus was not reached. These comments are tagged with one or more emotions from a broad selection of 27 categories or are marked as *neutral*.

The RECCON dataset (Poria et al., 2020), which incorporates elements from both the DailyDialog (Li et al., 2017) and IEMOCAP (Busso et al., 2008) datasets, is a manually annotated collection of dialogues and utterances. This dataset not only contains emotion labels but also provides information about the reasons behind these emotions.

Bostan and Klinger (2018) introduced a novel dataset that consolidates 14 pre-existing emotion corpora, encompassing datasets like ISEAR, EmoBank, and DailyDialog, into a unified labeling scheme comprising 11 emotion categories. This unified dataset facilitates direct comparisons between various datasets.

The DENS dataset comprises both traditional narratives and contemporary stories in the English language, as mentioned in Liu et al. (2019). While this dataset exhibits an imbalance when categorizing emotions into 9 specific categories, it becomes reasonably balanced when the categories of surprise and disgust are removed.

Due to a limited number of studies conducted in languages other than English, we will now focus on examples of datasets used in the field of emotion and sentiment detection, specifically in Turkish, Chinese, French, and Spanish.

One instance of a Turkish dataset was created by Demirci (2014), where Twitter was the data source. This dataset was gathered by performing keyword-based searches for six different emotions, including fear, joy, sadness, surprise, anger, and disgust. Approximately 6,000 tweets, evenly distributed across these six emotions, were collected through Twitter searches utilizing hashtags.

In the research conducted by Boynukalın (2012), two datasets were formulated for text-based emotion extraction. The first dataset was constructed by selecting and translating a subset of sentences from the ISEAR dataset, containing only four emotions. This involved the efforts of 33 individuals who translated these sentences, resulting in a dataset comprising 4265 documents. The second dataset consisted of 25 Turkish fairy tales obtained from various websites.

Another notable Turkish dataset is TREMO (Tocoglu and Alpkocak, 2018), which was developed based on memories and experiences shared by 4709 individuals of varying ages and locations, following the emotion categories outlined by Ekman. Subsequently, this dataset was meticulously annotated by 3-5 annotators, resulting in a total of 27350 documents within the dataset.

Our first Chinese dataset example is Ren_CECps 1.0, developed by Quan and Ren (2009). This dataset experienced manual annotation at the document, paragraph, and sentence levels by a team of 11 annotators. It covers discrete emotion categories, including *love*, *expect*, *joy*, *anxiety*, *surprise*, *sorrow*, *angry* and *hate*.

Another Chinese dataset, assembled by Lee and Wang (2015), is composed of posts gathered from Weibo, a prominent Chinese social media platform. This dataset encompasses emotions such as textitsurprise, *happiness*, *fear*, *anger* and *sadness* For posts containing English text, they employed English-to-Chinese translations.

Moving to French, Blandin et al. (2021) introduced a dataset based on newsletters. In this dataset, each word is associated with an emotion vector, constructed using the FEEL emotion lexicon (Abdaoui et al., 2017), and it encompasses the six basic emotions as defined by Ekman.

For a multilingual resource that includes both English and Spanish texts, the EmoEvent dataset is a notable example (del Arco et al., 2020). It originates from a collection of Tweets that were annotated with Ekman's six emotion categories, as well as neutral and other emotion categories, by a group of three annotators.

As demonstrated by various examples in the literature, the process of collecting datasets and performing annotations can be performed in diverse ways. Some datasets involve individuals being prompted to compose texts or respond to questions related to various emotion categories. Alternatively, data sources can be compiled by extracting text from websites, including news articles, blogs, or social media platforms. The annotation process for certain datasets is carried out by the authors of the texts themselves, while in other cases, dedicated annotators assess each text to determine its emotional or sentiment content.

For instance, the ISEAR dataset is derived from responses to questionnaires provided by 1096 participants across various emotion categories (Scherer KR, 1994). Consequently, this dataset comprises texts labeled with emotions based on the perspectives of the text authors. Similarly, the TEC dataset is a self-labeled collection of Tweets categorized into one of Ekman's six emotion categories (Mohammad, 2012).

In contrast, some datasets involve presenting the collected text to multiple annotators, and emotion labels are determined through their collective evaluations. For instance, in the case of EmoBank, approximately 10000 sentences sourced from various origins were annotated by five annotators (Buechel and Hahn, 2017).

3.1.2.2 Lexicons

Within the literature, a multitude of lexicons have been developed specifically for sentiment and emotion analysis for lexicon-based approaches. Similar to datasets, there is a notable discrepancy in the availability of lexicons between English and other languages. In this section, we will provide a concise overview of some of the lexicons that have been previously explored.

In essence, an emotion or sentiment lexicon serves as a categorized list of words. In sentiment lexicons, words are typically assigned polarity values such as *positive*,

negative, or *neutral*. Conversely, emotion lexicons encompass collections of words that have been labeled with discrete emotion categories like *joy*, *love* or *sadness* or they may involve dimensions like *pleasure* and *dominance*.

Table 2 presents a comparative overview of various lexicons. For instance, upon examination of the third column, it is seen that EmoWordNet (Badaro et al., 2018) and DepecheMood (Staiano and Guerini, 2014) stand out as English lexicons with a significantly larger number of terms compared to others. DepecheMood is annotated automatically by extracting news with readers' emotional selections about them while EmoWordNet is the extended version of DepecheMood. On the other hand, DUTIR, a non-English lexicon, has an extensive dataset in comparison to the other lexicons featured in the list (Chen, 2008). It is a Chinese Sentiment lexicon having 27466 words.

Table 2. Lexicons from Literature.

Lexicon	**Type**	**Number of Units**	**Language**
NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2013)	Emotion and Sentiment	14182 unigrams	English
WordNet-Affect (Strapparava and Valitutti, 2004)	Emotion	4787 words, 2874 synsets	English
Affective Norms of English Words (ANEW) (Bradley and Lang, 1999)	Emotion	1034 words	English
Extended Affective Norms of English Words (E-ANEW) (Warriner et al., 2013)	Emotion	13915 lemmas	English
SentiWordNet (Baccianella et al., 2010)	Sentiment	117000 synsets	English
Dalian University of Technology Information Retrieval (DUTIR) (Chen, 2008)	Sentiment	27466 words	Chinese
SentiStrength (Thelwall et al., 2012)	Sentiment	2489 terms	English
The Semantic Orientation Calculator (SO-CAL) (Taboada et al., 2011)	Sentiment	2252 adjectives, 745 adverbs, 1142 nouns, and 903 verb entries.	English
Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto and Gilbert, 2015)	Sentiment	7500 words	English
DepecheMood (Staiano and Guerini, 2014)	Emotion	37 thousand terms	English
EmoWordNet (Badaro et al., 2018)	Emotion	67 thousand terms	English
Turkish Emotion Lexicon (TEL) (Tocoglu and Alpkocak, 2019)	Emotion	1320 terms	Turkish
FEEL (Abdaoui et al., 2017)	Emotion and Sentiment	14127 terms	French
(Gala and Brun, 2012)	Sentiment	7483 nouns, verbs, adjectives and adverbs	French

(Navarrete et al., 2021)	Emotion	1892 words	Spanish
(Redondo et al., 2007)	Emotion	1034 words	Spanish
Spanish Emotion lexicon (Sidorov et al., 2012)	Emotion	2036 words	Spanish

Table 2 also highlights that WordNet-Affect (Strapparava and Valitutti, 2004) and SentiWordNet (Baccianella et al., 2010) lexicons are subsets of the WordNet lexicon (Miller, 1995). This English dataset encompasses nouns, adjectives, adverbs, and verbs organized into synonym sets known as synsets, totaling 117000 synsets in the lexicon. WordNet-Affect (Strapparava and Valitutti, 2004) was compiled through manual emotion annotation of specific words from dictionaries, and they expanded their list by selecting synsets from WordNet containing at least one word from their original list. Similarly, SentiWordNet (Baccianella et al., 2010) is another lexicon that originates from WordNet. This lexicon is created by applying automatic sentiment tagging through a semi-supervised learning approach, which involves a random walk mechanism to enhance the sentiment scores of all WordNet synsets, categorizing them into *positive, negative*, or *neutral* classifications.

From these lexicons, the National Research Council Canada (NRC) lexicon is one of the widely utilized emotion lexicons in the literature (Mohammad and Turney, 2013) (e.g., Waspodo et al., 2022; Benchimol et al., 2021; Seyeditabari et al., 2019; Agrawal et al., 2018). The manually annotated lexicon terms are based on several emotion categories of Plutchik's eight emotions.

Affective Norms of English Words (ANEW) (Bradley and Lang, 1999) is another English lexicon having 1034 manually labeled terms by *arousal, dominance* and *valence*. Following, Warriner et al. (2013) extended the ANEW lexicon such that it will have 13915 lemmas rated according to the dimensional model.

The Semantic Orientation Calculator (SO-CAL) (Taboada et al., 2011) is an English sentiment lexicon having adjectives, adverbs, nouns and verbs labeled according to sentiment polarity and strength.

In Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto and Gilbert, 2015) features from well-known lexicons are collected and extended. The lexicon is labeled based on sentiment polarity in the interval of (-4, +4).

Turkish Emotion Lexicon (TEL) (Tocoglu and Alpkocak, 2019) is the first Turkish Emotion lexicon that is formed from a Turkish dataset (TREMO (Tocoglu and Alpkocak, 2018)). The lexicon comprises four distinct versions, each considering different lemmatization and stemming techniques.

FEEL (Abdaoui et al., 2017) is a French lexicon that is formed by a semi-automatic translation of NRC-EmoLex. First, online translation is utilized then entries and related emotions are validated by a human.

The lexicon proposed by Gala and Brun (2012) is another French lexicon. It has nouns, verbs, adjectives, and adverbs that are semi-automatically labeled by polarity information.

Finally, in Table 2, the Spanish lexicons Navarrete et al. (2021), Redondo et al. (2007), and Sidorov et al. (2012) are listed. The first of these, the Navarrete et al. (2021) is constructed by translating the EmoLex, and then it is expanded with the synonyms from WordNet. It has 1892 words based on discrete categories. Redondo et al. (2007) is the Spanish-translated and collectively evaluated version of the ANEW lexicon. Lastly, in order to construct the Spanish Emotion Lexicon (SEL) (Sidorov et al., 2012), words are chosen from the SentiWordNet lexicon, automatically translated into Spanish, and then manually verified in accordance with Ekman's six emotions.

3.1.3. Emotion/Sentiment Detection Evaluation Metrics

In the concept of emotion and sentiment detection, the assessment of performance is carried out using well-established evaluation metrics from the field of information retrieval, similar to other natural language processing tasks. These commonly employed metrics include *accuracy*, *precision*, *recall*, and a composite measure known as the *F-score*. *Accuracy*, which is utilized in nearly all previous studies (e.g. Shi et al. (2018), Agrawal et al. (2018), Su et al. (2018), Tang et al. (2014)), calculates the proportion of correct predictions in relation to the total samples in the experiment. Essentially, it quantifies the ratio of accurately classified samples, particularly in tasks like emotion labeling, which often involve more than two classes.

On the other hand, *precision (P)* and *recall (R)* serve as alternative evaluation metrics in emotion, sentiment, and polarity detection problems and have been employed in numerous prior studies (e.g. Jiwung Hyun and Cheong (2020), Chang et al. (2019), Mao et al. (2019)). In binary classification scenarios, where only two outcomes such as *positive* or *negative* exist, *precision* is defined as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

where "TP" stands for true positive and "FP" represents false positive. In binary classification, "TP" represents the count of correctly labeled samples belonging to one of the classes, while "FP" denotes the number of samples incorrectly predicted to belong to the same class. As previously discussed, emotion detection typically involves a multi-class problem, where samples can be assigned to one of three or more classes. In such instances, two common techniques are employed: micro averaging and macro averaging. Microaveraging involves aggregating all correctly classified samples, regardless of their class, to calculate the total "TP". Similarly, for each class, samples incorrectly assigned to that class by the classifier are counted and the sum of all class FPs yields the total "FP" value. Finally, Equation 1 is applied using the total "TP" and "FP" values to derive the micro-averaged *precision*. In contrast, macro-averaging calculates the *precision* value for each class independently and then averages these individual *precision* values.

As for the *recall* metric, it represents the ratio of correctly predicted samples to the total number of samples belonging to a specific class in the given experiment, and it is expressed as:

$$R = \frac{TP}{TP + FN} \tag{2}$$

In Equation 2, TP stands for true positive, while FN corresponds to false negative. TP (True Positives) represents the number of samples correctly predicted as belonging to the designated class, while FN (False Negatives) counts the samples that should have been identified as positive but were incorrectly classified as negative. In the context of multi-class emotion detection, a similar approach as previously described for the *precision* metric is applied, involving micro and/or macro averaging, with the aim of producing a single performance value.

The *F-score*, also referred to as the *F1-score*, is another metric that combines both *precision* and *recall* (e.g. Naderalvojoud and Sezer (2020), Alshahrani et al. (2019), Shi et al. (2018), Wang and Meng (2018)). Essentially, it is the harmonic mean of these two metrics, as illustrated in Equation 3.

$$F = 2 * \frac{P * R}{P + R} \tag{3}$$

3.2. Vector Space Models

The primary objective in the field of natural language processing is to establish a connection between computers and human language, enabling computers to comprehend, analyze, and generate language. However, computers inherently operate with numerical data in the form of zeros and ones. Consequently, the intricate elements of language, encompassing grammatical rules, vocabulary, and various linguistic components, must be translated into numerical representations. Typically, both written and spoken language examples are stored and processed in textual format. This underscores the significance of text as a primary mode of human communication, housing a wealth of valuable information employed in diverse fields such as emotion and sentiment analysis, text similarity, summarization, classification, and clustering. In the domain of natural language processing, the process of converting textual data into numerical representations is commonly referred to as vectorization. The collective

representation of documents in a shared vector space is denoted as the vector space model, as elucidated by Manning et al. (2008). In this model based on linear algebra, there is the capability to perform vector-based operations such as addition, subtraction, and similarity measurements.

Within the framework of vector space modeling, one of the initial methods employed was the technique known as *one-hot encoding*, which is recognized for its reliance on word count and frequency. In this approach, a binary vector is constructed for each word in the vocabulary, contingent upon its frequency of occurrence. Here, the term "vocabulary" pertains to the compilation of distinct terms or words found within the given collection of documents. Each word corresponds to a vector of size n, where n corresponds to the size of the vocabulary. In this vector, the *jth* position is marked as 1, while the remaining positions remain 0. To illustrate, consider a set of three sentences within a text document: "Birds fly in the sky", "Birds are animals", and "Animals are our friends". In our vocabulary "V", we identify six words: "birds", "fly", "in", "the", "sky", "are", "animals", "our", "friends". Consequently, the word vectors can be visualized as presented in Table 3.

V = {birds, fly, in, the, sky, are, animals, our, friends}

Table 3. Word vectors for example vocabulary V.

Word	**Vector**								
birds	1	0	0	0	0	0	0	0	0
fly	0	1	0	0	0	0	0	0	0
in	0	0	1	0	0	0	0	0	0
the	0	0	0	1	0	0	0	0	0
sky	0	0	0	0	1	0	0	0	0
are	0	0	0	0	0	1	0	0	0
animals	0	0	0	0	0	0	1	0	0
our	0	0	0	0	0	0	0	1	0
friends	0	0	0	0	0	0	0	0	1

Even though the implementation of *one-hot encoding* is straightforward, it poses a significant challenge in terms of space complexity, especially when dealing with a large vocabulary size. For instance, in our example, there is a limited number of dimensions in the vector space. However, in a vocabulary containing 10 million words, representing these words using one-hot encoding would demand a substantial

amount of memory storage and make vector operations like addition and comparison computationally intensive. Moreover, the resulting vectors would be highly sparse, consisting mainly of zeros with only a few ones. This sparse representation may not capture an adequate amount of semantic information, making it difficult for vector operations to effectively reveal relationships between words. For instance, distance metrics (e.g., cosine distance) would struggle to accurately depict the degree of semantic similarity between words.

Another count-based method, *co-occurrence matrix representation*, offers an alternative method based on term weights. This approach is grounded in the concept that words with similar meanings often appear in similar contexts. Here, the method involves counting the number of times a term appears alongside its neighboring words within a fixed window size. In this context, "neighboring words" refers to the list of words that appear immediately before or after the target term in the text. The term weights are determined by counting how often two terms co-occur within the specified window size. For example, if the window size is set to one, the co-occurrence of each term with its preceding and succeeding terms is computed for every word in the vocabulary. Consequently, a square matrix is constructed, the dimensions of which are contingent upon the vocabulary's size. Similar to *one-hot encoding*, a notable drawback of this approach is the substantial memory requirement associated with the utilization of the co-occurrence matrix.

Term frequency (*tf*) and *inverse document frequency* (*idf*) are the other two alternatives in vector space modeling. They assign weights to words based on their frequency of occurrence. In *tf*, the weight of a term t within document D corresponds to how often t appears in D (Manning et al., 2008). In short, it's computed by determining how many times the term t appears relative to the overall term count in document D. In this approach, all terms are considered equally relevant to the document. On the other hand, *idf* calculates a term's weight as its frequency in document D divided by the number of documents where the term is found. This method assigns a higher weight to terms that appear frequently in one document but rarely in others, assuming they carry more meaningful and document and context-specific information. When combined, *tf* and *idf* form the well-known *tf-idf* weighting technique. In *tf-idf*, a term

t in a document *D* has the highest weight when *t* occurs frequently in a small number of documents and the lowest weight when *t* is present in all documents (Manning et al., 2008). However, *tf-idf* doesn't capture semantic relationships between terms and can be computationally expensive when dealing with large vocabularies, similar to the problem observed in the previously listed methods.

Previously mentioned early approaches of vector space models come with three significant limitations. Firstly, they demand more memory space as the vocabulary size expands. Secondly, the computational burden escalates as operations increase with higher dimensions in the vector space. To address these shortcomings, the concept of dimensionality reduction has been introduced, as suggested by Raunak et al. (2019). When employing dimensionality reduction, some terms may be excluded from the matrix, taking into account contextual words. Techniques for reducing dimensions, especially in bag-of-words models (such as one-hot encoding and co-occurrence matrix), are commonly discussed in the context of feature selection. In bag-of-words representations, every individual word is treated as a separate feature, and it's possible to utilize feature selection methods to pick out more informative terms within the sparse representations. For example, in the study of Erenel et al. (2020), various selection strategies like chi-square, Gini-Text relevance frequency (as presented by Park and Kwon (2011)), and class-wise feature selection proposed by Kumar and Harish (2020) are employed. While these approaches have demonstrated success in text-based emotion and sentiment recognition, they also run the risk of information loss. Thirdly, when new documents are added to the dataset, the vocabulary size expands with the inclusion of new terms, necessitating the reconstruction of the vector space. To address these limitations, models that incorporate word embeddings are introduced, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The primary distinction of these alternatives lies in their ability to construct vectors prior to performing operations in the document space.

In recent research, word embeddings have emerged as the most widely adopted models for representing text documents as fixed-length vectors in space. These models facilitate the transformation of words into vector representations by capturing their semantic, syntactic, and contextual nuances. Each term within the dataset is

depicted as a distinct vector, and terms sharing similar contexts exhibit comparable representations. Word embeddings enable the identification of word similarity or relationships within a document through straightforward operations. Examples of word embedding methods include models like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).

Word2Vec, an open-source pre-trained model created by Mikolov and his team at Google, has gained popularity for generating static word embeddings using the Google News dataset (Mikolov et al., 2013). This model consists of 3 million English word vectors, each having 300 dimensions, and is structured as a shallow, two-layer neural network. To explain it briefly, the model includes an input layer, an output layer of the same size as the input layer, and a hidden layer with various parameters like window size and embedding size. The number of neurons in the hidden layer determines the dimensionality of the word embeddings. Word2Vec takes a significant text corpus as input and represents each term within the corpus as a vector. Words with similar contextual meanings are positioned close to each other in the vector space. This unsupervised model works based on prediction and doesn't rely on labeled data.

Word2Vec provides two model architectures: skip-gram and continuous bag-of-words (CBOW), both of which are depicted in Figure 3. In the skip-gram model, each word acts as input to a log-linear classifier and the model predicts words within a window before and after the current word (Mikolov et al., 2013). Mikolov et al. noted that increasing the word range can improve the quality of word vectors, but it comes with increased computational complexity. Conversely, the CBOW model predicts the current word vector based on the surrounding words. This architecture is also referred to as the bag-of-words model because word order doesn't affect the prediction of the target word vector (Mikolov et al., 2013). They assessed the performance of their word representation model in word similarity tasks and compared it with other top-performing neural network-based methods.

Another example of static word embedding models is GloVe (Global Vectors for Word Representation) which is a vector space model that goes beyond local information by also incorporating global context. The algorithm, developed as an

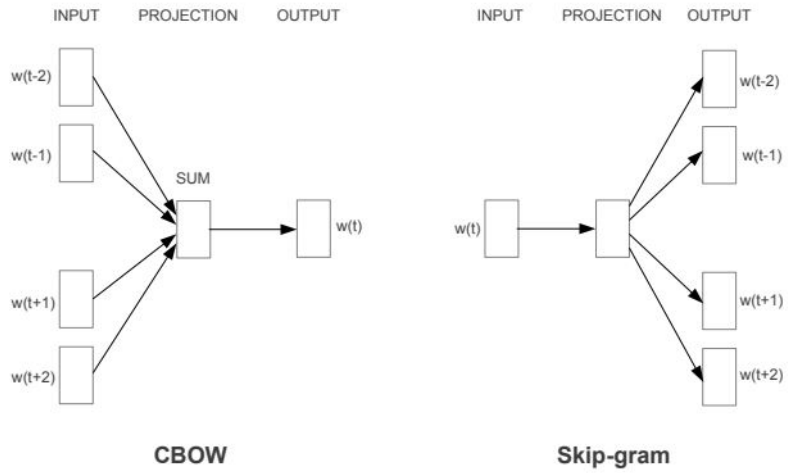


Figure 3. Word2Vec model architectures (Mikolov et al., 2013).

open-source project at Stanford University by Pennington et al. (2014), represents words in a vector space, wherein the proximity of word vectors reflects the degree of semantic similarity. When word vectors are close together, it signifies a strong semantic relationship between the corresponding words. The training process involves using co-occurrence statistics of word pairs from a corpus. They conducted multiple experiments, including tasks related to word similarity, word analogies, and named entity recognition, to assess its performance (Pennington et al., 2014).

As opposed to these static word embedding models, more recent models have emerged that create contextualized embeddings (e.g. Peters et al. (2018), Devlin et al. (2018), Sanh et al. (2019)). These embeddings consider the content, word sense, and the concept of polysemy. This implies that when a word has multiple meanings, its vectors can adapt based on the specific context or sentence in which it is employed.

ELMo (Embeddings from Language Models), introduced by Peters et al. (2018), creates contextualized representations of words by considering both their semantic and syntactic characteristics, as well as their multiple meanings if applicable (polysemy). Unlike traditional word embeddings like GloVe or Word2Vec, ELMo generates multiple vectors for words that can have diverse meanings or appear in different contexts within a sentence. These embeddings are computed based on the entire sentence, distinguishing ELMo from its counterparts. To train ELMo, a bi-directional Long Short-Term Memory (LSTM) model is utilized, trained on a substantial corpus

comprising 30 million sentences (Peters et al., 2018). ELMo has found application in various NLP tasks, consistently enhancing performance in areas such as question answering (Rajpurkar et al., 2016), named entity extraction (Tjong Kim Sang and De Meulder, 2003), and sentiment detection (Socher et al., 2013).

In 2018, Google introduced a recent contextualized embedding model called Bidirectional Encoder Representations from Transformers (BERT) through a transfer learning process (Devlin et al., 2018). The idea behind transfer learning is to leverage a pre-trained base model for a new task instead of training entirely new models from scratch, thereby reducing the computational cost. This approach aims to enhance the performance of the model on the new task by leveraging knowledge acquired from previous tasks (Verwimp and Bellegarda, 2019). The process is particularly useful when there is a limited amount of labeled data available for supervised training in a specific task (Verwimp and Bellegarda, 2019).

Transfer learning plays a pivotal role in various NLP tasks, diminishing the reliance on domain-specific data and presenting its own set of challenges and solutions. One such challenge is known as multi-source learning, which involves selecting the appropriate pre-trained model or data source for a given task. Some approaches involve using a combination of source datasets. For instance, in text categorization (Gupta and Ratinov, 2008), heterogeneous datasets like Wikipedia, Open Directory Project, and Yahoo are employed, though they may have practical access limitations. As an alternative strategy, these limitations are addressed by employing pre-trained models instead of relying solely on the source data, as demonstrated in works such as Nguyen et al. (2022) and Lee, Sattigeri and Wornell (2019). In the case of sentiment analysis, for example, Nguyen et al. (2022) introduces a method where N embeddings are input into a gating network to create a merged embedding for sentiment classification. The authors compare their model with alternative embeddings such as BERT, recurrent CNN, and concatenation using a large-scale Vietnamese database. Another aspect of transfer learning is multi-level learning, as described in Hung and Chang (2021). This approach involves adding one or more fine-tuning levels that leverage domain-specific knowledge to improve the performance of pre-trained embeddings. In Hung and Chang (2021), fine-tuning is applied to tasks in different fields, including facial emotion

recognition and named entity recognition. Additionally, in studies related to text-based emotion/sentiment detection, enriching pre-trained embeddings with emotional content, typically obtained from a different resource, can also be viewed as a form of fine-tuning for the specific domain.

As transfer learning techniques evolve for NLP tasks, they introduce certain limitations, such as the issue of catastrophic forgetting. It occurs when a neural network model, as it learns new tasks, adjusts its model parameters and weights for previous tasks, potentially leading to a decline in performance on those earlier tasks (McCloskey and Cohen, 1989), (Ke et al., 2020). Essentially, catastrophic forgetting is the phenomenon where a neural network gradually loses access to previously learned information as it acquires new knowledge. The literature has explored various methods to mitigate catastrophic forgetting (Ke et al., 2020), (Arora et al., 2019), (Goodfellow et al., 2015), (Srivastava et al., 2013). In a study conducted by Arora et al. (2019), they investigated the factors contributing to forgetting during the training process and compared the forgetting tendencies of CNN and LSTM architectures. Through a series of experiments, they observed that CNNs exhibit less forgetting compared to LSTMs. Furthermore, they found that using ELMo contextual word embeddings Peters et al. (2018) in both architectures helps to address the problem of forgetting. Additionally, there is research focused on sentiment classification that uses incremental learning to handle a series of classification tasks and reduce the problem of catastrophic forgetting, as shown in the works of authors such as Qin et al. (2020) and Lv et al. (2019).

BERT considered one of the notable instances of transfer learning models, generates contextualized embeddings by utilizing a multi-layer bidirectional transformer encoder architecture, as represented in Devlin et al. (2018). Similar to ELMo, BERT's model provides word embeddings that can vary based on the sentences in which those words appear. To create pre-trained deep bidirectional word representations, the model employs a masked language model approach Devlin et al. (2018). Devlin et al. conducted their pretraining on a substantial dataset comprising 800 million words from BooksCorpus (Zhu et al., 2015) and an additional 2,500 words from the English Wikipedia. The authors demonstrated the model's adaptability to a range of tasks, such as language inference and question answering. In 2019, Google's search engine began

using the BERT model to process English queries. However, it's important to note that BERT, being a transformer-based model, presents challenges in terms of computational cost due to its extensive training dataset and sub-word tokenization method, as pointed out by Moon and Okazaki (2021).

Several BERT variants, such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and DistilBERT (Sanh et al., 2019), have emerged after the state-of-the-art performance and potential of BERT. These variants were developed to address issues with the original BERT model, such as its high time and resource requirements due to its millions of parameters and complex architecture. For instance, RoBERTa (Liu et al., 2019) is a variant that utilizes dynamic masking instead of static masking in BERT. Liu et al. (2019) researched hyperparameter tuning and training set size, revealing that BERT was undertrained. In a study by Kumar and Albuquerque (2021), a derivative of RoBERTa was employed to investigate the performance of cross-lingual contextual word embeddings and zero-shot transfer learning on two Hindi datasets. The model was trained on a resource-rich language (English) and applied to a resource-scarce target language (Hindi). To accomplish this, they used the cross-lingual XLM-RoBERTa (XLM-R) transformer model (Conneau et al., 2020) and zero-shot transfer learning, a machine learning approach that allows the classification of unlabeled samples even when their categories are not present in the training data. In summary, Conneau et al. (2020) conducted fine-tuning on the XLM-R model using one language and evaluated its performance on another language, all without relying on machine translation. Their research revealed that XLM-R surpassed the performance of other studies when applied to Hindi datasets. ALBERT, proposed by Lan et al. (2019), introduced two parameter reduction techniques to reduce memory consumption and enhance training speed in BERT. This variant reduced the parameter size of BERT by approximately one-tenth. DistilBERT (Sanh et al., 2019), on the other hand, was designed as a compressed, faster, and lighter version of the BERT model. For DistilBERT, the aim was to create a model with a reduced size while preserving the power of BERT. The researchers behind DistilBERT achieved this by employing knowledge distillation (Sanh et al., 2019) as a compression technique, training a compact model while reducing the number of parameters. While the general

architecture of DistilBERT is similar to BERT, it has half the number of layers. In the research of Batra et al. (2021), three different strategies for sentiment analysis using BERT-based models and their variations are outlined. To start, the BERT model undergoes fine-tuning with 2-4 training epochs and then, they employ an ensemble technique that combines base BERT, RoBERTa, and ALBERT models, with the final prediction determined through a voting mechanism. Lastly, they introduce a distillation framework that incorporates both the base BERT and DistilBERT models. According to their findings, these ensemble and distillation approaches have led to substantial improvements, with F1 measure enhancements ranging from 6% to 12% across various datasets.

Apart from BERT and its variations, several other transfer learning-based models have been utilized in sentiment and emotion analysis. For instance, Malhotra et al. (2021) introduced a bidirectional model based on Universal Language Model Fine Tuning (ULMFit) for sentiment classification. Their proposed model incorporates pre-trained language modeling and fine-tuning on target data by adjusting the recurrent weights of Average Stochastic Gradient Descent Weight Dropped LSTM. In their study, they compared their model with classical machine learning techniques like logistic regression and linear SVM, semantic embeddings such as FastText, and other transfer learning methods including ULMFit and BERT. Tao and Fang (2020) introduced an alternative transfer learning approach for aspect-based sentiment analysis. Their model, as described by Tao and Fang (2020), expands upon existing methods by including multi-label classification. They utilized BERT and XLNet as transfer learning techniques and conducted a thorough comparison against 27 baseline deep learning and machine learning methods. Their models consistently demonstrated higher accuracy levels across three distinct datasets. Furthermore, their results indicated that XLNet outperforms BERT on two datasets, primarily due to XLNet's better performance with limited labeled data during fine-tuning of pre-trained data (Tao and Fang, 2020), (Yang et al., 2019). On the contrary, when assessing training time, BERT was found to run faster than XLNet.

In addition to the frequent usage of contextual word embedding models in NLP research, semantic embeddings like Word2Vec and GloVe are still being used.

However, they face challenges related to polysemic words. This is because these neural networks can only produce a single static vector for each word. Moreover, certain word embeddings are susceptible to encountering the out-of-vocabulary (OOV) issue, which essentially means dealing with words that are not included in the dataset specific to the given NLP problem domain. In NLP tasks, word embeddings are typically trained on extensive corpora with numerous words, but encompassing the entire vocabulary of a language may not be feasible, leading to the OOV problem. Consequently, in various tasks, performance may decline due to the presence of unknown vectors for certain words. In the literature, specific studies, such as the research conducted by Moon and Okazaki (2021), have investigated the impact of OOV specifically in the context of transfer learning for languages with large alphabets like Japanese, Chinese, and Korean, utilizing BERT as a word embedding model. Moreover, some studies have proposed solutions to address this problem, including assigning the embedding of an unknown token to OOV words or using the average of context words' embeddings as the vector for OOV words (Won et al., 2021). Additionally, certain established methods offer their own remedies for the OOV problem. For example, the BERT model tackles this issue through sub-word tokenization. Conversely, Word2Vec does not address the OOV problem, as it generates word embeddings equal to the size of the vocabulary in the corpus on which the model is trained. It's important to point out that the OOV problem has consequences, including its influence on the performance of transfer learning, which entails applying a model trained on one task to another related task (Pan and Yang, 2010).

Considering a broad spectrum of vector space models, including those presented here in conjunction with others in the field, it can be asserted that the angular distance between vectors is a commonly employed metric for representing the similarity or dissimilarity of words. Consequently, most vector representations operate under the assumption that words frequently observed in similar contexts are likely to share similar meanings. Nevertheless, certain emotionally dissimilar words may yield higher similarity scores than emotionally similar ones, primarily due to their frequent co-occurrence (e.g. *happy* and *sad*). This issue can lead to unexpected outcomes in emotion detection studies. To address this challenge, researchers have proposed

various methods to enhance the vector representation of words, such as incorporating emotional or sentiment-related information into the original representations. In summary, these studies aim to place words closer to each other in vector space not just based on their semantic or co-occurrence statistics but also based on the emotions they convey. Subsequent sections 3.2.1 and 3.2.2. will provide details about studies concerning this subject matter found in the literature.

3.2.1. Emotion Enriched Vectors

In this study, the vectors created by combining original word vectors with emotional information, as named in the work of Agrawal et al. (2018), will be referred to as *emotion-enriched word vectors*. When reviewing previous research, emotion-enriched word vector models can be categorized based on three main groups. The first one involves the utilization of an external emotion lexicon to enhance the vectors, which is exemplified by studies such as Agrawal et al. (2018), Seyeditabari et al. (2019), and Wang and Meng (2018). Secondly, another group of proposed methods employs a distance measure, commonly using cosine similarity (e.g., (Mao et al., 2019)). The final category encompasses studies that utilize datasets based on either a categorical model (e.g., Wang and Meng (2018), Mao et al. (2019)) or a dimensional model (e.g., Wu and Jiang (2019), Su et al. (2018)). Table 4 presents several different studies proposing emotion-enriched vectors, along with details about the base model employed, data sources, emotion models, and utilized evaluation results.

Table 4. Emotion Enriched Vectors.

Reference	Base Model	Dataset	Lexicon	Evaluation
(Agrawal et al., 2018)	LSTM	Amazon reviews dataset (McAuley et al., 2015)	WordNetAffect (Strapparava and Valitutti, 2004) NRC EmoLex (Mohammad and Turney, 2013)	Accuracy: 33.2%
(Seyeditabari et al., 2019)	RNN		NRC (Mohammad and Turney, 2013)	Similarity metrics. (29% improvement for GloVe vectors)
(Lee, Park and Choi, 2019)	LSTM	EmotionLines (Chen et al., 2018)	NRC (Mohammad and Turney, 2013)	
(Mao et al., 2019)	Word2Vec skip-gram	Collection of Weibo texts	DUTIR (Chen, 2008)	P: 72.11% R: 72.9% F1: 70.9%
(Wang and Meng, 2018)	pre-trained word vectors	Chinese & English reviews and Weibo texts.	Combination of Chinese lexicons	Accuracy: 74.3% F1: 77.5%
(Alshahrani et al., 2019)	CMA-ES	TEC (Mohammad, 2012)		F1: 33.8%-66.88%
(Su et al., 2018)	LSTM	NLPCC (*NLPCC Evaluation Tasks*, 2014)	CVAW (Yu et al., 2016)	Accuracy: 70.66%
(Jiwung Hyun and Cheong, 2020)	textCNN (Kim, 2014)	OffMyChest (Jaidka et al., 2020)		Accuracy: 69.6%
(Wu and Jiang, 2019)	Autoregressive linear model	SEND (Ong et al., 2019)	NRC (Mohammad and Turney, 2013)	Average similarity increase in positive valence emotions: 16.38%, negative valence emotions: 21.88%
(Wongpatikaseree et al., 2021)	Word2Vec	Twitter keyword search		F1: 0.76
(Matsumoto et al., 2022)	DistilBERT	Japanese-English bilingual sentiment corpus McInnes et al. (2020) Tweets and blog with emotion tags	F1: 0.33%-0.73%	

As an illustrative example of employing lexicons in word representation, Agrawal et al. (2018) introduced a technique with the goal of arranging vectors of words with similar emotional terms close to each other in a spatial context, while positioning words with dissimilar emotional ones farther apart. Their approach leverages recurrent neural networks, where the initial word vectors serve as inputs to a Long Short-Term Memory model. These word vectors are refined during the training process to acquire embeddings enriched with emotional information, using Ekman's six emotional categories as the basis. The experiments conducted encompass both single and multi-emotion labels for text samples. To determine the emotions associated with individual words, they draw upon the WordNetAffect and NRC Emolex lexicons. Their findings reveal that utilizing both lexicons simultaneously yields superior results compared to using either one in isolation. Sentence vectors are constructed by

summing the word vectors. In the experimentation phase, the word vectors are utilized as input features for L2-regularized multi-class logistic regression and support vector machine classifiers to predict emotional labels. The proposed method is benchmarked against traditional embeddings like GloVe and Word2Vec, and the results demonstrate that the model incorporating single emotion labels and a combination of both lexicons outperforms other established baseline approaches.

The second example from Table 4 is the work of Seyeditabari et al. (2019). This refinement is achieved by introducing a secondary training stage to embedding models. The findings presented suggest that this strategy notably improves the initial Word2Vec and GloVe models, leading to a 13% enhancement in the case of Word2Vec and a substantial 29% improvement for GloVe in tasks related to emotion identification. This re-training process primarily relies on three objective functions. The first function aims to reduce the angular distance between words associated with the same emotion while the second function seeks to increase the distance between words belonging to opposite emotion categories. The final function is employed to maintain the overall structure of the original vector space. According to Seyeditabari et al. (2019), models incorporating emotional information outperformed the original model, and the average similarity between opposite groups decreased across all vector space models employed in the study, including Word2Vec, GloVe, fastText, and ConceptNet Numberbatch (Speer et al., 2018).

Lee, Park and Choi (2019) introduced a comprehensive model for detecting emotions in English texts. Their approach covered word-level, utterance-level, and dialogue-level emotion analysis and relied on the NRC lexicon. To create emotion vectors encompassing Ekman's six emotions and a neutral category, they made adjustments to the skip-gram model. Since utterances consist of words, the first step was to determine the emotions associated with individual words. The study emphasized that a word could express multiple emotions across different utterances. Building on this concept, the authors assumed that words within an utterance shared similar emotions. Consequently, they modified the skip-gram model to include emotional information in word vectors. To facilitate their semi-supervised learning algorithm's training, they used the NRC emotion lexicon to label each term. Then,

to determine the overall emotion of utterances of varying lengths, they combined the vectors of constituent words obtained from their previous word-level embedding model. However, detecting the emotion of dialogues posed a challenge due to contextual complexities. To address this issue, they incorporated contextual LSTM (Long Short-Term Memory) into their approach.

Mao et al. (2019) presented an alternative approach to emotion detection, distinct from NRC lexicon-based methods. Their method forms a composite word representation by merging semantic and emotional embeddings. This method calculates the cosine similarity between the vectors of all words in the vocabulary and those of all emotional words in the lexicon. For each word, it identifies the vectors of the n nearest emotion-related words, adjusts them using a normalized weight value, and then generates the emotional word vector by averaging this collection of weighted vectors. The weight of each word in the lexicon is determined by multiplying its cosine similarity score with the degree of emotion as designated by the lexicon annotators. Then, these emotional vectors are integrated with the original embeddings to produce hybrid word representations. These hybrid vectors are then assessed for their performance in emotion identification using classifiers like support vector machine, logistic regression, decision tree, and gradient boost models. The study highlights that the classification accuracy for negative emotions, such as fear and disgust, is lower when compared to emotions like happiness and trust. According to Mao et al. (2019), this disparity may arise because there are more texts within the positive emotion categories than in the negative ones. Additionally, the research finds that employing hybrid vectors as input for classifiers yields better results than using the original semantic embeddings (Mao et al., 2019).

Wang and Meng (2018) conducted a study involving the utilization of a combined lexicon and vector similarity measure. They introduced a multi-emotion category model, which clusters word vectors based on both semantic similarity and shared emotional information. This model assigns an 8-dimensional vector, following Plutchik's emotional wheel, to each word and can be applied to any pre-existing word vector. The researchers merged multiple existing lexicons, resulting in a unified lexicon of 14450 words rather than devising a lexicon exclusively for the Chinese

language. Despite the distinct dimensions of these lexicons, they harmoniously integrated them to provide each word with a vector representing Plutchik's eight emotions. The process involves determining the semantic similarity between the target word's semantic vector and that of other words using cosine distance. Identifying the top k nearest neighbors follows this calculation. Subsequently, these neighbors undergo further ranking based on their emotional similarity, which is assessed through Euclidean distance. Following the initial ranking of nearest neighbors, a refinement step takes place. During this phase, adjustments are made to pre-trained word vectors to bring them closer to words with similar emotional characteristics and move them farther away from words with dissimilar emotional characteristics. This adjustment is executed by applying the objective function provided below in Equation 4:

$$arg\,minF(V) = arg\,min \sum_{i=1}^{n} [p_1 dis(v_i^{s+1}, v_i^s) + p_2 \sum_{j=1}^{k} w_{ij} dis(v_j^{s+1}, v_j^s)] \tag{4}$$

Here, $V = \{v_1, v_2, ..., v_n\}$ represents n number of vectors. v_j stands for one vector from the top k nearest neighbors of a target vector v_i. $dis(v_i, v_j)$ represents the distance between two target vectors, w_{ij} defines the weight of the vector v_j concerning the target vector v_i. The word vector in sequential steps s and $s+1$ are represented by v^s and v^{s+1}. Lastly, two parameters, denoted as p_1 and p_2, serve the purpose of preventing an excessive concentration of words in the same location. The researchers assessed the performance of the multi-emotion category model across various datasets, employing CNN and Bi-LSTM classifiers in their experiments. Throughout these experiments, they compared several word embedding techniques, including Word2Vec and HyRank (sentiment embedding), against their proposed approach. As reported in the study by Wang and Meng (2018), their model, as well as HyRank, outperformed the traditional Word2Vec model. Moreover, their fine-tuning model yielded enhancements for both embedding methods, namely HyRank and Word2Vec.

In the research of Alshahrani et al. (2019), they departed from the commonly utilized cosine distance and instead employed the Euclidean distance metric for emotion detection. In essence, their approach involved the creation of "idealized" word vectors for the purpose of emotion detection using Word2Vec (Alshahrani et al., 2019).

To achieve this, they employed the covariance matrix adaptation evolution strategy (CMA-ES), an iterative evolutionary algorithm developed by Hansen et al. (2003). The method, grounded in CMA-ES, continuously updated these "idealized" vectors for each specific emotion. They utilized the Twitter Emotion Corpus (TEC) (Mohammad, 2012) dataset to determine the optimal vector for each emotion. To ascertain the emotional content in texts by measuring the distance between word embeddings and the "idealized" emotional vectors, they relied on the Euclidean distance function.

In contrast to previous research, the study of Su et al. (2018) employed a dataset that included valence and arousal measures. They introduced an LSTM model for emotion detection, using both semantic and emotional vectors as inputs to the learning machine. Semantic word vectors were generated with Word2Vec, and emotional vectors were derived by projecting lexical words into the emotion space using Chinese Valence Arousal Words (CVAW) (Yu et al., 2016). Their approach encompassed various tasks during both training and testing phases, including word segmentation, word embedding, emotion space projection, and bottleneck feature extraction. Specifically, an auto-encoder was employed in the bottleneck feature extraction task to reduce the dimensionality of emotion word vectors and obtain bottleneck features. Su et al. (2018) focused on seven basic emotion categories: anger, sadness, happiness, disgust, boredom, anxiety, and surprise. As no dataset containing all seven emotion categories was available, they expanded the Natural Language Processing and Chinese Computing (NLPCC) dataset from the Shared Tasks 2014 (*NLPCC Evaluation Tasks*, 2014), which included five emotion categories but lacked boredom and anxiety. Their LSTM-based approach, which incorporated both semantic and emotional vectors, outperformed models that relied on single-feature vectors. Additionally, their model demonstrated a significant 5.33% improvement in accuracy when compared to a CNN-based method.

Jiwung Hyun and Cheong (2020) presented a solution for the CL-AFF (Computational Linguistics - Affect Understanding) shared task in 2020, which integrates deep learning techniques to combine emotion and language embedding models. In this approach, they leveraged the models introduced by Seyeditabari et al. (2019) for emotion embeddings and BERT (Devlin et al., 2018) for word embeddings.

Additionally, they utilized a dataset aimed at exploring the role of emotions in conversations (Jaidka et al., 2020). The experimental findings revealed that BERT models outperformed GloVe models, and combining BERT and GloVe with emotional GloVe embedding preferences led to enhancements in classification performance.

Wu and Jiang (2019) conducted research with the objective of emotion identification and successfully predicted valence ratings of text by incorporating latent semantic information from neural network layers and using GloVe embeddings. They employed the Stanford Emotional Narratives Dataset (SEND) (Ong et al., 2019), which consists of transcripts of video recordings containing emotional narratives. In analyzing the experimental results of their proposed model, they observed that not all dimensions of GloVe embeddings were equally informative for emotional valence. Surprisingly, they found that the 34th dimension of GloVe word embeddings was particularly expressive in terms of emotion valence. The transformation of the original GloVe embeddings into the proposed emotional space yielded improved results for emotional arithmetic. Lastly, they demonstrated that vectors exhibited better projection in terms of their emotional polarities in the proposed method when compared to the raw GloVe vectors.

The approach of Wongpatikaseree et al. (2021) involves training an embedding model that is sensitive to sentiment words. This model is then used as input for a convolutional neural network, enabling the classification of four emotions using Word2Vec. Their findings highlight the embedding model's improved ability to distinguish between words with similar or contrasting emotional connotations. A drawback, however, is the limited amount of pretraining data available for transformer models.

Matsumoto et al. (2022) introduced a model that merges emotion and semantic knowledge. To acquire emotional embeddings, they employed the pre-trained DistilBERT model and predicted the emotions associated with the given words. They conducted two experiments for performance evaluation, focusing on classifying emotional expressions and utterances. Their findings indicated enhanced performance when using emotional embeddings as opposed to the original vectors from DistilBERT. However, it's important to note that the performance improvement achieved with emotional embeddings was not uniform across all the datasets they employed.

3.2.2. Sentimental Vectors

Sentiment analysis, also known as opinion mining, involves the computational examination of people's opinions and attitudes towards various entities, including individuals, events, or topics, as discussed in (Medhat et al., 2014). It is a common observation within this field that sentiment detection studies typically consider fewer categories compared to emotion identification. Specifically, sentiment analysis commonly revolves around the identification of three categories: positive, negative, and neutral attitudes towards the subject entity, hence why it is often referred to as polarity detection.

The detection of sentiment or polarity within a text unit can be accomplished through various resources, such as lexicons, keyword lists, or labeled corpora used as learning inputs. In the following discussion, our focus will be on studies where text units are represented by vectors containing sentimental information. In these studies, existing vectors, like Word2Vec and GloVe, are enhanced by integrating sentimental data obtained from the resources mentioned earlier.

Table 5 provides an overview of several studies introducing sentimental vectors, outlining the base models used, data sources, and their evaluation results. Afterward, we will provide comprehensive insights into the related studies.

Table 5. Sentimental Vectors

Reference	**Base Model**	**Dataset**	**Lexicon**	**Evaluation**
(Naderalvojoud and Sezer, 2020)	Feedforward neural network model	SemEval-2013 (Nakov et al., 2019) SST (Socher et al., 2013)	E-ANEW (Warriner et al., 2013) Subjectivity clue lexicon (Wilson et al., 2005)	For binary task, accuracy increased from 85.8% to 87.4% for LSTM, from 86.3% to 87.2% for BiLSTM
(Yu et al., 2017)	CNN, LSTM, Deep averaging network (DAN)	SST	E-ANEW (Warriner et al., 2013)	In binary classification, 3.5% increase in accuracy between Word2Vec and its refined WordVec when using DAN model.
(Tang et al., 2014)	C&W model	SemEval-2013	MPQA (Wilson et al., 2005) HL (Hu and Liu, 2004)	Accuracy: 71.74%-77.33%
(Chang et al., 2019)	CBOW	NLPCC (*NLPCC Evaluation Tasks*, 2014)	CVAW (Yu et al., 2016)	F1 score: 72.39%
(Sweeney and Padmanabhan, 2017)	Word2Vec	Collection of Tweets	SentiWordNet (Baccianella et al., 2010)	Accuracy: 69% - 71%

(Wu, Wu, Liu, Huang and Xie, 2019)	LSTM	SemEval, Amazon product reviews, Chinese product reviews (Tan and Zhang, 2008)	(Hu and Liu, 2004)	Accuracy: 82.18%-88.06% for Amazon Reviews.
(Lei et al., 2018)	CNN	SST and Movie Reviews(MR) (Pang et al., 2002)	(Qian et al., 2017), (Hu and Liu, 2004)	49.7% for SST 84.3% for MR
(Shi et al., 2018)	Skip-gram	Amazon reviews (Blitzer et al., 2007)	HL (Hu and Liu, 2004), MPQA (Wilson et al., 2005)	Accuracy: 79.4%-85.6%
(Wang et al., 2021)	Word2Vec GloVe	SemEval (Nakov et al., 2019) SST1, SST2 (Socher et al., 2013) IMDB (Pang and Lee, 2005) Amazon (Health) YELP (Restaurant)	Fusion Sentiment Intensity Lexicon (FSIL) (Wang et al., 2021)	F1: 46.9 % - 90.1%
(Kasri et al., 2022)	CBOW	IMDB (Maas et al., 2011)	SentiWordNet (Baccianella et al., 2010) SenticNet (Cambria et al., 2018) VADER (Hutto and Gilbert, 2015)	Accuracy: 88.6%
(Rezaeinia et al., 2019)	Word2Vec and GloVe	Movie Reviews (Pang and Lee, 2005) Customer Reviews (Hu and Liu, 2004) SST, SST1 (Socher et al., 2013) Rotten Tomatoes movie reviews dataset (Hu and Liu, 2004)	Combination of Lexicons	Accuracy: 43.4%- 80.3%
(Sharma, 2022)	Word2Vec	SST (Socher et al., 2013)	(Warriner et al., 2013)	Accuracy Fine-grained: 49.2% Binary: 88.6%

The study of Naderalvojoud and Sezer (2020) exemplifies the construction of sentimental vectors through the use of lexicons. Their proposed methodology comprises two distinct approaches. The first approach focuses on refining existing pre-trained embeddings for sentiment classification, employing a feed-forward neural network that predicts the polarity of embeddings by incorporating a combination of two sentiment lexicons: an extended version of Affective Norms of English Words (E-ANEW) (Warriner et al., 2013) and Subjectivity Clue lexicons (Wilson et al., 2005). This method not only includes sentiment information but also the semantic understanding of words. The second approach aims to differentiate sentimental relationships between words instead of dealing with their contextual associations. Various classifiers, such as convolutional neural networks, long short-term memory networks, bidirectional LSTMs, and logistic regression, were applied, yielding promising results in comparison to sentiment analysis studies that utilize deep learning methods. It is worth noting that the vocabulary of documents often contains more words than the lexicons, resulting in some words being assigned a neutral sentiment polarity, which

could limit the model's performance.

Yu et al. (2017) introduced an alternative refinement method for pre-existing word vectors, Word2Vec, and GloVe, leveraging the E-ANEW sentiment lexicon (Warriner et al., 2013). In their model, vector representations of words are expected to be closer if the words share both semantic and sentimental similarities. To refine the pre-trained vectors, the model initially computes the semantic similarity (using cosine distance) between each target word in a given text unit and the lexicon words. Subsequently, it selects the k semantically most similar words as the nearest neighbors for each target word. By re-ranking the semantically similar words based on sentiment scores obtained from the lexicon, the words that share both semantic and sentimental similarities are brought closer in the vector space. Let $V = \{v_1, v_2, ..., v_n\}$ represent the word vectors in the sentiment lexicon. The model iteratively minimizes the distance between each target word and its top-k neighbors. The objective function denoted as $\Phi(V)$, is defined in Equation 5 where n is the number of words that will be refined, v_i and v_j is the target word vector and is its nearest neighbors vector, respectively. The weight, denoted as w_{ij} and as explained by Yu et al. (2017), is determined based on the ranked list. Their experimental findings indicate that their proposed approach outperforms traditional and sentimental word embeddings.

$$\Phi(V) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} * dist(v_i, v_j) \tag{5}$$

Tang et al. (2014) introduced a method that integrates sentiment information as an extension to an existing word embedding approach. Their study involved the development of three neural networks designed to learn word embeddings specifically for sentiment analysis on Twitter data. They utilized widely recognized sentiment lexicons, namely Multi-perspective Question Answering (MPQA) (Wilson et al., 2005) and HL (Hu and Liu, 2004), as part of their methodology. In contrast to other research, they selected a distinctive approach by using n-grams as input to the neural network and predicting the sentiment polarity using a sliding window-based technique. Furthermore, they collected Tweets containing positive and negative emoticons through distant supervised corpora, thereby avoiding the need for manual annotation (Tang

et al., 2014). Their findings indicated that traditional word embeddings like C&W and Word2Vec exhibit lower performance when compared to their embedding model enriched with sentiment information.

Chang et al. (2019) proposed a refinement technique that differs from the previously discussed methods by utilizing a dataset featuring arousal and valence values. In their study, they focused on the classification of Chinese movie reviews using support vector machines, with word representations incorporating sentiment information. These word representations were combined with sentiment information obtained from arousal and valence predictions. The 2016 Chinese Valence-Arousal Words (CVAW) (Yu et al., 2016) containing words annotated with arousal and valence values are utilized as training data. Due to the insufficiency of the dataset for sentiment analysis of movie reviews, they extended it using two distinct approaches. In the first approach, they assumed that synonyms would have similar valence and arousal, meaning they would be close in terms of valence and arousal values. The second approach involved the use of Word2Vec, where the valence of the target word was computed as the average of the valence values of its neighboring words. Arousal prediction relied on the average results of two prediction-based methods, specifically linear regression and support vector machine-based predictions. For their refined embedding model, they combined word embeddings generated from the CBOW model with their arousal and valence values. The study encompassed experiments to evaluate the performance of predicting valence and arousal values for words, as well as sentiment analysis of movie reviews. As noted in their research, one limitation they encountered was the handling of negation words.

Unlike most sentiment analysis studies that assign a single sentiment to an entire text, Sweeney and Padmanabhan (2017) took a different approach by attempting to attribute sentiments to individual entities within the text. They recognized that entities within the text may have varying sentiments. To achieve this, they performed preprocessing steps, including part-of-speech tagging and dependency parsing, and classified the given text as either a single-entity or multi-entity text. For single-entity texts, they utilized a Word2Vec model to determine the overall sentiment polarity of the entire text, which is then classified using a random forest classifier. On the other

hand, in multi-entity texts, they identified descriptor words that can carry different sentiment information and assign polarity scores to them using the SentiWordNet lexicon (Baccianella et al., 2010). In multi-entity texts, a vector containing polarity scores is generated for each descriptor entity, and the text's sentiment is determined by considering all of these composing descriptors. The proposed model's performance is assessed using Twitter data, and the reported accuracy ranges between 0.69 and 0.71 in various experiments.

Wu et al. introduced two approaches for sentiment classification in both English and Chinese languages. Firstly, sentiment lexicons are employed to classify the sentiments of words by examining their hidden representations in a neural sentiment classifier. This model was trained to identify words that carry sentiment, as these are crucial for determining the sentiment of the text they are part of. The second approach involved using a sentiment lexicon to obtain word embeddings that are aware of sentiment. Both of their models incorporated LSTM for sentiment detection at the sentence level. They evaluated their methods using three datasets: SemEval 2016 [1], Amazon product reviews (He and McAuley, 2016) and Chinese reviews (Tan and Zhang, 2008). Based on their experimental results, their models demonstrated superior performance compared to traditional machine learning techniques and methods relying on sentiment lexicons for sentiment classification.

Lei et al. (2018) proposed a three-layer network called the sentiment-aware attention network, which performed word-level correlation, phrase-level correlation, and sentence-level semantic modeling in each layer. In the first layer, they examined the correlation between the context words in a given text and sentiment-related words using GloVe embeddings. In the next layer, a dynamic attention mechanism was used to identify important and distinctive phrases. The final layer integrated these phrase-level and word-level correlations to create a sentiment-specific sentence representation. They evaluated their model on the Movie Review (Pang et al., 2002; Pang and Lee, 2004, 2005) and Stanford Sentiment Treebank (Socher et al., 2013) datasets, reporting accuracy scores of 84.3% and 49.7%, respectively.

In a group of studies focusing on sentiment analysis, researchers considered

[1] https://alt.qcri.org/semeval2016/task4/

separate word embeddings for the same word in different domains (Yang et al., 2017; Bollegala et al., 2014, 2015). The idea behind this was the recognition that words may have different sentiments (positive or negative) in various domains. Exemplifying the study of Shi et al. (2018), the word "good" generally conveys a positive sentiment, but the word "lightweight" is domain-specific, implying positivity in the electronics domain and negativity when describing movies that lack depth. In the research, to distinguish between domain-common and domain-specific words, sentiment labels and context words were utilized (Shi et al., 2018). If a word had similar context words and sentiments across several domains, it was categorized as a domain-common word. Their model incorporated both domain specificity and sentiment information for words. The researchers carried out experiments using Blitzer et al.'s dataset of Amazon product reviews (Blitzer et al., 2007). They compared their word embedding approach with alternative models and established benchmarks ((Mikolov et al., 2013; Yang et al., 2017)). The results demonstrated that their method surpassed the performance of the other models in sentence-level sentiment classification.

(Wang et al., 2021) introduced a sentiment enhancement technique that was implemented on both Word2Vec and GloVe models on datasets SemEval (Nakov et al., 2019), SST1, SST2 (Socher et al., 2013), IMDB (Pang and Lee, 2005), Amazon[2] and YELP[3]. This approach allowed the model to distinguish between sentiment words and context words within a sentence, thereby enabling the consideration of various sentiments for a target word based on its context. The paper combines various features with the original embeddings, such as part of speech tagging, position information, sentiment, and concept information. However, a limitation of the approach is that while it incorporates sentiment concepts for nouns, it does not extend this coverage to verbs, adjectives, and adverbs.

Kasri et al. (2022) employed a neural network architecture resembling CBOW to generate sentiment embeddings by combining it with an emotion lexicon. Principal component analysis (PCA) is then applied to manage the fusion of both semantic and sentiment embeddings. Notably, when modifying the CBOW model, the context words

[2]http://snap.stanford.edu/data/amazon-meta.html
[3]https://www.yelp.com/

yield sentiment embeddings instead of semantic ones. The researchers suggest that enhancing the quality of these embeddings can be achieved by incorporating additional information, like part of speech tagging.

In their work Rezaeinia et al. (2019), a word embedding model was introduced with sentiment enhancement achieved through a series of vectorization techniques. First, they integrated syntactic information into the embeddings by utilizing part of speech tagging. Second, they incorporated six different lexicons containing sentiment intensity scores. For a given word, sentiment scores were extracted from each lexicon and concatenated to form a vector. Additionally, they created a vector by considering the relative position of target words within sentences, transforming these positions into vectors using a position embedding table. The proposed model combines these three vectorization methods with the semantic Word2Vec and GloVe vectors to generate improved word embeddings. Their findings indicate that these combined embeddings resulted in a sentiment classification accuracy increase of more than 2%. In the study by Sharma (2022), they introduced a model that incorporates both neighbor ranking and refinement techniques. To begin with, an extended version of the E-ANEW emotion lexicon (EANEW) (Warriner et al., 2013) was employed for computing cosine similarity scores between the target words requiring refinement and the entries in the lexicon. During this distance calculation, they leveraged the semantic Word2Vec representations of the lexicon words. After identifying the top k similar words for each target word, they rearranged the list based on sentiment intensity scores. Notably, the intensity score ranking was determined concerning the proximity to the intensity score of the target words, as opposed to sorting the list from largest to smallest. In the refinement phase of the model, the pre-trained word vectors of the target words were enhanced by employing neighbor ranking. This process aimed to bring the vectors of the target word closer to neighbors that were both sentimentally and semantically similar. When calculating the refined vector for a target word, greater weight was assigned to the words ranked highest in the list of the top k similar neighbors. Based on their findings, the use of refined Word2Vec embeddings led to an accuracy increase of 4.4% in classification tasks with CNN and 2.2% with Bi-LSTM.

CHAPTER 4: ENRICHING VECTORS WITH EMOTIONAL CONTENT

The term *text unit* is a flexible concept that refers to distinct or meaningful portions of text in the studies of language and text analysis. Depending on the context, a text unit can encompass various levels of linguistic analysis. It may be as basic as a single character, useful for text processing and character-level tasks, or extend to a word, which is a fundamental unit in language processing and machine learning. Text units can also refer to larger segments like phrases or sentences, relevant in tasks such as text comprehension and sentiment analysis, or even encompass entire paragraphs, documents, or articles, which are significant in document classification, information retrieval, and summarization tasks. The choice of text unit depends on the specific objectives and requirements of the particular linguistic or text analysis task at hand.

For instance, word vectors, often referred to as word embeddings, are numerical representations of words. These representations consist of a series of numeric values that serve as a way to identify and capture various relationships involving the word. One practical application of these word vectors is to determine the semantic similarity between words by assessing the spatial distance between their corresponding vectors in a vector space. While word embeddings have proven effective for extracting semantic similarities, their performance in the realm of emotion detection has been somewhat limited for a variety of reasons. An illustration of this is seen in approaches that assess the semantic proximity of words through their co-occurrence frequencies, exemplified by techniques like Word2Vec and GloVe, where words are positioned near each other in the vector space. These methods tend to treat emotionally opposing words (e.g., *love* and *hate*) as similar words due to their frequent usage in similar contexts, which, as one might predict, significantly impairs the ability to differentiate/categorize the emotions of words or other text units.

Intending to advance natural language processing applications, the integration of emotional understanding into textual data has emerged as a significant research

area. This integration allows us to apprehend not merely the literal meaning of words and sentences, but also the emotional subtle nuances within them. To achieve this objective, researchers initiated a process to augment vector representations with emotional content. In this thesis, this process progressed in two distinct phases: *word-level enrichment* and *sentence-level enrichment*. While *word-level enrichment* tries to better represent the emotional connotations of individual words, *sentence-level enrichment* aims to better capture the emotional context of entire sentences. In the context of these studies, we addressed the categorical emotions defined by Plutchik (1980).

4.1. Enriching Word and Sentence Vectors

Recent methods for enhancing the emotional content of text involve comparing a selected set of words in the text to a list of words associated with specific emotions found in an emotion lexicon. The comparison is performed using various techniques, and the resulting similarity scores are used to adjust the vector representations of the words in the text. The effectiveness of this emotion enrichment process is typically assessed using metrics such as improvements in similarity within emotion categories or performance in emotion classification. The alternative method that demonstrates the most significant improvement is generally preferred.

In the existing literature, emotion enrichment is commonly carried out using emotion lexicon words (e.g., Agrawal et al. (2018); Su et al. (2018); Wu and Jiang (2019)). This involves evaluating the similarity between the vector representation of words in a sentence and the lexicon words. This approach essentially has two inputs: one representing the words in the text and the other representing the lexicon words. We have reservations about this approach as it introduces considerable uncertainties into emotion detection systems. To the best of our knowledge, the impact of enriching emotions based on text units other than individual words has not been adequately explored. Below, we highlight three key issues with the current methods:

1. The conventional approach of constructing emotion lexicons involves selecting random sentences from a dataset and labeling their emotions, associating specific

emotions with the words in those sentences. This process often neglects the potential variations in emotion that a word can convey in different contexts. Consequently, when a word in a given sentence is found in the emotion lexicon, it is assigned with the specified emotion label, without considering other potential emotions it might express in that specific context.

2. Previous studies that use individual words within a sentence as the basis for emotion enrichment assume that a specific group of words collectively represents the emotion conveyed in the entire sentence. The validity of this assumption, however, remains uncertain. Moreover, it is unclear how these approaches establish, recover, or modify sentence representations.

3. Studies that focus on enriching individual words for emotion often ignore the potential changes in emotion when these words combine to form collocations or multiword expressions within a sentence.

In this thesis, in addition to experimenting with previous approaches for out Turkish dataset, we suggest a different approach to address these limitations in emotion enrichment. Our proposal involves using sentence vectors as the input for the emotion enhancement process, as opposed to using vectors of individual words. Additionally, we recommend evaluating the entire emotion enrichment process at the sentence level by employing emotion sentences instead of emotion lexicon words. To sum up, in contrast to previous studies on emotion enrichment, we conducted experiments in which we:

1. Enhanced with sentence vectors rather than vectors of individual composing words.

2. Utilized emotion sentences instead of emotion lexicon words.

3. Utilized emotion-enriched versions of the original emotion lexicon words.

CHAPTER 5: EMOTION ENRICHMENT EXPERIMENTS

In this chapter, we present word-level and sentence-level emotion enrichment experiments. Firstly, Section 5.1 details the experimental setup, including the original embedding models, datasets, and emotion enrichment models utilized. In Section 5.2, we provide the results of the emotion enrichment experiments. This chapter addresses the research questions outlined below:

RQ1 - What is the most efficient original word/sentence embedding method for enhancing the detection of emotions in Turkish texts, thereby improving the performance of emotion detection studies?

RQ2 - Can enhanced representations of words and sentences outperform their original counterparts?

RQ3- Is the efficacy of original and enhanced representations subject to variation based on emotion categories?

RQ4 - Which emotion enrichment methods give better results on word-level and sentence-level emotion detection?

5.1. Experimental Setup

This research encompasses a range of experiments in emotion enrichment conducted at both the word and sentence levels. These experiments explore various configurations and settings, taking into account multiple parameters, including language, embedding models, emotion enrichment models, and evaluation methods. The following sections will provide comprehensive information on the datasets used for Turkish and English, summarize the embedding and enrichment models applied in the study, and outline the evaluation methodologies employed.

5.1.1. Embedding Models

For applying any emotion enrichment technique, the first step is utilizing an original embedding model with a specified dimension, denoted as d. In this research

for *emotion-enriched word vectors*, we employed two semantic embedding models, Word2Vec (with d=400) and GloVe (with d=300), along with a contextualized embedding model, BERT (with d=768 default vector length of BERT-base). Each of these word embedding techniques has its distinct advantages and limitations. The choice of which technique to utilize may vary depending on the specific NLP application. One advantage of utilizing GloVe, Word2Vec, and BERT vectors for enriching emotions is that they are widely adopted and have demonstrated effectiveness in numerous natural language processing tasks.

GloVe and Word2Vec represent unsupervised techniques that acquire embeddings by analyzing co-occurrence patterns within extensive textual data collections. They are particularly suited for tasks related to enhancing emotions and are frequently used as benchmark methods due to their ability to capture the semantic connections between words, as evidenced in prior research like the works by (e.g., Agrawal et al., 2018; Wongpatikaseree et al., 2021; Kasri et al., 2022).

In contrast, BERT vectors exhibit a higher degree of context sensitivity and are presently employed in a variety of emotion analysis tasks, as demonstrated in studies such as those conducted by (Chiorrini et al., 2021) and (Singh et al., 2021). To generate Word2Vec vectors, we trained a Continuous Bag of Words (CBOW) architecture using comprehensive Turkish Wikipedia articles. Subsequently, we extracted pre-trained Word2Vec and GloVe vectors for all words associated with one of the 8 emotions present in the dataset.

In the context of *enriching emotions at the sentence level*, two transfer learning-based models, BERT Devlin et al. (2018) and its variation known as DistilBERT (DBERT) (Sanh et al., 2019), are employed. The decision to utilize BERT and BERT-based models in this study is driven by two primary considerations. Firstly, these models stand out from their predecessors by offering both word and sentence-level embeddings. Secondly, they have demonstrated impressive performance in emotion classification for Turkish and English, as evidenced by previous research (Uçan et al., 2021; Abas et al., 2022; Abubakar et al., 2022). DistilBERT was specifically chosen among the various BERT variants due to its streamlined architecture, which allows for 60% faster operations while retaining 95% of BERT's functionality (Sanh et al., 2019).

To generate sentence embeddings using BERT and DistilBERT, we made use of pre-trained models from the Hugging Face library. For English sentences, we utilized model versions that were trained on English texts, and for Turkish sentences, we employed models trained on Turkish texts (Schweter, 2020).

5.1.2. Datasources

To create *emotion-enriched word vectors*, the initial step involved translating the NRC emotion lexicon (Mohammad and Turney, 2013) into Turkish (referred to as TT-NRC (Aka Uymaz and Kumova Metin, 2023a)). NRC words with matching Turkish translations were excluded from the dataset. As a result, there remained 4825 word-emotion pairs in TT-NRC, each of which was annotated with Plutchik's 8 emotions, as shown in Table 6.

Table 6. The statistics of TT-NRC.

Emotion	**# of words**
Anger	703
Anticipation	491
Disgust	595
Fear	888
Joy	403
Sadness	708
Surprise	292
Trust	745
Total	4,825

In this research, we utilized BERT's dynamic and context-aware embeddings, which enable it to generate varied embeddings for the same word depending on its contextual usage. To obtain BERT word vectors for words associated with emotions in our dataset, we initially compiled a set of sentences that have emotion labels. These sentences were gathered from three datasets: TEI (Mohammad and Bravo-Marquez, 2017), TEC (Mohammad, 2012), and TREMO (Tocoglu and Alpkocak, 2018). TEI and TEC datasets were originally in English, but we translated them into Turkish for our analysis. TREMO, on the other hand, is a Turkish dataset. You can find a quick overview of these datasets in Table 7 and more detailed explanations in Sub-section 3.1.2.

Table 7. Utilized datasets.

Dataset	Emotion categories	Data Size
TEI (Mohammad and Bravo-Marquez, 2017)	anger, fear, joy, sadness	7,097
TEC (Mohammad, 2012)	anger, fear, disgust, joy, sadness, surprise	21,051
TREMO (Tocoglu and Alpkocak, 2018)	happiness, anger, fear, sadness, disgust, surprise	27,350

As can be seen in Table 7, the emotion categories across the datasets exhibit variation. Consequently, we identified four emotions that are consistent among these datasets: anger, fear, joy, and sadness. We acquired BERT vectors for words from the TT-NRC lexicon associated with these emotional categories. To obtain the BERT vector for a specific lexicon word, we followed these procedures:

1. We counted the sentences in the collection that contain the target word. On average, each word in the lexicon is found in approximately 9 sentences in the datasets. As a result, we established a threshold of 9 and randomly selected a maximum of 9 sentences that include the target lexicon word. During the search for these sentences with target lexicon words, we employed a Turkish lemmatizer (Akın and Akın, 2007) for both the search words and the constituent words of the sentences.

2. The target lexicon words and the sentences containing related words from the collection were provided as inputs to BERT, and BERT generated word vectors for each lexicon word. To illustrate, if the word ($word_a$) is found in a set of 9 sentences (($sentence_1$, $sentence_2$,...$sentence_9$), ($word_a$) is associated with each sentence in the set. Each pair ($word_a$, $sentence_i$) is then processed by BERT individually, and the sum of the vector outputs is accepted as the BERT vector for ($word_a$).

For *sentence-level emotion enrichment*, we again employed the English datasets listed in Table 7. To work with both English and Turkish languages, we performed translations: converting the English datasets into Turkish and the Turkish datasets into English. We identified four prevalent emotions across these datasets: anger, fear, joy, and sadness. Following this, we randomly handpicked 500 sentences from each emotion category within the dataset collection.

5.1.3. Emotion enrichment models

The process of emotion enrichment, as defined in Seyeditabari et al. (2019), involves refining the embeddings to capture emotional content. In short, the common goal of studies in this field is to represent words that are semantically and emotionally close to each other in vector space. In this study, three methods were employed for emotion enrichment at both the word and sentence levels. These methods will be named EEA1, EEA2 and EEA3.

The first method, EEA1, utilized is the proposed approach of Seyeditabari et al. (2019). Plutchik's wheel of emotions was chosen as the emotional model for their training. They modified the initial vector space by introducing emotional constraints and their approach involved utilizing the NRC emotional lexicon to establish two distinct sets of constraints. The sets are;

$$S = \{(w_1, e_1), (w_1, e_3), (w_2, e_2), ...\} \tag{1}$$

$$O = \{(w_1, e_1^{'}), (w_1, e_3^{'}), (w_2, e_2^{'}), ...\} \tag{2}$$

where (w_i), (e_i), and $(e_i^{'})$ represent the words, their related and opposite emotion categories, respectively. The opposite emotion category of words is found using Plutchik's emotion model, where each emotion has an opposite emotion in it, such as sadness being the opposite of joy. In the approach presented by Seyeditabari et al. (2019), the main goal of the objective function is to minimize the angular distance between words and their corresponding emotions, while concurrently expanding the distance from emotions regarded as opposites. Therefore, they intended to decrease the distance between word pairs in the positive relation set while augmenting the distance between word pairs in the negative relation set. In summary, as seen below in Equations 3 and 4, they defined two separate objective functions for positive relations (PR) and negative relations (NR), where u and w are pairs of words in the sets (S and O) and $d(v_u, v_w)$ represents the distance between 2 vectors (v) of words (u and w).

$$PR(V^{'}) = \sum_{u,w \in S} max(0, d(v_u^{'}, v_w^{'})) \tag{3}$$

$$NR(V^{'}) = \sum_{u,w \in O} max(0, 1 - d(v_u^{'}, v_w^{'})) \tag{4}$$

Then, to preserve the original vector space's properties, a third part is included in the final objective function which is VSP (Equation 5).

$$VSP(V, V^{'}) = \sum_{i=1}^{N} \sum_{j \in N(i)} max(0, | d(v_u^{'}, v_w^{'}) - d(v_u, v_w)) \tag{5}$$

$$Obj(V') = PR(V') + NR(V') + VSP(V, V') \tag{6}$$

As can be seen in Equation 6, the final objective function includes (PR), (NR) and (VSP).

In our experiments, secondly, we applied the emotion enrichment approach, EEA2, proposed by Mao et al. (2019). Their objective was to create a combined textual representation, referred to as sentiment-aware word embedding, specifically designed for emotion detection tasks (Mao et al., 2019). This combined representation, often referred to as the hybrid vector, was constructed by merging emotional vectors with word embeddings. The word embeddings were generated using the skip-gram model within the Word2Vec architecture and were considered to represent the semantic meaning of the given text unit. To construct the emotional vector, Mao et al. (2019) compared the embeddings of words in the text to the embeddings of emotional words contained in the DUTIR emotion lexicon (Chen, 2008). This lexicon, which consisted of emotion-related words in the Chinese language, was categorized into seven emotion classes, including happiness, fear, surprise, anger, trust, sadness, and disgust. In their study, the comparison process involved calculating the cosine distance between each constituent word in the text and each emotion word in the lexicon. The vector representations (V) of all words in the vocabulary having length of m are presented by $V = \{v_1, v_2, ..., v_m\}$ such that $V_E = [V_{E_1}, V_{E_2}, ..., V_{E_k}]$ is the vector space of lexicon words carrying emotions. The cosine similarity between V and E is represented with

the following Equation 7.

$$sim(V,E) = \frac{\sum_{i=1}^{d}(V_i * V_{E_i})}{\sqrt{\sum_{i=1}^{d} V_i^2 * \sum_{i=1}^{d} V_{E_i}^2}} \tag{7}$$

$Y \in \mathbb{R}^{mxk}$ represents the matrix encompassing the outcomes of cosine similarity. By evaluating the similarity among all vocabulary words and words in the lexicon, the matrix Y is utilized to rank and choose the top n emotional words. Word vectors having emotional information (EV) are calculated with Equation 8 and Equation 9 represents the *weight* calculations that are utilized in the calculations of EV.

$$EV_i = \frac{1}{n} * \sum_{i=1}^{n} weight_i * V_{E_i} \tag{8}$$

$$weight_i = Y_{ij} * score_j \tag{9}$$

In Equation 9, the *score* indicates the level of sentiment, and the data is collected from the emotion lexicon. Y_{ij} denotes the degree of similarity between each word in the vocabulary (i) and the corresponding lexical item (j). Lastly, to standardize the weight values, the subsequent equation is employed as a normalization step:

$$weight_i = \frac{weight_i}{\sum_{j=1}^{n} weight_j} \tag{10}$$

The method combines semantic Word2Vec vectors and emotional vectors to create hybrid vectors.

The last model EEA3 is proposed as an alternative to the second model. As previously mentioned, EEA2 is centered around the utilization of similarity scores for the top n lexicon words. In our proposed approach, EEA3, we extend this concept by also taking into account the ranking of lexicon words. To elaborate, after determining the top n words that are most similar to a target word, we arrange them in descending order of their similarity degree. In simple terms, if a lexicon word better represents its associated emotion, we ensure that it receives a higher weight during the normalization step, as described in Equation (11). Consequently, when calculating (EV_i) in accordance with Equation (8), the weight coefficient is updated and increased

for the most closely related word pairs.

$$weight_i = \frac{weight_i}{\sum_{j=1}^{i} weight_j} \tag{11}$$

The reason for selecting these two existing methodologies by Seyeditabari et al. (2019) and Mao et al. (2019) can be considered as follows. Firstly, our study involves experiments on both Turkish and English datasets. Moreover, our primary focus is on the Turkish language as it is a low-resourced language and thus, studied infrequently in the NLP field and emotion detection tasks. The methods, EEA1 and EEA2 were initially developed for the English and Chinese languages, respectively. Since our work involved a language that had not been previously subjected to emotion enrichment, we aimed to adapt these methods to a different linguistic context, seeking potential benefits for our research. Furthermore, these methods were originally designed for semantic embeddings and were applied to assess their impact. Furthermore, we explored how these emotion enrichment techniques affected a contextual embedding model, BERT as well as the semantic models. The selected methods had the advantage of being adaptable for enriching various types of embeddings, which was particularly valuable for comparing the performance of different embedding models. In addition to the use of EEA1 and EEA2, we also introduced an enhanced version of the method proposed by Mao et al. (2019) (EEA3) to facilitate comparisons with the aforementioned techniques.

There are several procedures applied to adapt/utilize emotion enrichment methods both at word and sentence levels. The details can be presented as follows:

1. Considering the *word level emotion enrichment* EEA1, EEA2 and EEA3 methods are utilized. Following the approach introduced by Seyeditabari et al. (2019), three emotionally enriched representations were generated for GloVe, Word2Vec, and BERT vectors, denoted as *EEA1_GloVe*, *EEA1_Word2Vec*, and *EEA1_BERT*.

   In the work of Mao et al. (2019), a text dataset and an emotion lexicon were employed. To construct emotionally enriched vectors (referred to as EEA2), we utilized the TT-NRC emotion lexicon (Turkish manuallty translated version of

NRC lexicon 5.1.2), just as we did for EEA1.

Initially, the TT-NRC lexicon, containing 4825 words labeled with 8 different emotions, was divided into four equal parts, each comprising an equal number of words associated with a specific emotion. At each stage of the process, one of these parts was considered as the vocabulary, representing the words for which we aimed to derive enriched representations. The combined content of the remaining parts was treated as a list of emotional words. After computing the cosine similarity between the words in the vocabulary and the emotional words, we selected the top 10 (n) similar word pairs. We then applied the procedure defined by Mao et al. (2019) and its modified version to create EEA2 and EEA3 vectors, respectively. As outlined in the methods, it was necessary to extract emotion intensity values for the lexicon words when calculating EEA2 and EEA3 vectors. To achieve this, we utilized the NRC Emotion Intensity Lexicon (Mohammad, 2018). In conclusion, in addition to the original embeddings, we generated the following representations: EEA2_GloVe, EEA2_Word2Vec, EEA2_BERT, EEA3_GloVe, EEA3_Word2Vec, and EEA3_BERT. Figure 4 provides a visual representation of the stages involved in creating emotionally enriched word embeddings.

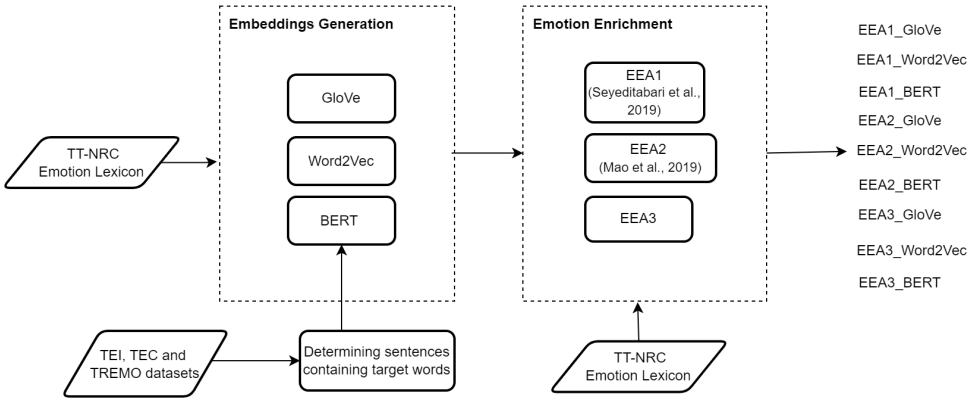


Figure 4. Framework for the *word-level emotion enrichment* experimental study.

2. *Emotion enrichment on sentences*, differs from the original methods used for word-level enrichment (EEA1, EEA2, and EEA3). In the first set of experiments, we enriched sentences using emotion lexicon words, and in the second set, we

enriched sentences using emotion sentences. The procedures for both methods were mostly identical, except for the change in the text unit being enriched. For instance, when enriching sentences with emotion sentences in EEA2 or EEA3, we measured the similarity between emotion sentences and the target sentence to be enriched. The degree of emotion in an emotion sentence was calculated by computing the mean of emotion intensities of the words it comprises according to the emotion category of the sentence. Throughout our research, we employed the original English emotion lexicon and manually translated Turkish versions of the National Research Council Canada (NRC) emotion lexicon (TT-NRC) in all three emotion enrichment methods. Figure 5 presents the steps required to generate emotionally enriched sentence embeddings.

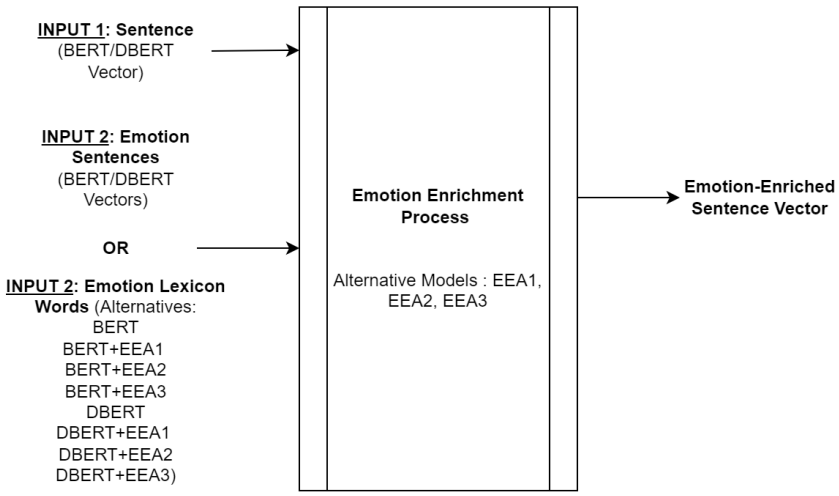


Figure 5. Framework for the *sentence-level emotion enrichment* experimental study.

All the preprocessing, vectorization, and enrichment procedures are implemented using Python 3.9 through Google Colaboratory (Google, 2017), a cloud-based platform based on Jupyter Notebooks. The vector operations are conducted through the NumPy library (Harris et al., 2020).

5.2. Experimental Results

5.2.1. Similarity Measurements and Classificaton

The effectiveness of the emotion-enrichment process can be evaluated by assessing the enhancement in emotion recognition within the text. This study employs two distinct methods for measuring this improvement. Initially, post-enrichment, the study examines the average change in *cosine distance* within the word/sentence set of each emotion category. The distance between two vectors can be quantified using various distance metrics, including *Euclidean, Manhattan, and Minkowski metrics*, as previously utilized in studies like Ratna et al. (2022) and Ranasinghe et al. (2019). However, in our research, we specifically selected the *pairwise cosine similarity score* as a means of assessing vector similarity. This choice is made for ease of comparison with prior research, given its widespread application in similar tasks. Cosine similarity between two text units a and b can be calculated through their vector representations V_a and V_b as follows:

$$cos(\theta) = \frac{\sum_{i=1}^{d}(V_{a_i} * V_{b_i})}{\sqrt{\sum_{i=1}^{d} V_{a_i}^2 * \sum_{i=1}^{d} V_{b_i}^2}} \tag{12}$$

In Equation 12, d represents the length of the vectors. In this context, when similarity scores approach 1, it signifies that a and b are highly similar, while scores close to -1 indicate that they have contrasting meanings. In our research, we employed cosine similarity in two specific contexts. The first is *in-category similarity*, which pertains to the similarity of text units containing words/sentences from the same emotional category. To illustrate, each word within the *joy* category is paired with all other words in the same category, and then similarity scores are computed for each pair. This process is repeated using all embeddings selected in the experiment design for each emotion. The expected outcome is that the average distance between words/sentences with the same emotion label will decrease due to the emotion-enrichment process. In other words, emotionally enriched representations

(e.g., EEA1_Word2Vec) should yield higher similarity scores for pairs labeled with the same emotion compared to their original counterparts (e.g., Word2Vec). Secondly, the cosine similarity score between words/sentences belonging to opposite emotion categories is measured. Finding the opposite emotion categories is handled by the rules of emotion theory that we used. In Plutchik (1980), every emotion is paired with an opposite pair such that *joy* being opposite of *sadness*.

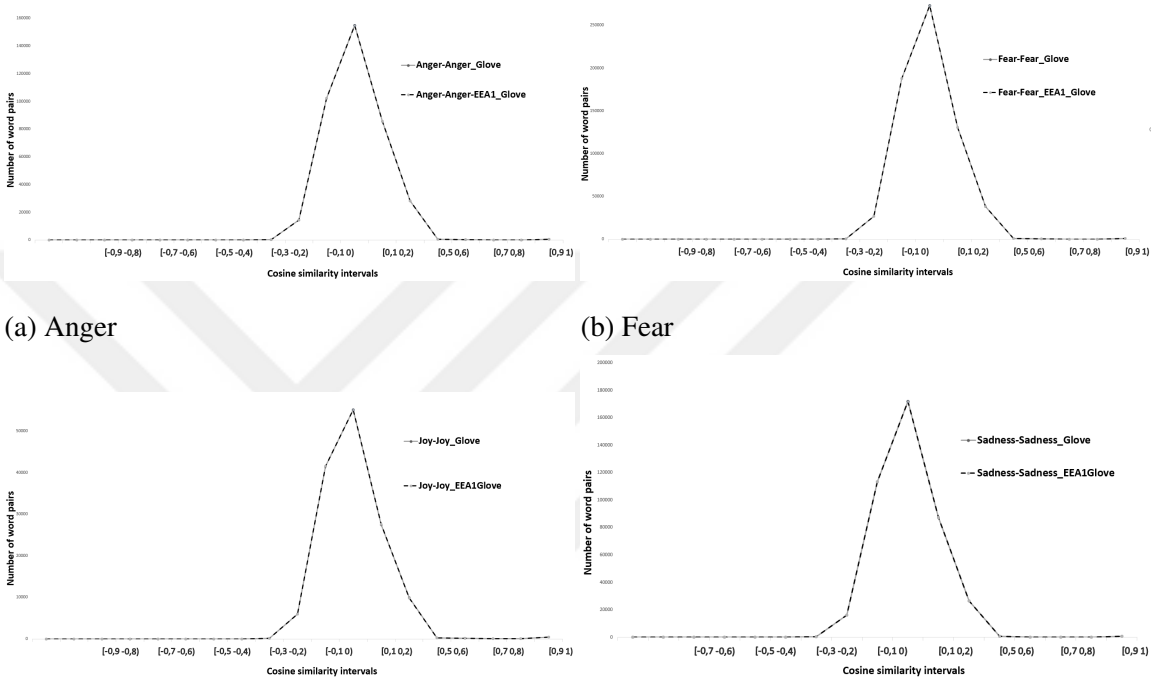
In the second approach, emotion identification is considered a classification task. In essence, text-based emotion identification involves classifying text segments into predefined emotion categories, with the number of emotions determining the task's complexity. For instance, if the given texts are to be categorized into two classes such as positive or negative, effectively framing the task as a binary classification problem, the identification process is referred to as polarity detection and is comparatively simpler compared to multi-class labeling. In this study, words enriched by semantic embeddings (GloVe and Word2Vec) are categorized into 8 categories defined by Plutchik (1980). On the other hand, words enriched by contextual embeddings and sentences enriched by BERT and DistilBERT are assigned 4 main emotions which are anger, fear, joy, and sadness, because of the nature of the data.

5.2.2. Word-Level Emotion Enrichment Experiments

To assess the success of *word-level emotion enrichment* in the Turkish language, we initially extracted GloVe, Word2Vec, and BERT vectors for all the words in the specified TT-NRC lexicon. Subsequently, as explained in Subsection 5.1.3, the original vectors were used to generate their emotion-enriched counterparts. To investigate whether these emotional vectors perform better, two sets of experiments were conducted. The details about cosine similarity and classification experiments are presented in 5.2.2.1 and 5.2.2.2, respectively.

5.2.2.1 Word Level Cosine Similarity Measurements

As mentioned previously, pairwise cosine similarity between each word belonging to the same category is measured using first the original, then emotion-enriched vectors. As illustrated in Figure 6, for instance, there are four histograms corresponding to different emotional categories, using both GloVe and EEA1_GloVe.



(a) Anger (b) Fear

(c) Joy (d) Sadness

Figure 6. Pairwise cosine similarity histograms (GloVe and EEA1_GloVe).

In the provided graphs, the horizontal axis (x) represents similarity score ranges, while the vertical axis y corresponds to the number of word pairs falling within those score ranges. A similarity score closer to 1 indicates a higher degree of similarity between two word vectors. Thus, when we employ emotionally enriched vector representations, we expect a shift to the right in the scores, which means an increase in the median value of the histogram or a greater concentration of values towards the right side.

The outcomes from Figure 6 reveal that the average similarity scores within the same emotion category have experienced a slight increase when using emotional vectors. However, this improvement may not be easily noticeable through graphs.

To make this improvement more apparent, we computed the disparity between the similarity values obtained using the original word vectors and their emotional counterparts for each word pair within in-category words. For example, in the case of words categorized under *fear*, we calculated the pairwise similarities using both the GloVe vectors and the emotionally enriched GloVe vectors (EEA1_GloVe). Subsequently, we subtracted the similarity value derived from the original vectors from the value obtained using the emotional vectors for each word pair. The resulting average differences for these comparisons, across four emotion categories (anger, fear, sadness, joy), are presented in Table 8.

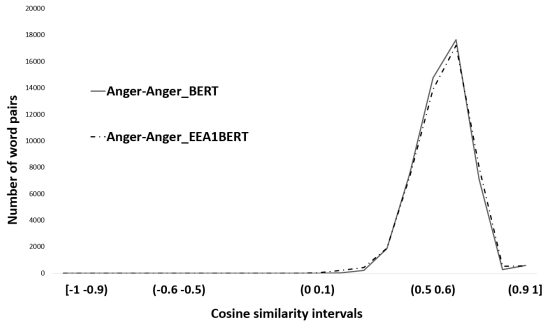
Table 8. The mean variation in similarity scores between word pairs within four emotion categories when comparing the original word embeddings with their emotionally enhanced counterparts.

	Average of Differences (# of word pairs)	
	> 0	**< 0**
EEA1_Word2Vec - Word2Vec	226,125	208,407
EEA2_Word2Vec - Word2Vec	92,619	14,852
EEA3_Word2Vec - Word2Vec	90,288	17,202
EEA1_GloVe - GloVe	206,659	202,941
EEA2_GloVe - GloVe	73,245	22,312
EEA3_GloVe - GloVe	74,280	26,944

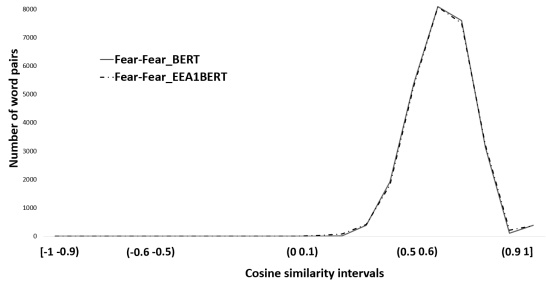
The results indicate that there are more word pairs with positive difference values than those with negative values in both Word2Vec and GloVe vectors. This suggests that, when employing emotional vectors, the increase in similarity outweighs the decrease in similarity across word pairs belonging to all emotion categories. Although only the results of GloVe and EEA1_Glove are presented here, graphs illustrating the cosine similarity values generated with the remaining GloVe and Word2Vec embedding models and their enriched counterparts can be found in Appendices A to E.

Following a similar approach, we calculated in-category similarity using original and emotionally enriched BERT vectors, and the resulting histograms are displayed in Figures 7, 8, and 9 for EEA1, EEA2, and EEA3, respectively. Upon examining these figures, it becomes evident that emotionally enhanced BERT vectors exhibit a noticeable improvement with a distinct rightward shift in the curves that represent the

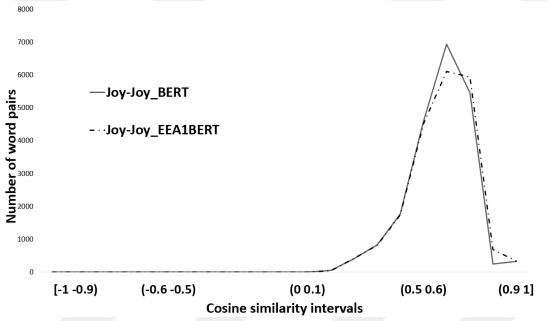
values derived from emotionally enriched BERT vectors. For instance, in Figure 9, it can be seen that both the number of word pairs having higher cosine similarity values are increased and the standard deviation is also decreased when usings EEA3_BERT embeddings.



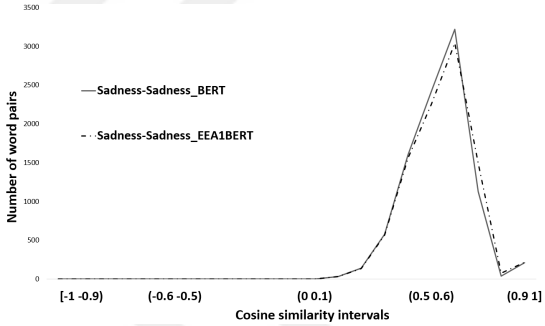
(a) Anger



(b) Fear

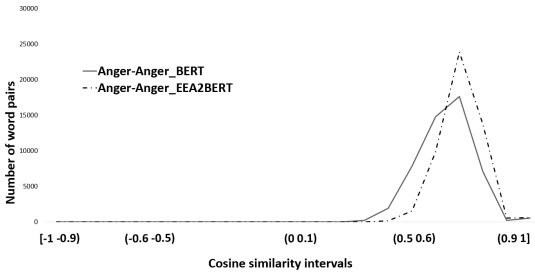


(c) Joy

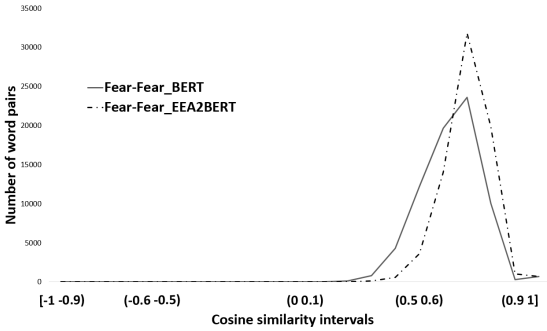


(d) Sadness

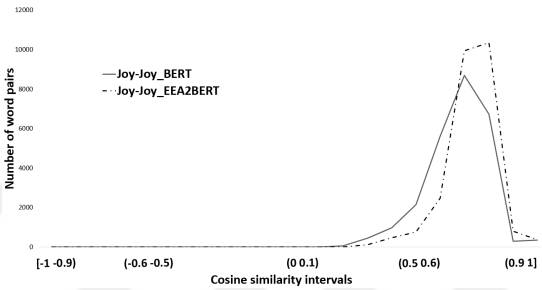
Figure 7. Pairwise similarity histograms (BERT - EEA1_BERT vectors)



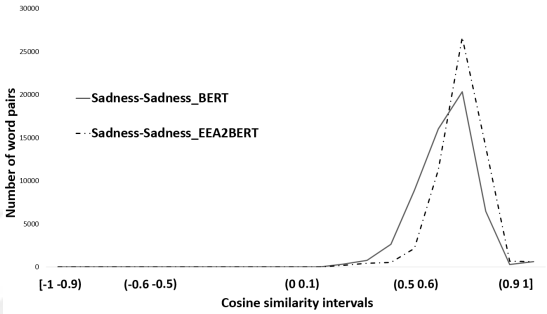
(a) Anger



(b) Fear

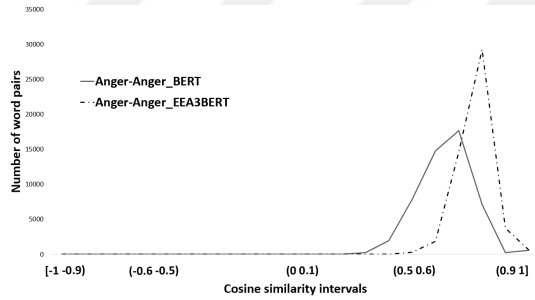


(c) Joy

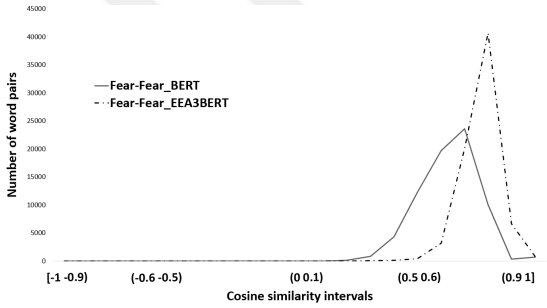


(d) Sadness

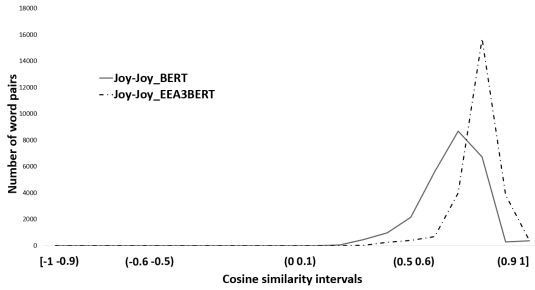
Figure 8. Pairwise similarity histograms (BERT - EEA2_BERT vectors).



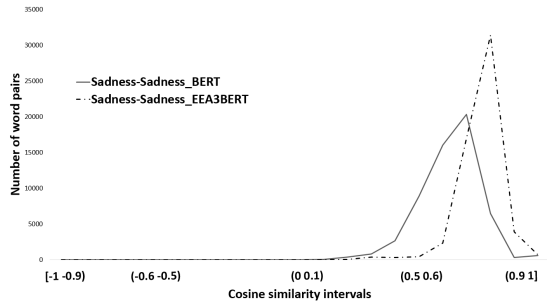
(a) Anger



(b) Fear



(c) Joy



(d) Sadness

Figure 9. Pairwise similarity histograms (BERT - EEA3_BERT vectors).

Additionally, the similarity metric is employed to assess the emotion enrichment

approach from a different angle, specifically through opposite-category similarity comparisons. Table 9 presents a heatmap showing the average in-category and opposite-category similarities. The heat map visualizes values ranging from 0 to 1. Smaller values are represented in shades of yellow, moderate values in shades of orange, and higher values in shades of green. The color intensity within each category may vary to convey subtle differences. In-category similarity scores are computed by averaging the cosine distance scores between word pairs within the same emotional category, employing 12 different embedding models as in previous experiments.

Analyzing the average in-category values in Table 9, it becomes evident that the average similarity scores among words within the same emotional category increase when compared to the original embeddings and their enriched versions. Generally, for emotional categories, EEA2 and EEA3 tend to yield higher scores than EEA1, possibly owing to their incorporation of additional emotion intensity information. In the calculation of average similarity in opposite emotion categories, two pairs of opposite emotional categories are selected based on Plutchik's (Plutchik, 1980) classification. The first pair is *joy-sadness*, involving the computation of the distance between words associated with joy and sadness for each emotion word. A similar procedure is applied to the second pair of emotions, "anger-sadness". While a reduction in average similarity scores was expected when employing emotionally enriched vectors, the results were not consistent with this expectation. This phenomenon could be attributed to the complex nature of language, particularly in the Turkish language. In Turkish, emotions like *anger* and *fear* may not be entirely opposite in meaning. For example, in the following Turkish word pairs labeled as *anger* and *fear* in the dataset, such as "psikoz" and "şizofreni" (English: "psychosis" and "schizophrenia"), "hastalık" and "enfeksiyon" (English: "illness" and "infection"), "sefalet" and "yoksulluk" (English: "poverty" and "misery"), "kargaşa" and "kaos" (English: "disturbance" and "chaos"), the words in each pair can often be used interchangeably in everyday language.

Following the experiments, paired t-tests were conducted on the cosine similarity values within in-category and opposite-category comparisons both before and after the enrichment process. The results revealed that in nearly all instances of enrichment approaches (except for EEA1_BERT in the case of the *fear* emotion), the p-values

for in-category similarity were below 0.001. This implies that the alterations in in-category cosine similarity values are statistically significant following emotional vector enrichment, with the single exception yielding a p-value of 0.37. Conversely, when assessing the opposite-category scores, the t-tests demonstrated that the emotional enrichment significantly modifies BERT vectors, with p-values consistently below 0.001.

Table 9. Average of in-category and opposite-category similarity scores.

	Average in-category similarity scores									Average similarity in opposite emotion categories	
	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Average	Anger vs. Fear	Sadness vs. Joy
Word2Vec	0.13456	0.08603	0.15486	0.12793	0.10448	0.12922	0.11189	0.08331	0.11654	0.12726	0.08907
EEA1_Word2Vec	0.13479	0.08607	0.15447	0.12822	0.10451	0.12909	0.112	0.08332	0.11656	0.12752	0.08901
EEA2_Word2Vec	0.14945	0.0979	0.17188	0.14063	0.11991	0.14391	0.13456	0.09476	0.13163	0.13659	0.09673
EEA3_Word2Vec	0.16454	0.1075	0.18925	0.15446	0.13145	0.15866	0.14408	0.10363	0.1442	0.15104	0.10886
GloVe	0.06401	0.03581	0.06749	0.053	0.05859	0.05791	0.04854	0.04145	0.05335	0.05435	0.02368
EEA1_GloVe	0.06404	0.03581	0.06751	0.05304	0.05867	0.05794	0.04856	0.04147	0.05338	0.05439	0.02369
EEA2_Glove	0.07293	0.04362	0.07670	0.06058	0.06999	0.0661	0.06338	0.04896	0.06278	0.05741	0.02546
EEA3_GloVe	0.07830	0.04732	0.08214	0.06524	0.07497	0.07126	0.06739	0.05313	0.06747	0.0623	0.02863
BERT	0.59632	-	-	0.57059	0.62375	0.58394	-	-	0.59365	0.57148	0.58216
EEA1_BERT	0.59781	-	-	0.57098	0.63157	0.59231	-	-	0.59817	0.57524	0.59075
EEA2_BERT	0.65657	-	-	0.65083	0.67896	0.64551	-	-	0.65797	0.64549	0.65757
EEA3_BERT	0.72201	-	-	0.72271	0.7392	0.71306	-	-	0.72425	0.7154	0.72132

5.2.2.2 Classification

Classification is a machine learning process in which a model is trained to assign input data to specific output categories or labels based on the information extracted from the input features. The primary goal of classification is to create a model capable of predicting the class or category of new data instances that it has not encountered previously. The output variable can involve only two distinct categories, known as binary classification, or multiple categories, referred to as multi-class classification.

Classification finds application in various domains, including speech recognition, image recognition, in the field of natural language processing, and in a range of real-world scenarios (e.g., William et al., 2022; Chen et al., 2022; Sarker, 2021; Afan et al., 2021; Bertolini et al., 2021; Banan et al., 2020).

In the context of emotion detection, when approached as a multi-class classification problem, classifiers can be utilized to assign a given text unit to one of the emotional categories, using either the original vectors or emotion-enhanced vectors as input. In our *word-level* experiments, we applied linear logistic regression (LLR), the sequential minimal optimization algorithm (SMO), and the multi-layer perceptron (MLP) on our Turkish dataset. These experiments were conducted using the Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009) tool with default settings, and all classification experiments were carried out with 10-fold cross-validation. We compared the performance of 12 different embeddings using F1 and accuracy metrics, as detailed in Table 10.

Table 10. Accuracy and F1-scores for a weighted average across four emotions (anger, fear, joy, sadness) are presented, with the top accuracy scores highlighted in bold and the highest F1-scores underlined for each model.

		Classification Models			
	Metric	**LLR**	**SMO**	**MLP**	**AVG**
GloVe	F1	0.5140 (100%)	0.4817 (100%)	0.4870 (100%)	0.4942 (100%)
	Accuracy	0.5165 (100%)	0.4850 (100%)	0.4872 (100%)	0.4962 (100%)
EEA1_GloVe	F1	0.5400 (+5.05%)	0.4810 (-0.15%)	0.4770 (-2.05%)	0.4993 (+1.03%)
	Accuracy	0.5421 (+4.97%)	0.4780 (-1.44%)	0.4780 (-1.88%)	0.4994 (+0.64%)
EEA2_Glove	F1	0.5130 (-0.20%)	0.4940 (+2.55%)	0.4830 (-0.82%)	0.4967 (+0.51%)
	Accuracy	0.5147 (-0.36%)	0.4908 (+1.21%)	0.4835 (-0.76%)	0.4963 (+0.03%)
EEA3_GloVe	F1	0.5300 (+3.11%)	0.4960 (+2.97%)	0.5000 (+2.67%)	0.5087 (+2.93%)
	Accuracy	0.5303 (+2.68%)	0.4936 (+1.77%)	0.5009 (+2.82%)	0.5083 (+2.43%)
Word2Vec	F1	0.4490 (100%)	0.4700 (100%)	0.4610 (100%)	0.4600 (100%)
	Accuracy	0.4628 (100%)	0.4823 (100%)	0.4592 (100%)	0.4681 (100%)
EEA1_Word2Vec	F1	0.4940 (+10.00%)	0.4850 (+3.19%)	0.4730 (+2.60%)	0.4840 (+5.22%)
	Accuracy	0.4965 (+7.28%)	0.4823	0.4734 (+3.09%)	0.4840 (+3.41%)
EEA2_Word2Vec	F1	0.4630 (+3.11%)	0.4800 (+2.13%)	0.4790 (+3.90%)	0.4740 (+3.04%)
	Accuracy	0.4734 (+2.3%)	0.4929 (+2.21%)	0.4787 (+4.25%)	0.4817 (+2.91%)
EEA3_Word2Vec	F1	0.4730 (+5.35%)	0.4800 (+2.13%)	0.4680 (+1.52%)	0.4737 (+2.98%)
	Accuracy	0.4734 (+2.3%)	0.4965 (+2.95%)	0.4663 (+1.55%)	0.4787 (+2.28%)
BERT	F1	0.6410 (100%)	0.6370 (100%)	0.6420 (100%)	0.6397 (100%)
	Accuracy	0.6409 (100%)	0.6358 (100%)	0.6414 (100%)	0.6394 (100%)
EEA1_BERT	F1	0.5330 (-16.85%)	0.5460 (-14.29%)	0.5260 (-18.07%)	0.5350 (-16.37%)
	Accuracy	0.5437 (-15.16%)	0.5493 (-13.61%)	0.5294 (-17.47%)	0.5408 (-15.42%)
EEA2_BERT	F1	0.6500 (+1.41%)	0.6380 (+0.15%)	0.6430 (+0.16%)	0.6437 (+0.63%)
	Accuracy	0.6498 (+1.41%)	**0.6369** (+0.18%)	**0.6425** (+0.17%)	0.6431 (+0.59%)
EEA3_BERT	F1	0.6510 (+1.56%)	0.6340 (-0.47%)	0.6430 (+0.16%)	0.6427 (+0.47%)
	Accuracy	**0.6504** (+1.49%)	0.6341 (-0.27%)	**0.6425** (+0.17%)	0.6423 (+0.47%)

Table 10 illustrates that the most significant accuracy and F1-scores across all classification methods were achieved when utilizing emotionally enriched representations of BERT vectors (EEA2_BERT and EEA3_BERT). The table also presents weighted average accuracy and F1-score results, considering the baseline performance of the original embeddings. For instance, in comparison to GloVe embeddings (set at 100% as the baseline), employing EEA1_GloVe vectors enhances the classification performance by +5.05 when using LLR. On average, except for EEA1_BERT vectors, there is an overall improvement in performance when utilizing emotionally enriched representations. EEA2_BERT produces the highest accuracy result when considering the average accuracy results of three classifiers.

To specifically assess the impact of emotion enrichment on opposite emotions, we focused on a binary classification task with only two categories. Tables 11 and 12 present the F1-scores for the *sadness* and *joy* categories, respectively. In the *sadness* category, the highest F1-score was achieved using EEA2_BERT vectors (0.6067), while in the *joy* category, EEA3_BERT vectors outperformed other embedding methods with the best F1-score of 0.7687.

Table 11. F1-scores of *sadness* emotion.

	LLR	**SMO**	**MLP**	**Average**
GloVe	0.5310 (100%)	0.5110 (100%)	0.4980 (100%)	0.5133 (100%)
EEA1_GloVe	0.5750 (+8.29%))	0.5060 (-0.98%)	0.4880 (-2.01%)	0.5230 (+1.89%)
EEA2_Glove	0.5270 (-0.76%)	0.5280 (+3.33%)	0.4960 (-0.41%)	0.5170 (+0.72%)
EEA3_GloVe	0.5310	0.5130 (+0.4%)	0.5150 (+3.42%)	0.5197 (+1.24%)
Word2Vec	0.4720 (100%)	0.4970 (100%)	0.4650 (100%)	0.4780 (100%)
EEA1_Word2Vec	0.4990 (+5.73%)	0.4860 (-2.22%)	0.4780 (+2.8%)	0.4877 (+2.03%)
EEA2_Word2Vec	0.4930 (+4.45%)	0.5130 (+3.22%)	0.4570 (-1.73%)	0.4877 (+2.03%)
EEA3_Word2Vec	0.4960 (+5.09%)	0.5130 (+3.22%)	0.4700 (+1.08%)	0.4930 (+3.14%)
BERT	0.6040 (100%)	0.5940 (100%)	0.6090 (100%)	0.6023 (100%)
EEA1_BERT	0.2700 (-55.3%)	0.3250 (-45.29)	0.3070 (-49.59%)	0.3007 (-50.09%)
EEA2_BERT	0.6180 (+2.32%)	0.5900 (-0.68%)	0.6120 (+0.5)	**0.6067 (+0.72%)**
EEA3_BERT	0.5970 (-1.16%)	0.5760 (-3.04%)	0.5990 (-1.65%)	0.5907 (-1.94%)

Table 12. F1-scores of *joy* emotion.

	LLR	**SMO**	**MLP**	**Average**
GloVe	0.5620 (100%)	0.5210 (100%)	0.5420 (100%)	0.5417 (100%)
EEA1_GloVe	0.5830 (+3.74%)	0.5210	0.5610 (+3.51%)	0.5550 (+2.47%)
EEA2_Glove	0.5770 (+2.67%)	0.5380 (+3.27%)	0.5360 (-1.11%)	0.5503 (+1.6%)
EEA3_GloVe	0.6430 (+14.42%)	0.5830 (+11.91%)	0.5980 (+10.34%)	0.6080 (+12.25)
Word2Vec	0.5260 (100%)	0.5350 (100%)	0.4810 (100%)	0.5140 (100%)
EEA1_Word2Vec	0.5170 (-1.72%)	0.5360 (+0.19%)	0.5030 (+4.58%)	0.5187 (+0.91%)
EEA2_Word2Vec	0.5250 (-0.2%)	0.5280 (-1.31%)	0.5100 (+6.03%)	0.5210 (+1.37%)
EEA3_Word2Vec	0.5180 (-1.53%)	0.5430 (+1.5%)	0.5000 (+3.96%)	0.5203 (+1.24%)
BERT	0.7540 (100%)	0.7380 (100%)	0.7480 (100%)	0.7467 (100%)
EEA1_BERT	0.7280 (-3.45%)	0.7350 (-0.41%)	0.7340 (-1.88%)	0.7323 (-1.92%)
EEA2_BERT	0.7670 (+1.73%)	0.7510 (+1.77%)	0.7510 (+0.41%)	0.7563 (+1.3%)
EEA3_BERT	0.7720 (+2.39%)	0.7720 (+4.61%)	0.7620 (+1.88%)	**0.7687** (+2.95%)

To summarize, when reviewing the *word-level* in-category cosine similarity and classification experiments and their findings:

1. We assessed the performance of the original word embeddings based on the experimental results presented in Tables 8 - 12. For instance, in Table 9, it's seen that, across all emotions, BERT, Word2Vec, and GloVe yield in-category similarity scores in descending order of magnitude. The average in-category scores (Table 9) for all emotion categories reveal a 0.5403 increase when using BERT compared to GloVe. According to Table 10, the highest accuracy score, 0.6394, is achieved with BERT vectors. Following BERT, GloVe and Word2Vec produced accuracy results of 0.4962 and 0.4681, respectively. When considering only two opposite emotions, *sadness* and *joy*, once again, BERT outperforms GloVe and Word2Vec (as seen in Tables 11 and 12). In summary, in both sets of experiments, BERT vectors consistently outperform semantic embeddings.

2. Upon reviewing the outcomes presented in Table 9, it is evident that all emotion enrichment methods result in increased in-category similarity scores for each emotion. When using Word2Vec, GloVe, and BERT scores as the baseline, the most significant score improvement is achieved with EEA3, while the least

improvement is observed with the EEA1 method. Table 10 demonstrates that, except EEA1_BERT, all emotion enrichment models outperform the original semantic or contextual embeddings in terms of classification performance. However, when examining the F1-scores provided in Table 11, it is important to note that while EEA3 enhances the performance of GloVe and Word2Vec embeddings, the improvement brought about by emotion enrichment does not show a consistent trend.

3. Table 9 demonstrates that when evaluating the similarity scores using vector representations derived from Word2Vec, GloVe, and their emotionally enriched counterparts, the emotion *disgust* exhibits the highest in-category similarity score among the eight emotions. Comparing similarity scores for four emotions calculated with BERT and its emotionally enriched versions reveals that *joy* obtains the highest similarity score. Likewise, in Tables 11 and 12, the highest average F1-scores across the three classification methods are observed for *joy* (with an F1-score of 0.7687) when contrasted with its opposing emotion, *sadness* (which has an F1-score of 0.6067).

4. In our investigation of the effectiveness of existing emotion enrichment methods in the Turkish language, we derived the following findings. In the research conducted by Seyeditabari et al. (2019), it was reported that in in-category cosine similarity calculations, there was a notable improvement in scores. When comparing the original versions, emotionally enriched Word2Vec exhibited a 13% score increase, while emotionally enriched GloVe showed a substantial 29% improvement. Moreover, the study achieved the highest average similarity score using an enriched version of the ConceptNet Numberbatch model (Speer and Chin, 2016), with a score of 0.57 compared to the original score of 0.47, representing a 22% enhancement.

   In our study, we attained the best average in-category similarity score of 0.72425 with emotionally enriched BERT vectors (EEA3_BERT), whereas the use of original BERT vectors yielded a score of 0.59365, marking a 22% improvement.

   In the work of Mao et al. (2019), the classification performances of their hybrid

representations were compared with the skip-gram model using SVM, LLR, decision tree, and gradient boost classifiers. For instance, when applying SVM, the F1-score increased from 0.6542 to 0.7099, and when using the LLR classifier, the F1-score improved from 0.6674 to 0.6969.

As previously mentioned, Table 10 in our study provides a detailed overview of the enhancements brought about by three applied enrichment methods over Word2Vec, GloVe, and BERT when utilizing three classification models. For example, when applying LLR, F1-scores increased from 0.5140 to 0.5400 for GloVe, from 0.4490 to 0.4730 for Word2Vec, and from 0.6410 to 0.6510 for BERT and their emotionally enriched versions, respectively. Just as observed in the aforementioned comparative studies, our study also demonstrates an increase in performance when employing EEA1, EEA2, and EEA3 on the Turkish dataset.

5.2.3. Sentence-Level Emotion Enrichment Experiments

To assess the impact of *sentence-level emotion enrichment* in Turkish and English, we initially acquired sentence vectors for 2000 sentences evenly distributed among four emotion categories: *anger, fear, sadness* and *joy*. These sentences, as described in sub-section 5.1.2, were gathered from various data sources. BERT and DBERT models are employed in sentence-level experiments due to two primary reasons. Firstly, they offer a pre-trained model, allowing the acquisition of sentence-level embeddings that take contextual information into account, all without the need for any pre-processing of individual words. Secondly, BERT has been shown to surpass other models in previous emotion detection studies (e.g., Adoma et al., 2020; Tanana et al., 2021; Savini and Caragea, 2022). Including our word-level experiments, in this study, our main focus was on the Turkish language. However, since *sentence-level emotion enrichment*, to the best of our knowledge, had not been previously studied to this extent, we also conducted the same experiments in English, one of the most extensively researched languages, to make comparisons between languages. The details about cosine similarity and classification experiments are presented in 5.2.3.1 and 5.2.3.2,

respectively.

5.2.3.1 Sentence Level Cosine Similarity Measurements

Table 13 presents the in-category similarity scores for sentences enriched with emotional content. To establish a baseline, original BERT/DBERT sentence vectors, without enrichment, were employed for measuring in-category similarity scores in both languages. The top three scores are highlighted in bold for each language. For instance, in the anger category, the DBERT vectors enriched with EEA3, EEA1, and EEA2 methods consistently outperform other configurations in both languages. Notably, the highest F1 scores show an improvement of 24.2% and 54.5% compared to the baseline scores (original BERT scores) for Turkish and English datasets, respectively. In Table 13, the emotion category with the highest similarity score for each configuration is indicated with an underline. The last two columns provide average similarity and improvement values (expressed as percentages) across all emotions.

Table 13. In-category similarity scores - Enriching sentences with emotional sentences.

Language	Vector types		EEA	In-category similarity, % improvement									
	Sentence	Emotion sentences		Anger		Fear		Joy		Sadness		Average	
Turkish	BERT	-	-	0.752	-	0.747	-	0.758	-	0.747	-	0.751	-
	BERT	BERT	EEA1	0.780	3.7%	0.773	3.5%	0.784	3.4%	0.772	3.3%	0.777	3.5%
	BERT	BERT	EEA2	0.771	2.5%	0.764	2.3%	0.776	2.4%	0.763	2.1%	0.769	2.4%
	BERT	BERT	EEA3	0.796	5.9%	0.787	5.4%	0.803	5.9%	0.786	5.2%	0.793	5.6%
	DBERT	-	-	0.915	21.7%	0.910	21.8%	0.922	21.6%	0.910	21.8%	0.914	21.7%
	DBERT	DBERT	EEA1	**0.930**	23.7%	**0.926**	24.0%	**0.935**	23.4%	**0.926**	24.0%	**0.929**	23.7%
	DBERT	DBERT	EEA2	**0.923**	22.7%	**0.919**	23.0%	**0.930**	22.7%	**0.919**	23.0%	**0.923**	22.9%
	DBERT	DBERT	EEA3	**0.934**	24.2%	**0.929**	24.4%	**0.941**	24.1%	**0.929**	24.4%	**0.933**	24.2%
English	BERT	-	-	0.610	-	0.593	-	0.623	-	0.597	-	0.606	-
	BERT	BERT	EEA1	0.624	2.3%	0.607	2.4%	0.637	2.2%	0.612	2.5%	0.620	2.3%
	BERT	BERT	EEA2	0.631	3.4%	0.613	3.4%	0.645	3.5%	0.617	3.4%	0.627	3.5%
	BERT	BERT	EEA3	0.655	7.4%	0.636	7.3%	0.672	7.9%	0.640	7.2%	0.651	7.4%
	DBERT	-	-	0.918	50.5%	0.916	54.5%	0.920	47.7%	0.917	53.6%	0.918	51.5%
	DBERT	DBERT	EEA1	**0.932**	52.8%	**0.931**	57.0%	**0.934**	49.9%	**0.931**	55.9%	**0.932**	53.8%
	DBERT	DBERT	EEA2	**0.927**	52.0%	**0.925**	56.0%	**0.929**	49.1%	**0.925**	54.9%	**0.927**	53.0%
	DBERT	DBERT	EEA3	**0.936**	53.4%	**0.934**	57.5%	**0.938**	50.6%	**0.934**	56.4%	**0.936**	54.5%

When we consider the in-category similarity scores in Table 13, it can be seen that,

1. The addition of emotional content consistently led to significantly higher scores in both the Turkish and English datasets, whether we examine individual emotions or average values. Furthermore, the improvement in similarity scores, when compared to the baseline scores, is notably more pronounced in the

English dataset. This suggests that enriching sentences with emotion has a more pronounced positive effect in English.

2. Comparing BERT and DBERT vectors, both the original and enriched DBERT vectors consistently yield higher similarity scores.

3. For the Turkish and English datasets, the enrichment process excels particularly in the joy category, and this difference is statistically significant (according to a paired t-test with a p-value of less than 0.02) compared to other emotion categories.

As the EEA3 method consistently outperforms other methods for all emotions in both languages, we employ this method to enrich sentences with emotion lexicon words in the subsequent comparative experiments. Table 14 provides in-category similarity results for these experiments, where we not only utilize original BERT/DBERT word vectors but also emotion-enriched versions of lexicon word vectors (e.g., BERT+EEA1, BERT+EEA2). Similar to Table 13, the top 3 scores are highlighted in bold for each emotion, and the emotion values that benefit most from enrichment are underlined in Table 14.

The experimental findings from enriching sentences with emotion-enriched lexicon words indicate the following:

1. Across all emotion categories, configurations utilizing DBERT vectors outperform those using BERT vectors.

2. When examining the top three scores, it's notable that two of them are associated with emotion-enriched lexicon words (specifically, DBERT+EEA3 and DBERT+EEA2), while one score is derived from original words represented by DBERT in both Turkish and English datasets. Consequently, it can be concluded that when incorporating lexicon words in the enrichment process, their original vectors can also be utilized.

3. The most significant improvement is observed in the *joy* category during the enrichment experiments with lexicon words, mirroring the results seen in

Table 14. In-category similarity scores - Enriching sentences with original and emotion-enriched lexicon words.

Language	Vector type		EEA Method	In-category similarity (% improvement)									
	Sentence	Emotion lexicon words		Anger		Fear		Joy		Sadness		Average	
Turkish	BERT	-	-	0.752	-	0.747	-	0.758	-	0.747	-	0.751	-
	BERT	BERT	EEA3	0.910	21.0%	0.922	23.4%	0.936	23.5%	0.916	22.6%	0.921	22.6%
	BERT	BERT+ EEA1	EEA3	0.756	0.5%	0.751	0.5%	0.761	0.4%	0.751	0.5%	0.755	0.5%
	BERT	BERT+ EEA2	EEA3	0.916	21.8%	0.926	24.0%	0.939	23.9%	0.921	23.3%	0.925	23.2%
	BERT	BERT+ EEA3	EEA3	0.922	22.6%	0.931	24.6%	0.943	24.4%	0.927	24.1%	0.931	24.0%
	DBERT	-	-	0.915	-	0.916	-	0.922	-	0.910	-	0.914	-
	DBERT	DBERT	EEA3	**0.971**	6.1%	**0.971**	6.0%	**0.976**	5.9%	**0.970**	6.6%	**0.972**	6.3%
	DBERT	DBERT+ EEA1	EEA3	0.915	0.0%	0.911	-0.5%	0.922	0.0%	0.911	0.1%	0.915	0.1%
	DBERT	DBERT+ EEA2	EEA3	**0.972**	6.2%	**0.972**	6.1%	**0.978**	6.1%	**0.972**	6.8%	**0.974**	6.6%
	DBERT	DBERT+ EEA3	EEA3	**0.973**	6.3%	**0.974**	6.3%	**0.979**	6.2%	**0.973**	6.9%	**0.975**	6.7%
English	BERT	-	-	0.610	-	0.593	-	0.623	-	0.597	-	0.606	-
	BERT	BERT	EEA3	0.837	37.2%	0.831	40.1%	0.872	40.0%	0.835	39.9%	0.844	39.3%
	BERT	BERT+ EEA1	EEA3	0.615	0.8%	0.598	0.8%	0.629	1.0%	0.602	0.8%	0.611	0.8%
	BERT	BERT+ EEA2	EEA3	0.840	37.7%	0.834	40.6%	0.876	40.6%	0.840	40.7%	0.848	39.9%
	BERT	BERT+ EEA3	EEA3	0.844	38.4%	0.838	41.3%	0.879	41.1%	0.845	41.5%	0.852	40.6%
	DBERT	-	-	0.918	-	0.916	-	0.920	-	0.917	-	0.918	-
	DBERT	DBERT	EEA3	**0.971**	5.8%	**0.969**	5.8%	**0.973**	5.8%	0.971	5.9%	**0.971**	5.8%
	DBERT	DBERT+ EEA1	EEA3	0.918	0.0%	0.917	0.1%	0.920	0.0%	0.917	0.0%	0.918	0.0%
	DBERT	DBERT+ EEA2	EEA3	**0.971**	5.8%	**0.969**	5.8%	**0.973**	5.8%	0.971	5.9%	**0.971**	5.8%
	DBERT	DBERT+ EEA3	EEA3	**0.970**	5.7%	**0.968**	5.7%	**0.971**	5.5%	0.969	5.7%	**0.970**	5.7%

sentence enrichment experiments. These results were confirmed using a paired t-test, showing a p-value of less than 0.02.

In summary, the experiments assessing similarity within the same emotion category have shown that enriching sentences with emotion-enriched lexicon words yields improved results in both languages, surpassing the baseline scores. In 9 (*word-level emotion enrichment experiments*), it's reported that when EEA3 enriches BERT vectors of lexicon words, the highest similarity score within the same emotion category reaches 0.72425 on the Turkish dataset. Comparing this score with the results of enriching Turkish sentences with either emotion sentences or emotion lexicon words, it becomes evident that enriching at the sentence level rather than the word level is more effective for the Turkish language.

In the study of Seyeditabari et al. (2019), where words are enriched with EEA1, in-category cosine similarity calculations on the English NRC (National Research Council) dataset resulted in the best average similarity score of 0.57, compared to

its original counterpart of 0.47, signifying a 22% improvement. In our research, the corresponding (highest) score increased from 0.606 to 40.6% and reached 0.852 with BERT, and from 0.918 to 0.971 with a 5.8% increase using DBERT vectors when considering the average scores. When comparing these results, even though our baseline scores are higher due to the utilization of BERT/DBERT vectors, it's evident that emotion enrichment at the sentence level holds greater potential in enhancing the emotion detection process compared to word-level enrichment.

Conversely, when analyzing the top average scores in Table 13 and Table 14, it becomes evident that, for both languages, methods of emotion enrichment that incorporate lexicon word vectors achieve superior results compared to those utilizing emotion sentence vectors. Consequently, it can be inferred that the settings involving sentence enrichment with emotion-enriched and/or original lexicon words are more effective compared to alternative approaches.

5.2.3.2 Classification

In this experiment, we aim to categorize each text sample into one of four primary emotions: anger, fear, joy, and sadness. Our experiments involve using various classification models, including logistic regression (LLR), sequential minimal optimization (SMO), multilayer perceptron (MLP), convolutional neural network (CNN), and deep neural network softmax dense layer (DNN-SM).

For the LLR, SMO, and MLP classifications, we employed the default parameters of the WEKA tool (Hall et al., 2009). Meanwhile, for CNN and DNN-SM, we followed the model architectures and parameters detailed in the study of Shaaban et al. (2021). The classification with CNN and DNN-SM is implemented using Python 3.9 and the Keras library (Chollet et al., 2015) in Google Colaboratory. To ensure robust results and prevent overfitting, all classification experiments were conducted using 5-fold cross-validation. We assessed the performance by computing the F1 metric based on the average results from the five folds.

Tables 15 and 16 display the mean F1 scores from classification experiments conducted with a 5-fold cross-validation method. Additionally, they present the percentage improvement in comparison to the baseline results obtained with original

vectors for each configuration. In these tables, cells that are shaded highlight the three highest scores within each respective language category.

The "Average_CL" column in both tables contains the average F1 scores of the classifiers, and each cell in the "Average" row holds the average F1 score for that specific classifier. For instance, when sentences are represented using original BERT vectors (without any enrichment), the average F1 value in the Turkish dataset is 0.673, irrespective of the type of classifier used. In the same dataset, regardless of the embedding and enrichment methods applied, the LLR classifier exhibits an average performance of 0.613. The highest values for "Average_CL" (found in the rightmost two columns of Tables 3 and 4) are indicated by underlining.

Table 15. Classification F1 scores - Enriching sentences with emotion sentences.

Language	Vector type		EEA	Classification F1 score (% improvement)											
	Sentence	Emotion sentences		LLR		SMO		MLP		CNN		DNN-SM		Average_CL	
Turkish	BERT	-	-	0.628	-	0.597	-	0.613	-	0.751	-	**0.774**	-	0.673	-
	BERT	BERT	EEA1	0.618	-1.6%	0.599	0.3%	0.619	1.0%	0.747	-0.5%	0.750	-3.1%	0.667	-0.9%
	BERT	BERT	EEA2	0.632	0.6%	0.596	-0.2%	0.614	0.2%	0.748	-0.4%	**0.778**	0.5%	**0.674**	0.1%
	BERT	BERT	EEA3	0.626	-0.3%	0.594	-0.5%	0.619	1.0%	0.747	-0.5%	0.772	-0.3%	0.672	-0.1%
	DBERT	-	-	0.597	-	0.580	-	0.591	-	0.706	-	0.758	-	0.646	-
	DBERT	DBERT	EEA1	0.592	-0.8%	0.580	0.0%	0.589	-0.3%	0.679	-3.8%	0.724	-4.5%	0.633	-2.1%
	DBERT	DBERT	EEA2	0.608	1.8%	0.574	-1.0%	0.586	-0.8%	0.726	2.8%	0.756	-0.3%	0.650	0.6%
	DBERT	DBERT	EEA3	0.601	0.7%	0.585	-0.9%	0.596	0.8%	0.733	3.8%	**0.780**	2.9%	0.659	1.9%
		Average		0.613		0.588		0.603		0.711		0.755		0.630	
English	BERT	-	-	0.663	-	0.615	-	0.646	-	0.752	-	**0.788**	-	0.693	-
	BERT	BERT	EEA1	0.658	-0.8%	0.617	0.3%	0.656	1.5%	0.757	0.7%	0.770	-2.3%	0.692	-0.2%
	BERT	BERT	EEA2	0.663	0.0%	0.616	0.2%	0.652	0.9%	0.757	0.7%	**0.803**	1.9%	**0.698**	0.8%
	BERT	BERT	EEA3	0.658	-0.8%	0.615	0.0%	0.643	-0.5%	0.748	-0.5%	0.787	-0.1%	0.690	-0.4%
	DBERT	-	-	0.616	-	0.589	-	0.617	-	0.740	-	0.778	-	0.668	-
	DBERT	DBERT	EEA1	0.617	0.2%	0.587	-0.3%	0.612	-0.8%	0.727	-1.8%	0.748	-3.9%	0.658	-1.5%
	DBERT	DBERT	EEA2	0.618	0.3%	0.596	1.2%	0.615	-0.3%	0.750	1.4%	**0.797**	2.4%	0.675	1.1%
	DBERT	DBERT	EEA3	0.620	0.6%	0.591	0.3%	0.622	0.8%	0.751	1.5%	0.776	-0.3%	0.672	0.6%
		Average		0.639		0.603		0.633		0.742		0.775		0.655	

The results presented in Table 15 and Table 16 indicate the following findings:

1. The classifier DNN_SM consistently achieved the highest F1 scores, both in terms of the highest and average classification scores across experiments for both languages. Notably, the maximum F1 scores of DNN-SM outperformed other classifiers for both languages. When considering the average F1 scores of individual classifiers (as shown in the "Average" row), DNN_SM emerged as the top-performing classifier, closely followed by CNN.

2. A paired samples t-test was conducted to compare the performance of DNN_SM and CNN. The results revealed a significant difference in F1 scores between DNN_SM and CNN for experiments on the Turkish dataset ($p<0.02$). Similarly,

Table 16. Classification F1 scores - Enriching sentences with original and emotion-enriched lexicon words.

Language	Vector type		EEA	Classification F1 score (% improvement)											
	Sentence	Emotion lexicon words		LLR		SMO		MLP		CNN		DNN_SM		Average_CL	
Turkish	BERT	-	-	0.628	-	0.597	-	0.613	-	0.750	-	**0.786**		0.675	
	BERT	BERT	EEA3	0.582	-7.3%	0.587	-1.7%	0.585	-4.6%	0.669	-10.8%	0.674	-14.2%	0.619	-8.2%
	BERT	BERT+ EEA1	EEA3	0.629	0.2%	0.597	0.0%	0.616	0.5%	0.751	0.1%	**0.791**	0.6%	**0.677**	0.3%
	BERT	BERT+ EEA2	EEA3	0.578	-8.0%	0.580	-2.8%	0.576	-6.0%	0.680	-9.3%	0.666	-15.3%	0.616	-8.7%
	BERT	BERT+ EEA3	EEA3	0.574	-8.6%	0.595	-0.3%	0.581	-5.2%	0.654	-12.8%	0.654	-16.8%	0.612	-9.4%
	DBERT	-	-	0.597	-	0.580	-	0.591	-	0.706	-	0.758	-	0.589	-
	DBERT	DBERT	EEA3	0.580	-7.6%	0.571	-1.6%	0.566	-4.2%	0.700	-0.8%	0.698	-7.9%	0.572	-2.9%
	DBERT	DBERT+ EEA1	EEA3	0.593	-5.6%	0.580	0.0%	0.589	-0.3%	0.712	0.8%	**0.770**	1.6%	0.587	-0.3%
	DBERT	DBERT+ EEA2	EEA3	0.576	-8.3%	0.560	-3.4%	0.568	-3.9%	0.663	-6.1%	0.707	-6.7%	0.568	-3.6%
	DBERT	DBERT+ EEA3	EEA3	0.587	-6.5%	0.573	-1.2%	0.578	-2.2%	0.684	-3.1%	0.698	-7.9%	0.579	-1.7%
		Average		0.592		0.582		0.586		0.697		0.720		0.609	
English	BERT	-	-	0.663	-	0.615	-	0.646	-	0.752		**0.788**		0.693	
	BERT	BERT	EEA3	0.605	-8.7%	0.591	-3.9%	0.581	-10.1%	0.677	-10.0%	0.637	-19.2%	0.618	-10.8%
	BERT	BERT+ EEA1	EEA3	0.665	0.3%	0.616	0.2%	0.646	0.0%	0.759	0.9%	**0.798**	1.3%	**0.697**	0.6%
	BERT	BERT+ EEA2	EEA3	0.603	-9.0%	0.590	-4.1%	0.591	-8.5%	0.700	-6.9%	0.672	-14.7%	0.631	-8.9%
	BERT	BERT+ EEA3	EEA3	0.597	-10.0%	0.583	-5.2%	0.586	-9.3%	0.684	-9.0%	0.677	-14.1%	0.625	-9.7%
	DBERT	-	-	0.616	-	0.589	-	0.617	-	0.740	-	0.778	-	0.607	-
	DBERT	DBERT	EEA3	0.598	-2.9%	0.585	-0.7%	0.611	-1.0%	0.700	-5.4%	0.757	-2.7%	0.598	-1.5%
	DBERT	DBERT+ EEA1	EEA3	0.612	-0.6%	0.590	0.2%	0.613	-0.6%	0.741	0.1%	**0.806**	3.6%	0.605	-0.4%
	DBERT	DBERT+ EEA2	EEA3	0.592	-3.9%	0.585	-0.7%	0.602	-2.4%	0.712	-3.8%	0.746	-4.1%	0.593	-2.4%
	DBERT	DBERT+ EEA3	EEA3	0.598	-2.9%	0.573	-2.7%	0.598	-3.1%	0.714	-3.5%	0.736	-5.4%	0.590	-2.9%
		Average		0.615		0.592		0.609		0.718		0.740		0.626	

significant differences in F1 scores were observed for emotion sentence experiments on the English dataset, with a p-value<0.001.

3. When examining the "Average_CL" values in Table 15, it is seen that configurations utilizing the EEA2 enrichment model led to improvements in F1 scores. In Table 16, a similar pattern is observed, where an enhancement in "Average_CL" scores is achieved when enriching the data with emotion lexicon words using the EEA1 model. Nevertheless, it's important to note that no specific embedding or enrichment model dominantly outperforms others across all configurations.

4. Overall, the classification results demonstrate that certain enriched configurations do result in improved classification performance. However, when considering the average scores presented in the last two columns of Table 15 and Table 16, it becomes apparent that there is potential for improvement of up to 1.9% when compared to the base scores.

To assess the significance of improvements in the configurations that yielded the highest F1 scores in Tables 15 and 16, a 5-fold cross-validated paired t-test was employed to compare the performance of the top two configurations for both languages. Table 17 provides the results regarding these configurations, denoted as CF1 (the best-performing) and CF2 (the second best-performing).

For instance, in the Turkish dataset, the highest F1 score (0.791) was achieved by the DNN-SM classifier (CL). In this configuration, sentences were represented using original BERT vectors whereas emotion lexicon words were represented EEA1 BERT vectors. Sentence-level enrichment was accomplished through EEA3. On the other hand, the second-best configuration (CF2) for the Turkish dataset achieved an F1 score of 0.786 with the DNN_SM classifier, and it used original BERT sentences without any enrichment. The p-values for both the Turkish and English datasets fall below the 0.05 threshold, indicating a significant distinction in performance between the top and second-ranking configurations.

Table 17. The outcomes of the 5-fold cross-validated paired t-test comparing the top two performing configurations in experiments involving emotion sentences and emotion lexicon words.

	CF1						CF2						
Language	**Sentence**	**Emotion sentences**	**Emotion lexicon words**	**EEA**	**CL**	**F1**	**Sentence**	**Emotion sentences**	**Emotion lexicon words**	**method**	**CL**	**F1**	**p-value**
Turkish	BERT	-	BERT+ EEA1	EEA3	DNN_SM	0.791	BERT	-	-	-	DNN_SM	0.786	0.045
English	DBERT	-	DBERT+ EEA1	EEA3	DNN_SM	0.806	BERT	BERT	-	EEA2	DNN_SM	0.803	0.046

5.2.3.3 Comparison of Sentence-level and Word-level Classification

To assess the relative performance of sentence-level and word-level classification experiments, we pursued three distinct strategies to produce alternative word-level results. At the word level, we employed emotion lexicon sets (specifically, BERT+EEA1 and DBERT+EEA1 for Turkish and English) along with the DNN_SM classifier, which was utilized in the most successful configurations as shown in Table 17, to ensure comparability of results. Additionally, we applied EEA3 enrichment like the configurations in sentence-level experiments. Below, we provide a concise overview of the word-level approaches (A1, A2, A3):

A1 We generated BERT/DBERT embeddings for sentences that included at least one emotion word by averaging the embeddings of the emotion words from the enriched lexicon found within the sentence. As we directly used the EEA1 enriched lexicon words, no additional enrichment was applied. A drawback of this method is that it cannot create embeddings for sentences without lexicon words, leading to a decrease in the dataset size to 1718 and 1694 sentences for Turkish and English, respectively, from the initial 2000 sentences.

A2 In this method, we initially produced BERT/DBERT vectors for every token present in the sentence. These individual token vectors were then averaged to form sentence embeddings, and no further enhancements were applied.

A3 The final approach introduced an additional enhancement step compared to the second method. In this case, we first created BERT/DBERT vectors for the individual words in the sentence and enriched them with EEA3, and then we calculated the sentence vector by averaging these vectors.

Table 18 presents the average F1 scores at the word level from classification experiments conducted using a 5-fold cross-validation approach. The highest F1 score is 0.725 achieved with the A3 approach for the Turkish dataset and 0.693 with the A1 approach for the English dataset. When comparing these top F1 scores with the CF1 values in Table 17, it becomes evident that sentence-level methods outperform word-level approaches.

Table 18. F1 scores for classification at the word-Level experiments

Language	Approach	Composing words	Emotion Lexicon Words	CL	method	F1
Turkish	**A1**	BERT	BERT + EEA1	DNN_SM	-	0.633
	A2	BERT	-	DNN_SM	-	0.586
	A3	BERT	BERT + EEA1	DNN_SM	EEA3	**0.725**
English	**A1**	DBERT	DBERT + EEA1	DNN_SM	-	**0.693**
	A2	DBERT	-	DNN_SM	-	0.614
	A3	DBERT	DBERT + EEA1	DNN_SM	EEA3	0.650

CHAPTER 6: OPTIMIZING EMOTION ENRICHMENT: DIMENSIONALITY REDUCTION AND LEXICON FILTERING

This chapter aims to address the research question below:

RQ5 - In the context of representing emotionally enriched sentence vectors, how can we improve precision and effectiveness by optimizing the computational efficiency of vectors and refining emotion lexicons, while taking into account linguistic nuances in both Turkish and English languages?

In our previous experiments in Chapter 5, we conducted an assessment of sentence-level emotion enrichment using both emotion-lexicon words and emotion-lexicon sentences. We found that enriching sentences with emotional lexicon words yielded more promising results. Then, we focused even more intensively on this specific approach, in this context, we conducted two groups of studies to optimize sentence enrichment approaches. These studies are summarized below:

1. **Lexicon Filtering:** First, we discussed the potential issues that could arise within emotion lexicons. These issues encompassed a range of challenges, including the accuracy of word categorization within emotion categories, the inclusion of words with ambiguous or multiple emotional associations, and the potential presence of meaningless or less relevant terms. For instance, in the NRC lexicon, the word "grim" (tr. "acımasız") is associated with *anger, anticipation, disgust, fear* and *sadness* emotion categories at the same time. Indeed, it is essential to consider both the real-world context and the human factor when dealing with emotion lexicons. Emotions are intricately tied to various subjective, cultural, and ethnic nuances. When we consider this diversity, it is necessary to acknowledge the influence of human subjectivity and cultural factors in the annotation and interpretation of emotion lexicons.

   At this point, we emphasized the importance of capturing a connection between the general contexts in which words are used and their usage within sentences

from emotion-rich datasets. The idea here is that if a word's meaning in emotion-rich texts significantly diverges from its usage in general contexts, it may imply that the word is not being used accurately in emotional sentences. In this context, we conducted experiments to filter emotion lexicon words and work on a more accurate emotion lexicon.

2. **Dimensionality Reduction:** We recognized that using 768-dimensional BERT vectors can be computationally demanding, especially when dealing with large datasets. Thus, we wondered if some of the individual dimensions within these vectors might contain hidden emotional information. This led us to explore whether we could find emotional cues by dissecting and analyzing these dimensions. By addressing these concerns, our goal was to make our computations more efficient and potentially discover new insights into how emotions are hidden in language representations. To sum up, we utilize a sliding window technique to segment BERT vectors and analyze consistent patterns, aiming to improve computational efficiency and gain insights into how emotions are integrated within these vectors, ultimately providing a novel perspective on emotion representation.

The primary reason for our focus on these two solutions, reducing BERT vectors and lexicon filtering, is to enable more precise and effective results in representing *emotionally enriched sentence vectors*. Dimensionality reduction of BERT vectors helps reduce computational costs, allowing for faster and more efficient analyses, while lexicon filtering assists us in accurately identifying and providing emotional content. Proposed solutions are independent of lexicon and word embedding models. In this study, we experimented with the NRC emotion lexicon and BERT word embedding models. Just like in the same sentence embedding enrichment study, in this group study, we conducted experiments for both the Turkish and English languages.

6.1. Lexicon Filtering

In previous works in the field, the process of emotion enrichment is often carried out using emotion lexicon words. However, there might be certain issues about the

quality of the emotion lexicon that could impact the accurate detection or classification of emotions in texts. The quality can vary due to factors such as words belonging to each emotion category, the potential for incorrect categorization of words, or the inclusion of meaningless terms.

In this study, we proposed a method for creating a more refined and accurately categorized emotion lexicon by subjecting words belonging to emotion categories through a filter. The procedure can be explained as follows: We generated two BERT vectors for emotion words in the NRC lexicon and TT-NRC (Turkish-translated NRC). As BERT is a contextual embedding method, the context of words is important while obtaining embeddings. When obtaining vector *(VE)* for each word in the emotion lexicon, we used randomly selected sentences containing those words from a collection of datasets having sentences labeled by the same emotion categories. The dataset is the collection of TEC, TEI, and TREMO datasets. This procedure was conducted as described in Section 5. Secondly, we created vectors *(VN)* for each emotion word without a labeled dataset. For English, we retrieved sentences from the OpenWebTextCorpus dataset (Gokaslan and Cohen, 2019) where the words could have been used more frequently in their general contexts. OpenWebTextCorpus is a dataset containing randomly shuffled Reddit posts. We chose this dataset to explore the usage and meaning of emotion words in the lexicon. For Turkish, we utilized the Turkish news dataset (Özbay, 2019) containing Turkish news texts.

We used *(VN)* BERT vectors obtained from OpenWebTextCorpus and the Turkish news dataset as general meaning and usage vectors. Therefore, we obtained two distinct vectors for each emotion lexicon word from two different datasets. The vector generation procedures are the same for *(VE)* and *(VN)*, with the only difference being the datasets. Next, we compared the word representations from *VE* with *VN* using the cosine similarity metric for each word. The general procedure of regarding procedure is given in Figure 10.

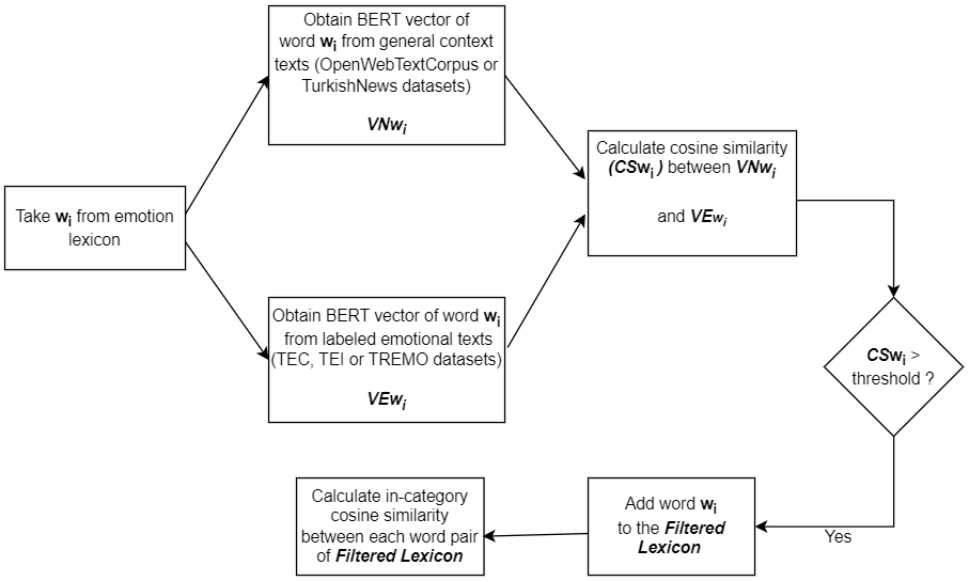


Figure 10. The framework of lexicon filtering procedures.

For instance, suppose the emotion lexicon $W = w_1, w_2, w_3, ..., w_n$ where n is the lexicon size. In this case, the procedure involves creating BERT vectors for each of these lexicon words, resulting in pairs of vectors for comparison. For instance, we would calculate cosine similarity between VE_{w_1} and VN_{w_1}, VE_{w_2} and VN_{w_2}, VE_{w_3} and VN_{w_3}, and so on, for all n emotion lexicon words. This comprehensive analysis allows us to understand how the contextual information from labeled datasets and the general context from OpenWebTextCorpus impact the vector representations of individual emotion lexicon words.

After calculating the cosine similarities for each word from 2 data sources, if the resulting cosine similarity values fell below a certain threshold, we considered removing those words from further analysis. To establish the appropriate threshold, we conducted multiple iterations of the same experiment.

Tables 19 and 20 present the before and after results of in-category cosine similarity averages following the filtering process for English and Turkish languages, respectively. Initially, under the "unfiltered" category, we provide the average cosine similarity values for word pairs within each emotion category, along with the count of words in each category without any filtering. Subsequently, we applied different thresholds to eliminate words such that, a word w_i will be filtered if the similarity between VE_{w_i} and VN_{w_i} fell below a certain value. After such removals, we

recalculated the in-category cosine similarity averages for all word pairs using vector space *VE* and reported the updated results. Accordingly, we repeated the same experiment with different thresholds to find the most suitable value for lexicon filtering. The increases in cosine similarity values and the reduction in the number of words are summarized in Tables 19 and 20.

Table 19. Changes in the number of words in four emotion categories and the amounts of increase in average cosine similarity (CS) values within each category based on different threshold values for English Lexicon Words.

		Anger	Fear	Sadness	Joy
Unfiltered	**CS**	0.448	0.447	0.451	0.512
	% Increase in CS	-	-	-	-
	# of words	526	650	526	740
	% Decrease in # of words	-	-	-	-
0.2	**CS**	0.466	0.469	0.469	0.527
	% Increase in CS	3.936%	4.971%	4.138%	2.880%
	# of words	462	570	464	650
	% Decrease in # of words	-12.167%	-12.308%	-11.787%	-12.162%
0.3	**CS**	0.484	0.485	0.485	0.537
	% Increase in CS	7.842%	8.391%	7.629%	4.748%
	# of words	376	493	395	576
	% Decrease in # of words	-28.517%	-24.154%	-24.905%	-22.162%
0.4	**CS**	0.500	0.500	0.499	0.560
	% Increase in CS	11.620%	11.951%	10.964%	9.490%
	# of words	290	402	333	481
	% Decrease in # of words	-44.867%	-38.154%	-36.692%	-35.000%
0.5	**CS**	0.530	0.539	0.5390	0.587
	% Increase in CS	18.313%	20.514%	19.636%	14.671%
	# of words	197	283	240	376
	% Decrease in # of words	-62.548%	-56.462%	-54.373%	-49.189%
0.6	**CS**	0.573	0.580	0.559	0.607
	% Increase in CS	27.886%	29.783%	24.002%	18.648%
	# of words	120	172	145	253
	% Decrease in # of words	-77.186%	-73.538%	-72.433%	-65.811%
0.7	**CS**	0.618	0.623	0.597	0.620
	% Increase in CS	37.895%	39.347%	32.598%	21.072%
	# of words	41	57	48	78
	% Decrease in # of words	-92.205%	-91.231%	-90.875%	-89.459%

Table 20. Changes in the number of words in four emotion categories and the amounts of increase in average cosine similarity (CS) values within each category based on different threshold values for Turkish Lexicon Words.

		Anger	**Fear**	**Sadness**	**Joy**
Unfiltered	**CS**	0.589	0.578	0.577	0.619
	% Increase in CS	-	-	-	-
	# of words	450	536	475	321
	% Decrease in # of words	-	-	-	-
0.1	**CS**	0.618	0.618	0.604	0.628
	% Increase in CS	5.029%	6.816%	4.763%	1.562%
	# of words	297	376	328	256
	% Decrease in # of words	-34%	-29.851%	-30.947%	-20.249%

The experiments with various thresholds could be repeated in English, but this was not possible in Turkish. Even at a threshold value of 0.1, there was a 30% decrease in the total lexicon words belonging to 4 emotional categories. Therefore, without conducting further experiments, we accepted the threshold of 0.1.

Table 21 presents the outcomes of the filtering process of the English lexicon based on varying cosine similarity thresholds. In the table, "Threshold" represents the cosine similarity threshold values experimented in the filtering, "Data Set Size and % Decrease in data set size" shows the number of words retained in the lexicon after applying each threshold and the percentage decrease relative to the original dataset size of 2442 words, "Weighted Average CS" reveals the weighted average in category cosine similarity values of all emotion categories post-filtering and "% Increase in CS" illustrates the percentage growth in cosine similarity compared to the unfiltered dataset. As the threshold increases, the dataset size consistently decreases while the weighted average cosine similarity experiences a steady rise. Considering the increase in cosine similarity scores and the decrease in the number of lexicon words, we selected 0.5 as the filtering threshold.

Table 21. Change in English data set size and average in-category cosine similarity scores for all emotion categories.

Threshold	Data set size	% Decrease in data set size	Weighted Average CS	% increase in CS
0.00	2442	-	0.47	-
0.20	2146	12.12%	0.49	3.89%
0.30	1840	24.65%	0.50	7.02%
0.40	1506	38.33%	0.52	11.07%
0.50	1096	55.12%	0.55	18.42%
0.60	690	71.74%	0.58	24.95%
0.70	224	90.83%	0.62	31.59%

6.2. Dimensionality Reduction

While BERT has proven highly effective in capturing contextual information, its utilization of 768-dimensional vectors can be computationally intensive for some tasks especially when the dataset size is very large. Besides, the parts of the 768-length vector may capture some information about some specific aspects of the language or the property of the text unit that it represented. In this study we present an alternative approach that focuses on identifying patterns within these word/sentence vectors, thereby reducing the complexity of the analysis. Our proposed methodology can be summarized as below:

1. To extract meaningful patterns from BERT vectors, we utilize a sliding window technique. This technique breaks down the vectors into smaller fixed-size segments, allowing us to capture local contextual information.

2. By conducting a detailed analysis of cosine similarity values and investigating patterns of segments that might have emotional information, we aimed to gain insights into the relationship between different dimensions of BERT vectors and the emotional scopes of texts.

To determine the reduced vector size we refer to the study conducted by Su et al. (2021). They utilized a technique called "whitening" mainly for enhancing the isotropy of sentence representations and also reducing the dimensionality of vectors. Their

methodology reduces the BERT embedding size to 256 and 384.

In our methodology, the window size in the sliding window technique is selected as 256, much like the way Su et al. (2021) partitioned the BERT vectors in their study. Consequently, in our study, we divided the vectors, each having a BERT vector size of 768, into segments (windows) with a window size of 256 as can be seen in Figure 11. To cover every dimension of a vector, we established the slide size as 64. Thus, the first window (sub-vector) starts at position 1 and ends at position 256, and the second window starts at position 65 and ends at position 321, and so forth.

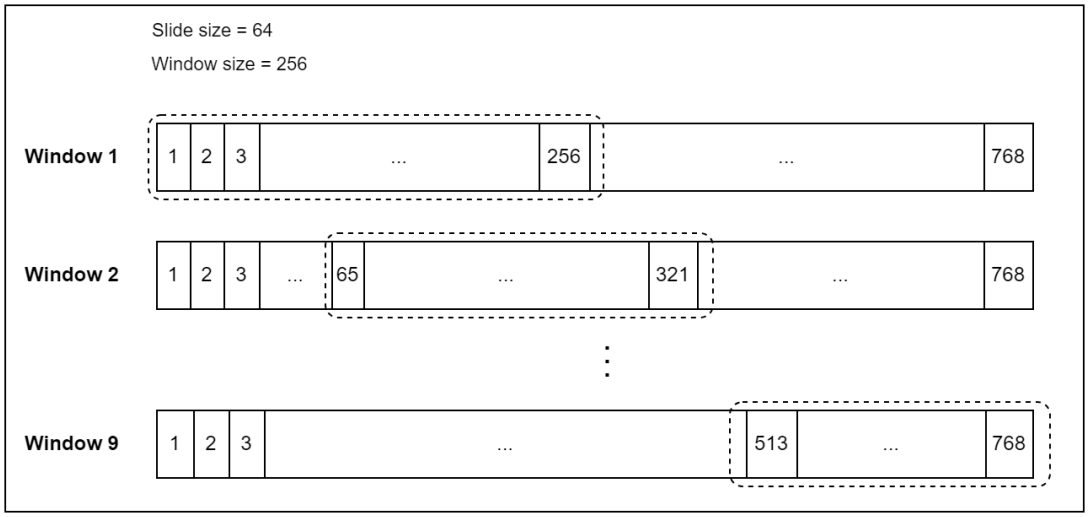


Figure 11. Framework for vector partitioning with sliding window technique.

6.2.1. Experimental Results

Firstly, because we were primarily looking for a pattern carrying emotional information within BERT vectors, we initially investigated the subvectors of lexicon words. When we divided the 768-dimensional word BERT vectors into segments using the sliding window technique, we obtained 9 sub-vectors from each vector. Initially, we used these 9 sub-vectors individually to represent the respective words. As in our previous experiments, we once again conducted pairwise cosine similarity measurements within the same emotions (in-category) and between opposite emotions (opposite category). However, this time, in the representation of a word labeled with an emotion, we operated with only the sub-vector corresponding to a single window of the 768-dimensional BERT vector. For example, when measuring cosine similarity

between two words labeled with *anger*, we first used only the sub-vectors between dimensions 0 and 256 and calculated the cosine similarity. Similarly, we repeated the similarity measurements with the sub-vectors corresponding to other windows. Assuming that words were represented with only 1 sub-vector, we repeated the cosine similarity experiments 9 times and presented the results as a heat map in Table 22 and 23 for English and Turkish words, respectively.

Table 22. Pairwise in-category and opposite-category cosine similarity results of *English words* while using only one window.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Anger-Anger	0.249	0.597	0.628	0.633	0.630	0.361	0.256	0.244	0.233
In-category	**Fear-Fear**	0.220	0.607	0.634	0.640	0.637	0.340	0.236	0.226	0.215
cosine similarity	**Sadness-Sadness**	0.236	0.598	0.629	0.636	0.633	0.357	0.254	0.250	0.242
	Joy-Joy	0.285	0.665	0.687	0.692	0.690	0.403	0.311	0.305	0.283
Opposite-category	**Joy-Sadness**	0.226	0.619	0.647	0.653	0.650	0.354	0.251	0.243	0.228
cosine similarity	**Anger-Fear**	0.223	0.596	0.625	0.631	0.627	0.340	0.233	0.222	0.212

Table 23. Pairwise in-category and opposite-category cosine similarity results of *Turkish words* while using only one window.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Anger-Anger	0.288	0.330	0.300	0.324	0.312	0.767	0.766	0.768	0.775
In-category	**Fear-Fear**	0.276	0.318	0.292	0.321	0.306	0.760	0.760	0.761	0.768
cosine similarity	**Sadness-Sadness**	0.275	0.317	0.295	0.321	0.302	0.760	0.760	0.762	0.770
	Joy-Joy	0.276	0.318	0.316	0.342	0.341	0.797	0.796	0.798	0.805
Opposite-category	**Joy-Sadness**	0.250	0.293	0.284	0.308	0.297	0.771	0.771	0.773	0.781
cosine similarity	**Anger-Fear**	0.276	0.318	0.289	0.316	0.302	0.761	0.761	0.763	0.769

As seen through the heat maps, it is evident that there are dimensions in BERT vectors that carry emotional information. Therefore, when certain sub-vectors are used in cosine similarity measurements, higher similarity is observed compared to other sub-vectors. This situation suggests that not all 768-dimensional vectors need to be used, but rather, sub-vectors can be utilized. More specifically, when we look at

English word vectors, we observed emotional information in windows 2,3,4 and 5, and in Turkish, emotional intensity can also be present in windows 6, 7, 8, and 9.

Later on, after adding filters (threshold is 0.5 for English and 0.1 for Turkish) to our experiments with lexicon words, we continued by repeating the previous experiment. After filtering both English and Turkish lexicon words (as described in 6.1), we once again split the word BERT vectors into 256-dimensional sub-vectors and conducted our experiment by measuring cosine similarity between the sub-vectors. We list the results in Tables 24 and 25 for both English and Turkish.

Table 24. Pairwise in-category and opposite-category cosine similarity results of *Filtered English words* while using only one window.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
In-category cosine similarity	**Anger-Anger**	0.301	0.687	0.714	0.717	0.713	0.427	0.308	0.292	0.281
	Fear-Fear	0.279	0.700	0.723	0.728	0.725	0.417	0.301	0.289	0.274
	Sadness-Sadness	0.289	0.697	0.723	0.727	0.723	0.420	0.305	0.304	0.297
	Joy-Joy	0.342	0.739	0.757	0.760	0.757	0.465	0.366	0.362	0.338
Opposite-category cosine similarity	**Joy-Sadness**	0.268	0.703	0.726	0.730	0.726	0.409	0.294	0.287	0.270
	Anger-Fear	0.276	0.688	0.712	0.716	0.712	0.407	0.287	0.274	0.264

Table 25. Pairwise in-category and opposite-category cosine similarity results of *Filtered Turkish words* while using only one window.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
In-category cosine similarity	**Anger-Anger**	0.303	0.344	0.316	0.346	0.333	0.792	0.791	0.792	0.799
	Fear-Fear	0.299	0.343	0.318	0.352	0.337	0.794	0.794	0.795	0.801
	Sadness-Sadness	0.284	0.324	0.306	0.334	0.315	0.783	0.783	0.785	0.793
	Joy-Joy	0.279	0.321	0.321	0.346	0.344	0.805	0.805	0.806	0.814
Opposite-category cosine similarity	**Joy-Sadness**	0.253	0.297	0.290	0.314	0.305	0.787	0.786	0.789	0.797
	Anger-Fear	0.294	0.336	0.309	0.341	0.327	0.791	0.790	0.791	0.798

Table 26. % increase in cosine similarity per window after filtering English lexicon words.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
In-category cosine similarity	**Anger-Anger**	20.884%	15.075%	13.694%	13.270%	13.175%	18.283%	20.313%	19.672%	20.601%
	Fear-Fear	26.818%	15.321%	14.038%	13.750%	13.815%	22.647%	27.542%	27.876%	27.442%
	Sadness-Sadness	22.458%	16.555%	14.944%	14.308%	14.218%	17.647%	20.079%	21.600%	22.727%
	Joy-Joy	20.000%	11.128%	10.189%	9.827%	9.710%	15.385%	17.685%	18.689%	19.435%
	Average % increase in in-category similarity	**22.540%**	**14.520%**	**13.216%**	**12.789%**	**12.729%**	**18.490%**	**21.405%**	**21.959%**	**22.551%**
Opposite-category cosine similarity	**Joy-Sadness**	18.584%	13.570%	12.210%	11.792%	11.692%	15.537%	17.131%	18.107%	18.421%
	Anger-Fear	23.767%	15.436%	13.920%	13.471%	13.557%	19.706%	23.176%	23.423%	24.528%
	Average % increase in opposite-category similarity	**21.175%**	**14.503%**	**13.065%**	**12.631%**	**12.624%**	**17.621%**	**20.154%**	**20.765%**	**21.475%**

Table 27. % increase in cosine similarity per window after filtering Turkish lexicon words.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
In-category cosine similarity	**Anger-Anger**	5.208%	4.242%	5.333%	6.790%	6.731%	3.259%	3.264%	3.125%	3.097%
	Fear-Fear	8.333%	7.862%	8.904%	9.657%	10.131%	4.474%	4.474%	4.468%	4.297%
	Sadness-Sadness	3.273%	2.208%	3.729%	4.050%	4.305%	3.026%	3.026%	3.018%	2.987%
	Joy-Joy	1.087%	0.943%	1.582%	1.170%	0.880%	1.004%	1.131%	1.003%	1.118%
	Average % increase in in-category similarity	**4.475%**	**3.814%**	**4.887%**	**5.417%**	**5.511**	**2.941**	**2.974%**	**2.903%**	**2.875%**
Opposite-category cosine similarity	**Joy-Sadness**	1.200%	1.365%	2.113%	1.948%	2.694%	2.075%	1.946%	2.070%	2.049%
	Anger-Fear	6.522%	5.660%	6.920%	7.911%	8.278%	3.942%	3.811%	3.670%	3.771%
	Average % increase in opposite-category similarity	**3.861%**	**3.513%**	**4.517%**	**4.930%**	**5.486%**	**3.009%**	**2.878%**	**2.870%**	**2.910%**

In both Tables 26 and 27, the increase percentages were provided when cosine similarity was measured using each subvector after filtering the English and Turkish lexicon words, respectively. When we examine Tables 26 and 27:

1. When a word is represented using only a BERT sub-vector corresponding to a single window in both languages and cosine similarity is measured among filtered lexicon words within the same emotions (in-category cosine similarity), an increase in the similarity score is observed when it is compared to non-filtered lexicon words.

2. The increase has been observed individually for all emotion categories; however, *joy* is the emotion with the least increase in both languages.

3. Despite no decrease in opposite-category similarity values after filtering, there has been a decrease in the magnitude of the increase.

4. When comparing the two languages, the cosine similarity increase rates in Turkish after filtering are lower than those in English.

In our experiments up to this point, we experimented with the effectiveness of the lexicon filtering and sub-vector determination methods on Turkish and English lexicon words. We observed the impact of using only certain portions of the vectors by conducting in-category cosine similarity measurements, as opposed to using all vectors, and later experienced that this effect was further amplified on filtered lexicon words. Subsequently, we adapted these results to our primary problem, which is the sentence emotion enrichment problem. In summary, in our previous experiments, we adapted various word-level emotion enrichment methods to sentence-level emotion enrichment (Sub-section 5.2.3). Finally, we took this sentence enrichment problem one step further with different experimental combinations using sub-vectors and filtered lexicons and then, applied various enrichment settings.

These combinations and the in-category cosine similarity results of sentences belonging to four different emotion categories are given in Tables 28 and 29. The first line of both tables gives the average in-category cosine similarity results of sentences belonging to emotion categories when the sentence embeddings are represented by BERT vectors having a length of 768 without any enrichment. We accepted these values as a baseline and compared the results of subsequent enrichment combinations with it, then provided the increased amounts as percentages in Tables 28 and 29 accordingly.

In all other methods, the sentence enrichment process was performed using the EEA3 method, as it yielded the best results in our previous experiments. Initially, in these tables, we added the configurations that provided the best results in cosine similarity values in our sentence-level enrichment experiments. In both Turkish and English, this method involved sentence vectors enriched with lexicon word BERT vectors represented by the EEA3 method.

Then, as seen in the third and fourth rows of the Tables 28 and 29, we enriched

sentence vectors using filtered lexicon words represented by BERT and EEA3_BERT vectors, respectively.

Finally, instead of the 768-dimensional BERT vectors, we used the sub-vectors we had determined in previous experiments for both English and Turkish languages. We took sub-vectors that were formed by concatenating the sub-vectors that provided the best results in both English and Turkish. For example, English BERT vectors contained more emotive data in sub-vectors 2, 3, 4, and 5. These sub-parts are concatenated and form a vector that starts from the second window's beginning dimension and ends in the fifth window's last dimension. The framework for extracting sub-vectors can be seen in Figure 12 As a result, we enriched sentence sub-vectors first with the sub-vectors of emotion lexicon words (EEA3_BERT). Lastly, we enriched the sentence sub-vectors with the sub-vectors of filtered lexicon words (EEA3_BERT).

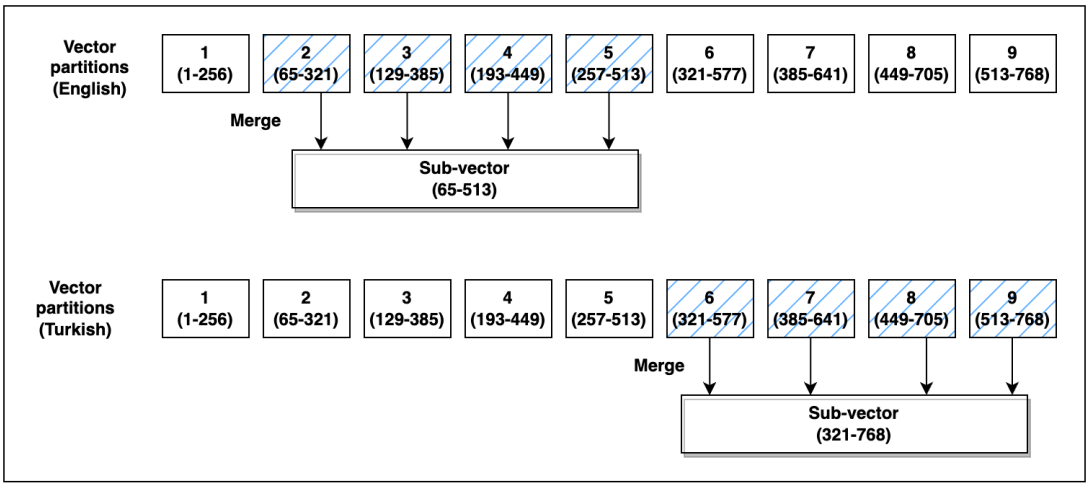


Figure 12. Framework for extracting sub-vectors.

Table 28. English sentence embeddings enrichment with several combinations. (The best results are shown in bold.)

Sentence embedding	Enrichment method	Enrichment by	In-category similarity (% improvement)									
			Anger		Fear		Joy		Sadness		Average	
BERT	**-**	**-**	0.610	-	0.593	-	0.623	-	0.597	-	0.606	-
BERT	**EEA3**	**Emotion Lexicon Words (BERT + EEA3)**	0.844	38.36%	0.838	41.32%	0.879	41.09%	0.845	41.54%	0.852	40.57%
BERT	**EEA3**	**Filtered Emotion Lexicon Words (BERT)**	0.697	14.25%	0.669	12.79%	0.628	0.74%	0.634	6.18%	0.657	8.43%
BERT	**EEA3**	**Filtered Emotion Lexicon Words (BERT + EEA3)**	0.863	41.50%	0.862	45.34%	0.899	44.24%	0.866	45.06%	0.872	44.02%
BERT Subvector	**EEA3**	**Emotion Lexicon Words Subvector (BERT + EEA3)**	0.885	45.09%	0.880	48.44%	0.905	45.28%	0.883	47.88%	0.888	46.65%
BERT Subvector	**EEA3**	**Filtered Emotion Lexicon Words Subvector (BERT + EEA3)**	**0.900**	**47.55%**	**0.899**	**51.68%**	**0.922**	**47.99%**	**0.900**	**50.83%**	**0.905**	**49.48%**

Table 29. Turkish sentence embeddings enrichment with several combinations. (The best results are shown in bold.)

Sentence embedding	Enrichment method	Enrichment by	In-category similarity (% improvement)									
			Anger		Fear		Joy		Sadness		Average	
BERT	**-**	**-**	0.752	-	0.747	-	0.758	-	0.747	-	0.751	-
BERT	**EEA3**	**Emotion Lexicon Words (BERT + EEA3)**	0.922	22.61%	0.931	24.63%	0.943	24.41%	0.927	24.10%	0.931	23.93%
BERT	**EEA3**	**Filtered Emotion Lexicon Words (BERT)**	0.916	21.76%	0.926	23.99%	0.939	23.93%	0.922	23.39%	0.926	23.27%
BERT	**EEA3**	**Filtered Emotion Lexicon Words (BERT + EEA3)**	0.927	23.31%	0.937	25.37%	0.948	25.01%	0.932	24.83%	0.936	24.63%
BERT Subvector	**EEA3**	**Emotion Lexicon Words Subvector (BERT + EEA3)**	0.953	26.67%	0.959	28.45%	0.966	27.45%	0.956	28.03%	0.959	27.65%
BERT Subvector	**EEA3**	**Filtered Emotion Lexicon Words Subvector (BERT + EEA3)**	**0.957**	**27.24%**	**0.962**	**28.84%**	**0.968**	**27.68%**	**0.960**	**28.50%**	**0.962**	**28.06%**

Examining the results in Tables 28 and 29, in all configurations performed, an increase in all emotion categories is observed based on the cosine similarity values compared with the base BERT vectors. The best sentence-level enrichment results were achieved when;

1. Representing sentences through sub-vectors,

2. Using filtered lexicon words for enriching sentence vectors,

3. Representing lexicon word vectors with EEA3_BERT sub-vectors.

CHAPTER 7: CONCLUSION

This thesis explores the impact of incorporating emotional content into a highly agglutinative and less-resourced language, Turkish. The study employs three different approaches to emotion enrichment, applied to both contextual and semantic embeddings. Additionally, the experimentation involved two different text units, namely words and sentences. Evaluation is conducted through cosine similarity and classification. The original embeddings are enriched by incorporating additional information, such as emotion intensity, into the models. Notably, this investigation represents the first comprehensive exploration of emotion enrichment in Turkish texts. The results suggest that integrating emotion enrichment has the potential to enhance word and sentence embeddings in the Turkish dataset, thereby contributing to the advancement of text-based emotion detection studies.

Furthermore, our research focused on improving the enrichment of emotions at the sentence level by tackling computational challenges associated with vectors and fine-tuning emotion lexicons. Our goal was to boost computational efficiency and precision in capturing emotional sentences by diminishing the dimensionality of vectors and refining lexicon words. The results of our experiments on BERT vectors demonstrated notable enhancements in cosine similarity values across all emotion categories when representing sentences through sub-vectors, utilizing filtered lexicon words, and incorporating enriched (EEA3_BERT) sub-vectors. These outcomes underscore the potential of our refined methodologies in advancing the representation of emotions in any language, providing valuable insights for future investigations in emotion detection.

In this thesis, which we specified in the introduction chapter, there were 5 research questions that we aimed to find the answers to. These are given below with the resulting answers/discussions.

RQ1 - What is the most efficient original word/sentence embedding method for enhancing the detection of emotions in Turkish texts, thereby improving the

performance of emotion detection studies?

In word-level experiments, GloVe and Word2Vec as semantic, and BERT contextual embedding models are utilized on the Turkish dataset. On the other hand, BERT and DistilBERT contextual embeddings are utilized in sentence-level experiments. In these experiments, we observe that transformer-based contextual embedding models achieve much better results compared to semantic models.

RQ2 - Can enhanced representations of words and sentences outperform their original counterparts?

When examining the results of the word-level experiments, we observe an improvement in cosine similarity experiments when emotion enrichment is applied. Looking at the classification results, we notice that the EEA1 method does not yield improvement for all the original embedding models, especially when the BERT word embedding model is used. However, when EEA2 and EEA3 methods are employed, an increase in classification performance is observed. We believe that the reason for this could be the additional emotion intensity information used by these two methods. At the sentence-level, we observed a consistent increase in in-category similarity in cosine similarity experiments when emotion enrichment methods were applied using two different enrichment setups. In classification experiments, a specific embedding or enrichment model consistently outperforms others in all configurations could not be observed.

RQ3- Is the efficacy of original and enhanced representations subject to variation based on emotion categories?

In word-level cosine similarity experiments, the highest results are obtained with *disgust* emotion category when experiments are conducted in 8 emotion categories. Considering 4 emotion categories, the average similarity scores of *joy* category surpass the other categories. In sentence level, the joy category demonstrates the highest performance in both configurations considering enrichment through lexicon words and emotion sentences.

RQ4 - Which emotion enrichment methods give better results on word-level and sentence-level emotion detection?

Examining the word-level experiments EEA2 and EEA3 generally produced

better results in two groups of experiments. In the first part of the sentence-level experiments (when sentences are enriched though emotion sentences), the EEA3 method outperforms continuously for all emotions in both Turkish and English languages and that is why we continued the second part of the experiments (enriching the sentences with emotion lexicon words) with using EEA3 as enrichment method.

RQ5 - In the context of representing emotionally enriched sentence vectors, how can we improve precision and effectiveness by optimizing the computational efficiency of vectors and refining emotion lexicons, while taking into account linguistic nuances in both Turkish and English languages?

To address the common characteristic of all vectorization methods, which is their high dimensionality and consequently computational demand, we investigated whether vectors representing any text unit carry emotional information in specific parts and whether it is possible to reduce vector dimensions. According to our results, we achieved higher in-category similarity in specific parts of vectors in both Turkish and English and continued our experiments using only those sub-parts in representing sentences. Additionally, addressing issues in emotion lexicons, we filtered lexicon words based on their usage in general contexts and emotion-containing contexts. According to our results, our proposed approach of both sub-vector usage and filtered lexicon usage yielded the best results in our comparative experiments. Figure 13 presents the framework for the best-performing configuration of optimized sentence-level enrichment. Also, Table 30 presents the in-category cosine similarity scores using the best-performing configurations for both languages in comparison with representing the sentences with only BERT embeddings.

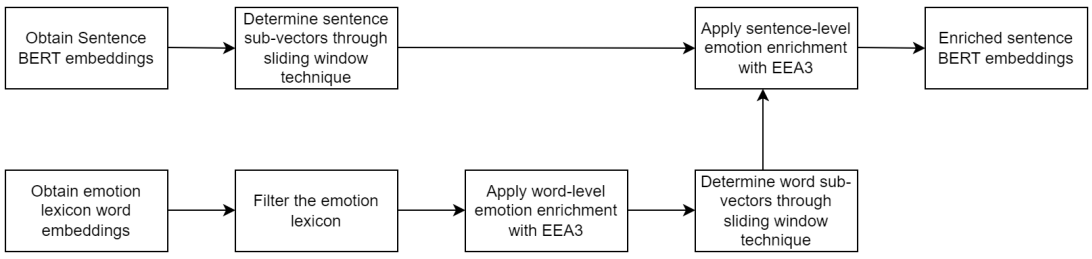


Figure 13. The framework of the best-performing sentence level emotion enrichment with optimization.

Table 30. Turkish and English sentence embeddings enrichment of best-performing combinations.

Language	Sentence embedding	Enrichment method	Enrichment by	In-category similarity (% improvement)									
				Anger		Fear		Joy		Sadness		Average	
Turkish	BERT	-	-	0,752	-	0,747	-	0,758	-	0,747	-	0,751	-
	BERT Subvector	EEA3	Filtered Emotion Lexicon Words Subvector (BERT + EEA3)	**0,957**	**27,24%**	**0,962**	**28,84%**	**0,968**	**27,68%**	**0,960**	**28,50%**	**0,962**	**28,06%**
English	BERT	-	-	0.610	-	0.593	-	0.623	-	0.597	-	0.606	-
	BERT Subvector	EEA3	Filtered Emotion Lexicon Words Subvector (BERT + EEA3)	**0.900**	**47.55%**	**0.899**	**51.68%**	**0.922**	**47.99%**	**0.900**	**50.83%**	**0.905**	**49.48%**

In summary, this thesis provides contributions to the field that can be outlined as follows;

- The study classifies and condenses commonly used data resources in text-based emotion detection, contrasts lexicons created or applied in research centered on sentiment and emotion detection through lexicon-based approaches and provides an overview of techniques suggested in the literature to enhance vector representation by incorporating emotional and sentiment information. These aspects are extensively explored in Aka Uymaz and Kumova Metin (2022).

- Three techniques for enriching emotionally the vector representations were employed on a Turkish dataset, and their effectiveness was assessed and compared. To our knowledge, the utilization of emotion-enriched vectors in Turkish was introduced for the first time in Aka Uymaz and Kumova Metin (2023a) and Aka Uymaz and Kumova Metin (2023b)

- While the literature has focused on enriching emotions at the word level, this study introduces the concept of sentence-level emotion enrichment. The approach involves applying various parameters to sentence vectors in both Turkish and English.

- The computational demands associated with 768-dimensional BERT vectors are addressed and subtle emotional cues within specific dimensions are explored. Our approach utilizes the sliding window technique to enhance computational

efficiency, providing novel perspectives on emotion representation by leveraging sub-vectors.

- To address potential challenges associated with emotion lexicons, a proposed approach suggests refining and more precisely categorizing the lexicon. The method involves filtering emotion lexicon words according to contextual distinctions, to enhance the emotion lexicons.

When examining our overall studies and findings, we believe that the use of labeled datasets or lexicons in many emotion enrichment methods can impact the success of the studies. The utilization of datasets labeled or controlled by native speakers of each language, rather than translation datasets, can minimize this effect. Additionally, the efforts we made in this study to distinguish parts of BERT vectors carrying emotional information among their dimensions can be applied to different languages and different embedding models. This approach may also provide a perspective for studies aimed at extracting different information from various vector patterns.

As future work, our studies can focus on aspect-based sentiment analysis, which is a recent concept that involves analyzing specific aspects rather than examining the entire text. This approach aims to capture different emotions or polarities for distinct terms. Comparisons can be made between methods targeting this objective and standard sentiment analysis techniques across different languages.

REFERENCES

Abas, A. R., Elhenawy, I., Zidan, M. and Othman, M. (2022) *BERT-CNN: A deep learning model for detecting emotions from text.*, Computers, Materials & Continua, Vol. 71 (2).

Abdaoui, A., Azé, J., Bringay, S. and Poncelet, P. (2017) *Feel: a French expanded emotion lexicon*, Language Resources and Evaluation, Vol. 51.

Abdel Razek, M. and Frasson, C. (2017) *Text-based intelligent learning emotion system*, Journal of Intelligent Learning Systems and Applications, Vol. 09, pp. 17–20.

Abubakar, A. M., Gupta, D. and Palaniswamy, S. (2022) Explainable emotion recognition from tweets using deep learning and word embedding models, *2022 IEEE 19th India Council International Conference (INDICON)*, pp. 1–6.

Acheampong, F., Wenyu, C. and Nunoo-Mensah, H. (2020) *Text-based emotion detection: Advances, challenges and opportunities*, Engineering Reports, Vol. .

Adoma, A. F., Henry, N.-M., Chen, W. and Rubungo Andre, N. (2020) Recognizing emotions from texts using a bert-based approach, *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 62–66.

Afan, H. A., Osman, A. I. A., Essam, Y., Ahmed, A. N., Huang, Y. F., Kisi, O., Sherif, M., Sefelnasr, A., wing Chau, K. and El-Shafie, A. (2021) *Modeling the fluctuations of groundwater level by employing ensemble deep learning techniques*, Engineering Applications of Computational Fluid Mechanics, Vol. 15 (1), pp. 1420–1439.

Agrawal, A., An, A. and Papagelis, M. (2018) Learning emotion-enriched word representations, *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Association for Computational Linguistics, pp. 950–961. https://www.aclweb.org/anthology/C18-1081

Ahmad, Z., Jindal, R., Ekbal, A. and Bhattachharyya, P. (2020) *Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding*, Expert Systems with Applications, Vol. 139, pp. 112851.

https://www.sciencedirect.com/science/article/pii/S0957417419305536

Aka Uymaz, H. and Kumova Metin, S. (2022) *Vector based sentiment and emotion analysis from text: A survey*, Engineering Applications of Artificial Intelligence, Vol. 113, pp. 104922.

Aka Uymaz, H. and Kumova Metin, S. (2023a) *Emotion-enriched word embeddings for Turkish*, Expert Systems with Applications, Vol. 225, pp. 120011. https://www.sciencedirect.com/science/article/pii/S0957417423005134

Aka Uymaz, H. and Kumova Metin, S. (2023b) *Enriching transformer-based embeddings for emotion identification in an agglutinative language: Turkish*, IT Professional, Vol. 25 (4), pp. 67–73.

Akın, A. A. and Akın, M. D. (2007) Zemberek, an open source nlp framework for turkic languages.

Alshahrani, M. (2020) Exploring embedding vectors for emotion detection. phd thesis, university of essex.

Alshahrani, M., Samothrakis, S. and Fasli, M. (2019) Identifying idealised vectors for emotion detection using cma-es, *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '19, New York, NY, USA, Association for Computing Machinery, p. 157–158. https://doi.org/10.1145/3319619.3322057

An, Y., Sun, S. and Wang, S. (2017) *Naive bayes classifiers for music emotion classification based on lyrics*, 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Vol. pp. 635–638.

Arora, G., Rahimi, A. and Baldwin, T. (2019) Does an lstm forget more than a cnn? an empirical study of catastrophic forgetting in nlp, *ALTA*.

Azizan, A., Jamal, N. N. S. A., Abdullah, M. N., Mohamad, M. and Khairudin, N. (2019) Lexicon-based sentiment analysis for movie review tweets, *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pp. 132–136.

Baali, M. and Ghneim, N. (2019) *Emotion analysis of arabic tweets using deep learning approach*, Journal of Big Data, Vol. 6.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010) Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining., Vol. 10, Proceedings of

the International Conference on Language Resources and Evaluation, LREC 2010, 17-23, Valletta, Malta.

Badaro, G., Jundi, H., Hajj, H. and El-Hajj, W. (2018) EmoWordNet: Automatic expansion of emotion lexicon using English WordNet, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 86–93. https://www.aclweb.org/anthology/S18-2009

Banan, A., Nasiri, A. and Taheri-Garavand, A. (2020) *Deep learning-based appearance features extraction for automated carp species identification*, Aquacultural Engineering, Vol. 89, pp. 102053. https://www.sciencedirect.com/science/article/pii/S0144860919302195

Batra, H., Punn, N., Sonbhadra, S. and Agarwal, S. (2021) Bert based sentiment analysis: A software engineering perspective.

Benchimol, J., Kazinnik, S. and Saadon, Y. (2021) *Federal reserve communication and the covid-19 pandemic*, Covid Economics, Vol. 79, pp. 218–256.

Bertolini, M., Mezzogori, D., Neroni, M. and Zammori, F. (2021) *Machine learning for industrial applications: A comprehensive literature review*, Expert Systems with Applications, Vol. 175, pp. 114820. https://www.sciencedirect.com/science/article/pii/S095741742100261X

Blandin, A., Saïd, F., Villaneau, J. and Marteau, P.-F. (2021) Automatic emotions analysis for french email campaigns optimization., Barcelone, Spain.

Blitzer, J., Dredze, M. and Pereira, F. (2007) Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, Association for Computational Linguistics, pp. 440–447. https://aclanthology.org/P07-1056

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) Enriching word vectors with subword information.

Bollegala, D., Maehara, T. and Kawarabayashi, K.-i. (2015) Unsupervised cross-domain word representation learning, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics, pp. 730–740. https://aclanthology.org/P15-1071

Bollegala, D., Weir, D. and Carroll, J. (2014) Learning to predict distributions of words across domains, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, Association for Computational Linguistics, pp. 613–623. https://aclanthology.org/P14-1058

Bostan, L. A. M. and Klinger, R. (2018) An analysis of annotated corpora for emotion classification in text, *COLING*.

Boynukalın, Z. (2012) *Emotion analysis of Turkish texts by using machine learning methods.*, Master's thesis, Department of Computer Engineering, Middle East Technical University, Ankara,Turkey.

Bradley, M. and Lang, P. (1999) Affective norms for english words (anew): Instruction manual and affective ratings.

Buechel, S. and Hahn, U. (2017) EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain, Association for Computational Linguistics, pp. 578–585. https://www.aclweb.org/anthology/E17-2092

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S. and Narayanan, S. (2008) *Iemocap: Interactive emotional dyadic motion capture database*, Language Resources and Evaluation, Vol. 42, pp. 335–359.

Calvo, M., Álvarez Sánchez, J., Ferrández, J. and Fernandez, E. (2020) *Affective robot story-telling human-robot interaction: Exploratory real-time emotion estimation analysis using facial expressions and physiological signals*, IEEE Access, Vol. PP, pp. 1–1.

Calvo, R. A. and Kim, S. M. (2012) *Emotions in text: Dimensional and categorical models*, Computational Intelligence, Vol. early view.

Cambria, E., Poria, S., Hazarika, D. and Kwok, K. (2018) Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings, *AAAI Conference on Artificial Intelligence*.

https://api.semanticscholar.org/CorpusID:13690470

Canales, L. and Martinez-Barco, P. (2014) Emotion detection from text: A survey, Conference: Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC), pp. 37–43.

Chang, Y.-C., Yeh, W.-C., Hsing, Y.-C. and Wang, C.-A. (2019) *Refined distributed emotion vector representation for social media sentiment analysis*, PLOS ONE, Vol. 14.

Chaumartin, F.-R. (2007) *Upar7: A knowledge-based system for headline sentiment tagging*, Proceedings of SemEval-2007, Vol. .

Chen, J. (2008) The construction and application of chinese emotion word onthology.master's thesis, dalian university of technology, china.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Huang, K. and Ku, L.-W. (2018) Emotionlines: An emotion corpus of multi-party conversations, Miyazaki, Japan, European Language Resources Association (ELRA).

Chen, W., Sharifrazi, D., Liang, G., Band, S. S., Chau, K. W. and Mosavi, A. (2022) *Accurate discharge coefficient prediction of streamlined weirs by coupling linear regression and deep convolutional gated recurrent unit*, Engineering Applications of Computational Fluid Mechanics, Vol. 16 (1), pp. 965–976.

Chiorrini, A., Diamantini, C., Mircoli, A. and Potena, D. (2021) Emotion and sentiment analysis of tweets using bert., *EDBT/ICDT Workshops*, Vol. 3.

Chollet, F. et al. (2015) Keras, `https://keras.io`.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020) Unsupervised cross-lingual representation learning at scale, pp. 8440–8451.

Darwin, C. (1872) *The expression of the emotions in man and animals*, , Vol. .

del Arco, F. M. P., Strapparava, C., López, L. A. U. and Valdivia, M. T. M. (2020) Emoevent: A multilingual emotion corpus based on different events, *LREC*.

Demirci, S. (2014) *Emotion analysis on Turkish tweets*, Master's thesis, Department of Computer Engineering, Middle East Technical University, Ankara, Turkey.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. and Ravi, S. (2020) Goemotions: A dataset of fine-grained emotions, pp. 4040–4054.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding.

Douiji, Y., Hajar, M. and Hassan, A. (2016) *Using youtube comments for text-based emotion recognition*, Procedia Computer Science, Vol. 83, pp. 292–299.

Ekman, P. (1992) *An argument for basic emotions*, Cognition and Emotion, Vol. 6 (3-4), pp. 169–200.

Ekman, P. and Cordaro, D. (2011) *What is meant by calling emotions basic*, Emotion Review, Vol. 3, pp. 364–370.

Ema, R., Islam, T. and Ahmed, H. (2018) *Detecting emotion from text and emoticon*, London Journal of Research in Computer Science and Technology, Vol. 17, pp. 8–13.

Erenel, Z., Adegboye, O. and Kusetogullari, H. (2020) *A new feature selection scheme for emotion recognition from text*, Applied Sciences, Vol. 10, pp. 5351.

Gaind, B., Syal, V. and Padgalwar, S. (2019) *Emotion detection and analysis on social media*, , Vol. . https://arxiv.org/abs/1901.08458

Gala, N. and Brun, C. (2012) Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse de sentiments (spreading polarities among word families: Impact of morphology on building a lexicon for sentiment analysis) [in French], *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, Grenoble, France, ATALA/AFCP, pp. 495–502. https://aclanthology.org/F12-2045

Gao, F., Sun, X., Wang, K. and Ren, F. (2016) Chinese micro-blog sentiment analysis based on semantic features and pad model, *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–5.

García-Martínez, B., Fernández-Caballero, A., Alcaraz, R. and Martínez-Rodrigo, A. (2021) *Cross-sample entropy for the study of coordinated brain activity in calm and distress conditions with electroencephalographic recordings*, Neural Comput & Applic 33, 9343–9352, Vol. .

Ghazi, D., Inkpen, D. and Szpakowicz, S. (2014) *Prior and contextual emotion of words in sentential context*, Computer Speech & Language, Vol. 28 (1), pp. 76 – 92.

Gokaslan, A. and Cohen, V. (2019) *Openwebtext corpus*. [Online]. Available at: http://Skylion007.github.io/OpenWebTextCorpus. (Accessed: 05 November 2023).

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A. and Bengio, Y. (2015) An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Google (2017) *Google colaboratory*. [Online]. Available at: `https://colab.google/`. (Accessed: 05 November 2023).

Grover, S. and Verma, A. (2016) Design for emotion detection of punjabi text using hybrid approach, *2016 International Conference on Inventive Computation Technologies (ICICT)*, Vol. 2, pp. 1–6.

Gupta, R. and Ratinov, L. (2008) Text categorization with knowledge transfer from heterogeneous data sources, *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, AAAI Press, p. 842–847.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009) *The weka data mining software: An update*, SIGKDD Explor. Newsl., Vol. 11 (1), pp. 10–18. https://doi.org/10.1145/1656274.1656278

Hamdi, E., Rady, S. and Aref, M. (2019) A convolutional neural network model for emotion detection from tweets, *in* A. E. Hassanien, M. F. Tolba, K. Shaalan and A. T. Azar (eds), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018*, Cham, Springer International Publishing, pp. 337–346.

Hansen, N., Müller, S. and Koumoutsakos, P. (2003) *Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es)*, Evolutionary computation, Vol. 11, pp. 1–18.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. and Oliphant, T. E. (2020) *Array programming with NumPy*, Nature, Vol. 585 (7825), pp. 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasan, M., Rundensteiner, E. A. and Agu, E. (2014) Emotex: Detecting emotions in twitter messages, ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, May 27-31.

He, R. and McAuley, J. (2016) *Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering*, Proceedings of the 25th

International Conference on World Wide Web, Vol. .

Hu, M. and Liu, B. (2004) Mining and summarizing customer reviews, pp. 168–177.

Hung, J. and Chang, J.-W. (2021) *Multi-level transfer learning for improving the performance of deep neural networks: Theory and practice from the tasks of facial emotion recognition and named entity recognition*, Applied Soft Computing, Vol. 109, pp. 107491.

Hutto, C. and Gilbert, E. (2015) Vader: A parsimonious rule-based model for sentiment analysis of social media text, Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Ibrahim, H., Abdou, S. and Gheith, M. (2015) *Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis*, International Journal of Computer Applications, Vol. 118, pp. 26–31.

Jaidka, K., Singh, I., Chhaya, N., Ungar, L. and Lu, J. (2020) A report of the cl-aff offmychest shared task: Modeling supportiveness and disclosure, AffCon@AAAI.

Jianqiang, Z., Xiaolin, G. and Xuejun, Z. (2018) *Deep convolution neural networks for twitter sentiment analysis*, IEEE Access, Vol. 6, pp. 23253–23260.

Jiwung Hyun, B.-C. B. and Cheong, Y.-G. (2020) [cl-aff shared task]multi-label text classification using an emotion embedding model.

Kasri, M., Birjali, M., Nabil, M., Beni-Hssane, A., El-Ansari, A. and El Fissaoui, M. (2022) *Refining word embeddings with sentiment information for sentiment analysis*, Journal of ICT Standardization, Vol. pp. 353–382.

Ke, Z., Liu, B., Wang, H. and Shu, L. (2020) Continual learning with knowledge transfer for sentiment classification, *ECML/PKDD*.

Kim, Y. (2014) Convolutional neural networks for sentence classification.

Klebanov, B., Burstein, J. and Madnani, N. (2013) *Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds*, ACM Transactions on Speech and Language Processing (TSLP), Vol. 10.

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S. and Prendinger, H. (2018) *Deep learning for affective computing: Text-based emotion recognition in decision support*, Decision Support Systems, Vol. 115, pp. 24–35. https://www.sciencedirect.com/science/ARTICLE/pii/S0167923618301519

Kumar, A. and Albuquerque, V. (2021) *Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language*, ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 20, pp. 1–13.

Kumar, H. M. K. and Harish, B. S. (2020) *A new feature selection method for sentiment analysis in short text*, Journal of Intelligent Systems, Vol. 29 (1), pp. 1122–1134. https://doi.org/10.1515/jisys-2018-0171

Kuta, M., Morawiec, M. and Kitowski, J. (2017) Sentiment analysis with tree-structured gated recurrent units, *in* K. Ekštein and V. Matoušek (eds), *Text, Speech, and Dialogue*, Cham, Springer International Publishing, pp. 74–82.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations.

Lech, M., Stolar, M., Best, C. and Bolia, R. (2020) *Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding*, Frontiers in Computer Science, Vol. 2, pp. 14. https://www.frontiersin.org/article/10.3389/fcomp.2020.00014

Lee, J., Sattigeri, P. and Wornell, G. (2019) Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks, *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (eds), *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc.

Lee, S. Y. M. and Wang, Z. (2015) Multi-view learning for emotion detection in code-switching texts, *2015 International Conference on Asian Language Processing (IALP)*, pp. 90–93.

Lee, Y., Park, C. and Choi, H. (2019) Word-level emotion embedding based on semi-supervised learning for emotional classification in dialogue, *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–4.

Lei, Z., Yang, Y. and Yang, M. (2018) Saan: A sentiment-aware attention network for sentiment analysis, pp. 1197–1200.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z. and Niu, S. (2017) Dailydialog: A manually labelled multi-turn dialogue dataset. https://arxiv.org/abs/1710.03957

Lim, N. (2016) *Cultural differences in emotion: differences in emotional arousal level between the east and the west*, Integrative Medicine Research, Vol. 5 (2), pp. 105–109.

Liu, C., Osama, M. and Andrade, A. (2019) Dens: A dataset for multi-class emotion analysis, pp. 6294–6299.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) *Roberta: A robustly optimized bert pretraining approach*, , Vol. .

Lv, G., Wang, S., Liu, B., Chen, E. and Zhang, K. (2019) Sentiment classification by leveraging the shared knowledge from a sequence of domains, *DASFAA*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. and Potts, C. (2011) Learning word vectors for sentiment analysis, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Association for Computational Linguistics, pp. 142–150. https://aclanthology.org/P11-1015

Malhotra, S., Kumar, V. and Agarwal, A. (2021) *Bidirectional transfer learning model for sentiment analysis of natural language*, Journal of Ambient Intelligence and Humanized Computing, Vol. 12, pp. 1–21.

Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*, Cambridge, UK, Cambridge University Press. http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Mao, X., Chang, S., Shi, J., Li, F. and Shi, R. (2019) *Sentiment-aware word embedding for emotion classification*, Applied Sciences, Vol. 9, pp. 1334.

Matsumoto, K., Matsunaga, T., Yoshida, M. and Kita, K. (2022) *Emotional similarity word embedding model for sentiment analysis*, Computación y Sistemas, Vol. 26 (2).

McAuley, J., Targett, C., Shi, Q. and van den Hengel, A. (2015) Image-based recommendations on styles and substitutes.

McCloskey, M. and Cohen, N. J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem, Vol. 24 of *Psychology of Learning and Motivation*, Academic Press, pp. 109–165. https://www.sciencedirect.com/science/article/pii/S0079742108605368

McInnes, L., Healy, J. and Melville, J. (2020) Umap: Uniform manifold approximation and projection for dimension reduction.

Medhat, W., Hassan, A. and Korashy, H. (2014) *Sentiment analysis algorithms and*

applications: A survey, Ain Shams Engineering Journal, Vol. 5 (4), pp. 1093–1113. https://www.sciencedirect.com/science/ARTICLE/pii/S2090447914000550

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) *Efficient estimation of word representations in vector space*, Proceedings of Workshop at ICLR, Vol. .

Miller, G. A. (1995) Wordnet: A lexical database for english. communications of the acm vol. 38, no. 11: 39-41.

Mohammad, S. (2012) *#emotional tweets*, Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM), Vol. .

Mohammad, S. and Bravo-Marquez, F. (2017) Emotion intensities in tweets, *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, Vancouver, Canada, Association for Computational Linguistics, pp. 65–77. https://www.aclweb.org/anthology/S17-1007

Mohammad, S. M. (2018) Word affect intensities, *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Mohammad, S. M. and Turney, P. D. (2013) *Crowdsourcing a word-emotion association lexicon*, Computational Intelligence, Vol. 29 (3), pp. 436–465.

Moon, S. and Okazaki, N. (2021) Effects and mitigation of out-of-vocabulary in universal language models, *Journal of Information Processing Vol.29 490–503*.

Naderalvojoud, B. and Sezer, E. A. (2020) *Sentiment aware word embeddings using refinement and senti-contextualized learning approach*, Neurocomputing, Vol. 405, pp. 149 – 160. http://www.sciencedirect.com/science/ARTICLE/pii/S0925231220304811

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V. and Wilson, T. (2019) *Semeval-2013 task 2: Sentiment analysis in twitter*, , Vol. . https://arxiv.org/abs/1912.06806

Navarrete, A., Martinez-Araneda, C. and Manzano, C. (2021) *A novel approach to the creation of a labelling lexicon for improving emotion analysis in text*, The Electronic Library, Vol. ahead-of-print.

Nguyen, C. V., Le, K. H., PHAM, H. Q., Pham, Q. H. and Nguyen, B. T. (2022) Learning for amalgamation: A multi-source transfer learning framework for sentiment classification.

NLPCC Evaluation Tasks (2014) [Online]. Available at: `http://tcci.ccf.org.cn/conference/2014`. (Accessed: 05 November 2023).

Ong, D., Wu, Z., Zhi-Xuan, T., Reddan, M., Kahhale, I., Mattek, A. and Zaki, J. (2019) *Modeling emotion in complex stories: the stanford emotional narratives dataset*, IEEE Transactions on Affective Computing, Vol. PP, pp. 1–1.

Pan, S. J. and Yang, Q. (2010) *A survey on transfer learning*, IEEE Transactions on Knowledge and Data Engineering, Vol. 22 (10), pp. 1345–1359.

Pang, B. and Lee, L. (2004) A sentimental education: Sentiment analysis using subjectivity, *Proceedings of ACL*, pp. 271–278.

Pang, B. and Lee, L. (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *Proceedings of ACL*, pp. 115–124.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? sentiment classification using machine learning techniques, *Proceedings of EMNLP*, pp. 79–86.

Park, H. and Kwon, H.-C. (2011) *Improved gini-index algorithm to correct feature-selection bias in text classification*, IEICE Transactions, Vol. 94-D, pp. 855–865.

Pennington, J., Socher, R. and Manning, C. D. (2014) Glove: Global vectors for word representation, *In EMNLP*.

Perikos, I. and Hatzilygeroudis, I. (2016) *Recognizing emotions in text using ensemble of classifiers*, Engineering Applications of Artificial Intelligence, Vol. 51, pp. 191–201. Mining the Humanities: Technologies and Applications. https://www.sciencedirect.com/science/ARTICLE/pii/S0952197616000166

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018) Deep contextualized word representations, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, New Orleans, Louisiana, Association for Computational Linguistics.

Plutchik, R. (1980) A general psychoevolutionary theory of emotion, *in* R. Plutchik and H. Kellerman (eds), *Theories of Emotion*, Academic Press, pp. 3–33.

Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S., Ghosh, R., Chhaya, N., Gelbukh, A. and Mihalcea, R. (2020) Recognizing emotion cause in conversations.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L. and Shulman, E. (2016) Modelling valence and arousal in Facebook posts, *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, San Diego, California, Association for Computational Linguistics, pp. 9–15. https://www.aclweb.org/anthology/W16-0404

Qian, Q., Huang, M., Lei, J. and Zhu, X. (2017) Linguistically regularized lstm for sentiment classification, pp. 1679–1689.

Qin, Q., hu, W. and Liu, B. (2020) Using the past knowledge to improve sentiment classification, pp. 1124–1133.

Quan, C. and Ren, F. (2009) Construction of a blog emotion corpus for Chinese emotional expression analysis, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, Association for Computational Linguistics, pp. 1446–1454. https://www.aclweb.org/anthology/D09-1150

Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) Squad: 100,000+ questions for machine comprehension of text, pp. 2383–2392.

Ranasinghe, T., Orasan, C. and Mitkov, R. (2019) Semantic textual similarity with Siamese neural networks, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria, INCOMA Ltd., pp. 1004–1011. https://aclanthology.org/R19-1116

Ratna, A. A. P., Purnamasari, P. D., Anandra, N. K. and Luhurkinanti, D. L. (2022) Hybrid deep learning cnn-bidirectional lstm and manhattan distance for japanese automated short answer grading: Use case in japanese language studies, *Proceedings of the 8th International Conference on Communication and Information Processing*, ICCIP '22, New York, NY, USA, Association for Computing Machinery, p. 22–27. https://doi.org/10.1145/3571662.3571666

Raunak, V., Gupta, V. and Metze, F. (2019) Effective dimensionality reduction for word embeddings, *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Florence, Italy, Association for Computational Linguistics, pp. 235–243.

Redondo, J., Fraga, I., Padron, I. and Comesaña, M. (2007) *The spanish adaptation of anew (affective norms for english words)*, Behavior research methods,

Vol. 39, pp. 600–5.

Rezaeinia, S. M., Rahmani, R., Ghodsi, A. and Veisi, H. (2019) *Sentiment analysis based on improved pre-trained word embeddings*, Expert Systems with Applications, Vol. 117, pp. 139–147.

Russell, J. (1980) *A circumplex model of affect*, Journal of Personality and Social Psychology, Vol. 39, pp. 1161–1178.

Russell, J. and Mehrabian, A. (1977) *Evidence for a three-factor theory of emotions*, Journal of Research in Personality, Vol. 11, pp. 273–294.

Sabini, J. and Silver, M. (2005) *Ekman's basic emotions: Why not love and jealousy?*, Cognition and Emotion, Vol. 19 (5), pp. 693–712.

Sailunaz, K., Dhaliwal, M., Rokne, J. and Alhajj, R. (2018) *Emotion detection from text and speech: a survey*, Social Network Analysis and Mining, Vol. 8 (1), pp. 28. https://doi.org/10.1007/s13278-018-0505-2

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*, ArXiv, Vol. abs/1910.01108.

Sarker, I. (2021) *Machine learning: Algorithms, real-world applications and research directions*, SN Computer Science, Vol. 2.

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., Alnumay, W. and Smith, A. P. (2021) *A lexicon-based approach to detecting suicide-related messages on twitter*, Biomedical Signal Processing and Control, Vol. 65, pp. 102355.

Savini, E. and Caragea, C. (2022) *Intermediate-task transfer learning with bert for sarcasm detection*, Mathematics, Vol. 10, pp. 844.

Scherer KR, W. H. (1994) *Evidence for universality and cultural variation of differential emotion response patterning*, , Vol. .

Schweter, S. (2020) Berturk - bert models for turkish. https://doi.org/10.5281/zenodo.3770924

Scollon, C., Diener, E., Oishi, S. and Biswas-Diener, R. (2004) *Emotions across cultures and methods*, Journal of Cross-Cultural Psychology, Vol. 35, pp. 304 – 326.

Seyeditabari, A., Tabari, N., Gholizadeh, S. and Zadrozny, W. (2019) *Emotional embeddings: Refining word embeddings to capture emotional content of words*, ArXiv, Vol. abs/1906.00112.

Shaaban, Y., Korashy, H. and Medhat, W. (2021) Emotion detection using deep learning, *2021 16th International Conference on Computer Engineering and Systems (ICCES)*, pp. 1–10.

Sharma, D. A. (2022) *Context-aware Sentiment Analysis on Refined Word Embeddings Word2Vec Model*, TechRxiv, Vol. .

Shastri, L., Parvathy, A. G., Kumar, A., Wesley, J. and Balakrishnan, R. (2010) *Sentiment extraction: Integrating statistical parsing, semantic analysis, and common sense reasoning*, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. .

Shaver, P., Schwartz, J., Kirson, D. and O'Connor, C. (1987) *Emotion knowledge: Further exploration of a prototype approach*, Journal of personality and social psychology, Vol. 52, pp. 1061–86.

Shi, B., Fu, Z., Bing, L. and Lam, W. (2018) Learning domain-sensitive and sentiment-aware word embeddings, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2494–2504.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A. and Gordon, J. (2012) Empirical study of machine learning based approach for opinion mining in tweets, Berlin, Heidelberg, Springer-Verlag.

Singh, M., Jakhar, A. K. and Pandey, S. (2021) *Sentiment analysis on the impact of coronavirus in social life using the bert model*, Social Network Analysis and Mining, Vol. 11 (1), pp. 33.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C. (2013) Recursive deep models for semantic compositionality over a sentiment treebank, Vol. 1631, pp. 1631–1642.

Speer, R. and Chin, J. (2016) *An ensemble method to produce high-quality word embeddings*, , Vol. . http://arxiv.org/abs/1604.01692

Speer, R., Chin, J. and Havasi, C. (2018) Conceptnet 5.5: An open multilingual graph of general knowledge.

Sreeja, P. S. and Mahalakshmi, G. S. (2017) *Emotion models: A review*, International Journal of Control Theory and Applications, Vol. 10, pp. 651–657.

Srivastava, R., Masci, J., Kazerounian, S., Gomez, F. and Schmidhuber, J. (2013)

Compete to compute, Advances in Neural Information Processing Systems, Vol. .

Staiano, J. and Guerini, M. (2014) Depeche mood: a lexicon for emotion analysis from crowd annotated news, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, Association for Computational Linguistics, pp. 427–433. https://www.aclweb.org/anthology/P14-2070

Stojanovska, F., Gievska, S. and Najdenkoska, I. (2018) Detecting emotions in tweets based on hybrid approach, Conference: 15th International Conference on Informatics and Information Technologies, CIIT.

Strapparava, C. and Mihalcea, R. (2008) Learning to identify emotions in text, *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, New York, NY, USA, Association for Computing Machinery, p. 1556–1560. https://doi.org/10.1145/1363686.1364052

Strapparava, C. and Valitutti, A. (2004) WordNet affect: an affective extension of WordNet, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf

Su, J., Cao, J., Liu, W. and Ou, Y. (2021) Whitening sentence representations for better semantics and faster retrieval.

Su, M.-H., Wu, C.-H., Huang, K.-Y. and Hong, Q.-B. (2018) Lstm-based text emotion recognition using semantic and emotional word vectors, *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–6.

Sweeney, C. and Padmanabhan, D. (2017) Multi-entity sentiment analysis using entity-level feature extraction and word embeddings approach, *RANLP*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) *Lexicon-based methods for sentiment analysis*, , Vol. 37 (2).

Tahon, M., Lecorvé, G. and Lolive, D. (2018) *Can we generate emotional pronunciations for expressive speech synthesis?*, IEEE Transactions on Affective Computing, Vol. pp. 1–1.

Tan, S. and Zhang, J. (2008) *An empirical study of sentiment analysis for chinese documents*, Expert Systems with Applications, Vol. 34, pp. 2622–2629.

Tanana, M., Soma, C., Kuo, P., Bertagnolli, N., Dembe, A., Pace, B., Srikumar, V., Atkins, D. and Imel, Z. (2021) *How do you feel? using natural language processing to automatically rate emotion in psychotherapy*, Behavior Research Methods, Vol. 53, pp. 1–14.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T. and Qin, B. (2014) Learning sentiment-specific word embedding for Twitter sentiment classification, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, Association for Computational Linguistics, pp. 1555–1565. https://www.aclweb.org/anthology/P14-1146

Tao, J. and Fang, X. (2020) *Toward multi-label sentiment analysis: a transfer learning based approach*, Journal of Big Data, Vol. 7, pp. 1.

Thelwall, M., Buckley, K. and Paltoglou, G. (2012) Sentiment strength detection for the social web, Journal of the American Society for Information Science and Technology, Volume 63,Issue 1,January, pp 163–173.

Tiwari, S., Raju, M., Phonsa, G. and Deepu, D. (2016) *A novel approach for detecting emotion in text*, Indian Journal of Science and Technology, Vol. 9.

Tjong Kim Sang, E. F. and De Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. https://aclanthology.org/W03-0419

Tocoglu, M. and Alpkocak, A. (2018) *Tremo: A dataset for emotion analysis in Turkish*, Journal of Information Science, Vol. 44, pp. 016555151876101.

Tocoglu, M. and Alpkocak, A. (2019) *Lexicon-based emotion analysis in Turkish*, Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 27, pp. 1213–1227.

Uçan, A., Dörterler, M. and Sezer, E. (2021) *A study of turkish emotion classification with pretrained language models*, Journal of Information Science, Vol. .

Verwimp, L. and Bellegarda, J. R. (2019) *Reverse transfer learning: Can word embeddings trained for different nlp tasks improve neural language models?*, ArXiv, Vol. abs/1909.04130.

Wang, S. and Meng, X. (2018) Multi-emotion category improving embedding for

sentiment classification, Association for Computing Machinery ,Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1719–1722.

Wang, Y., Huang, G., Li, J., Li, H., Zhou, Y. and Jiang, H. (2021) *Refined global word embeddings based on sentiment concept for sentiment analysis*, IEEE Access, Vol. 9, pp. 37075–37085.

Warriner, A., Kuperman, V. and Brysbaert, M. (2013) *Norms of valence, arousal, and dominance for 13,915 english lemmas*, Behavior research methods, Vol. 45.

Waspodo, B., Nuryasin, Bany, A. K. N., Kusumaningtyas, R. H. and Rustamaji, E. (2022) Indonesia covid-19 online media news sentiment analysis with lexicon-based approach and emotion detection, *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–6.

William, P., Gade, R., Chaudhari, R. e., Pawar, A. B. and Jawale, M. A. (2022) Machine learning based automatic hate speech recognition system, *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 315–318.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A. and Spasic, I. (2015) *The role of idioms in sentiment analysis*, Expert Systems with Applications, Vol. 10.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005) Recognizing contextual polarity in phrase-level sentiment analysis, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, Association for Computational Linguistics, pp. 347–354. https://aclanthology.org/H05-1044

Won, M.-S., Choi, Y., Kim, S., Na, C. and Lee, J.-H. (2021) An embedding method for unseen words considering contextual information and morphological information, SAC '21, New York, NY, USA, Association for Computing Machinery, p. 1055–1062. https://doi.org/10.1145/3412841.3441982

Wongpatikaseree, K., Kaewpitakkun, Y., Yuenyong, S., Matsuo, S. and Yomaboot, P. (2021) *Emocnn: Encoding emotional expression from text to word vector and classifying emotions—a case study in thai social network conversation*, Engineering Journal, Vol. 25 (7), pp. 73–82.

Wu, C., Wu, F., Liu, J., Huang, Y. and Xie, X. (2019) Sentiment lexicon enhanced neural sentiment classification, Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1091–1100.

Wu, C., Wu, F., Wu, S., Yuan, Z., Liu, J. and Huang, Y. (2019) *Semi-supervised dimensional sentiment analysis with variational autoencoder*, Knowl. Based Syst., Vol. 165, pp. 30–39.

Wu, Z. and Jiang, Y. (2019) Disentangling latent emotions of word embeddings on complex emotional narratives, Natural Language Processing and Chinese Computing, 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14.

Yang, W., Lu, W. and Zheng, V. (2017) A simple regularization-based algorithm for learning cross-domain word embeddings, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 2898–2904. https://aclanthology.org/D17-1312

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R. and Le, Q. V. (2019) Xlnet: Generalized autoregressive pretraining for language understanding, *NeurIPS*.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. and Zhang, X. (2016) Building chinese affective resources in valence-arousal dimensions, Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)At: San Diego, CA, USA.

Yu, L.-C., Wang, J., Lai, K. and Zhang, X. (2017) Refining word embeddings for sentiment analysis, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 534–539.

Zhang, S., Zhang, X., Chan, J. and Rosso, P. (2019) *Irony detection via sentiment-based transfer learning*, Information Processing & Management, Vol. 56 (5), pp. 1633–1644. https://www.sciencedirect.com/science/article/pii/S0306457318307428

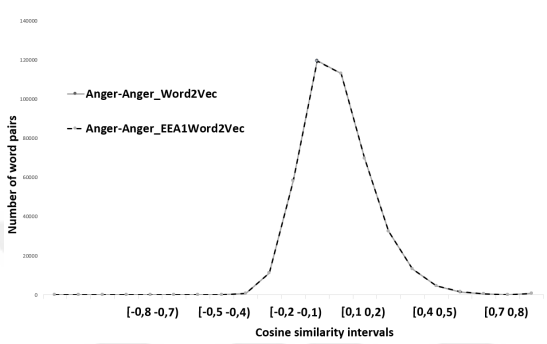
Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S. (2015) *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*, , Vol. .

Zimbra, D., Abbasi, A., Zeng, D. and Chen, H. (2018) *The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation*, ACM Trans. Manage. Inf. Syst., Vol. 9 (2). https://doi.org/10.1145/3185045

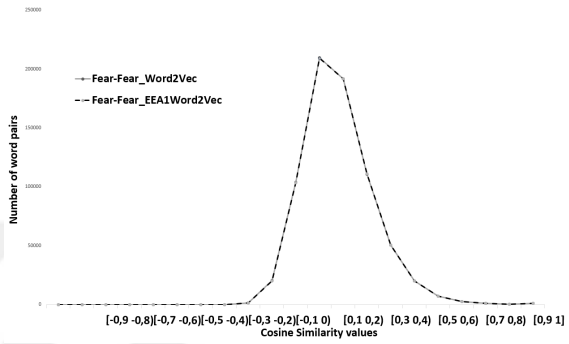
Özbay, F. (2019) *Turkish news dataset*. [Online]. Available at: https://www.kaggle.com/datasets/furkanozbay/turkish-news-dataset/. (Accessed: 05 November 2023).

APPENDICES

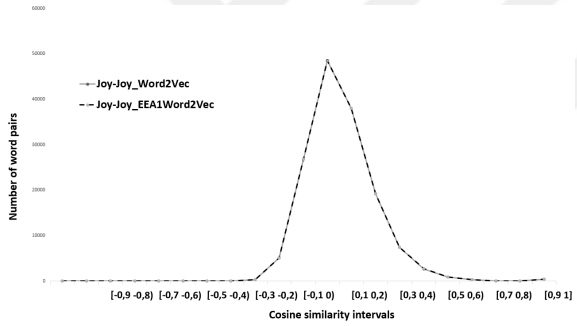
Appendix A - Pairwise Cosine Similarity Histograms (Word2Vec and EEA1_Word2Vec)



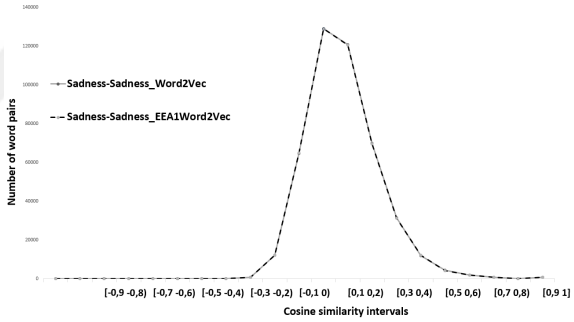
(a) Anger



(b) Fear



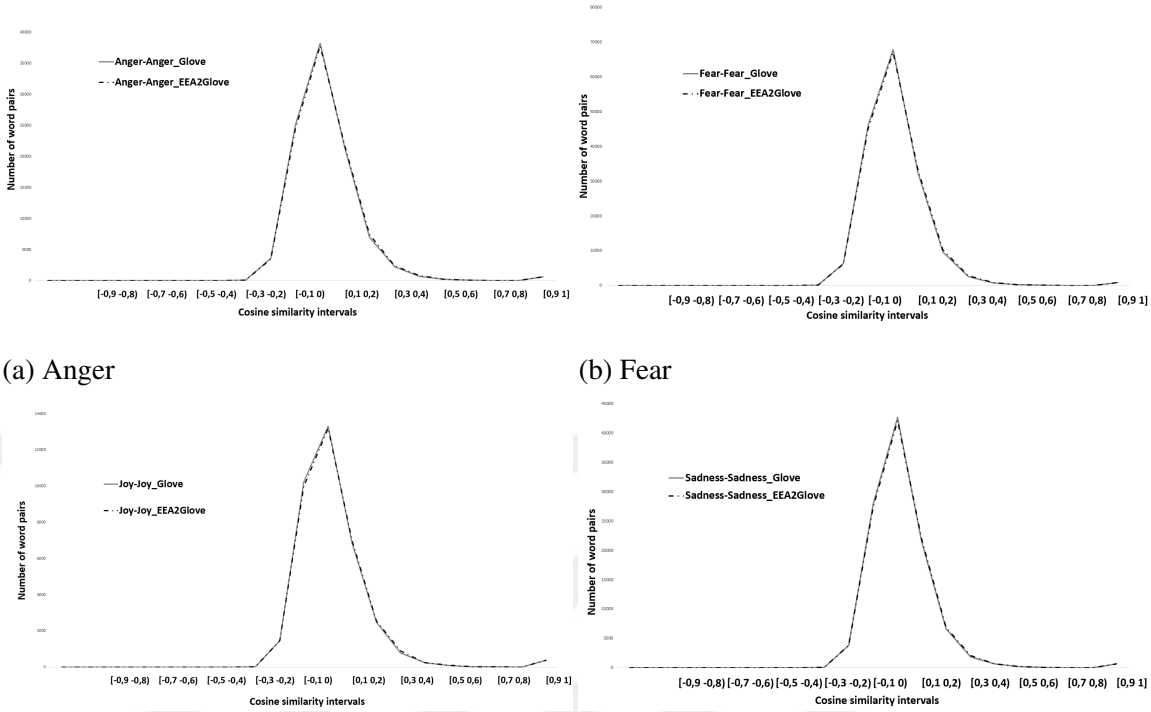
(c) Joy



(d) Sadness

Figure 14. Pairwise CS histograms (Word2Vec and EEA2_Word2Vec).

Appendix B - Pairwise cosine similarity histograms (GloVe and EEA2_GloVe)



(a) Anger

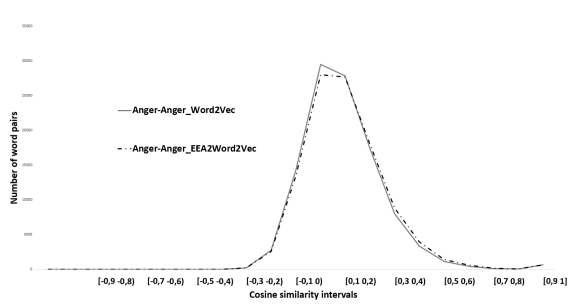
(b) Fear

(c) Joy

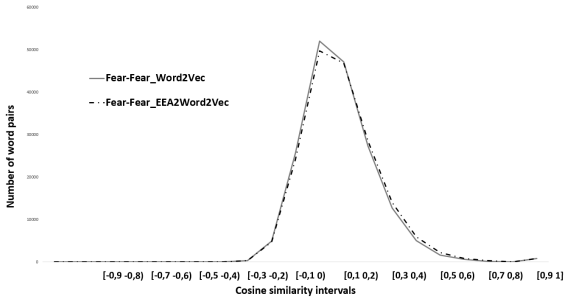
(d) Sadness

Figure 15. Pairwise CS histograms (GloVe and EEA2_GloVe).

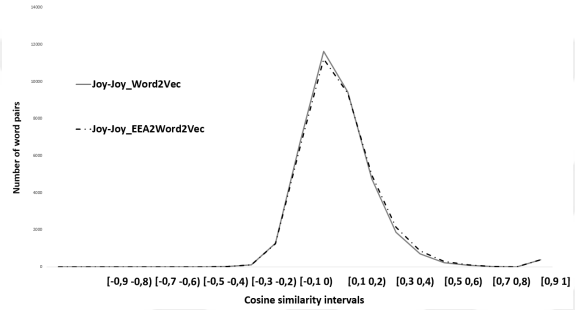
Appendix C - Pairwise cosine similarity histograms (Word2Vec and EEA2_Word2Vec)



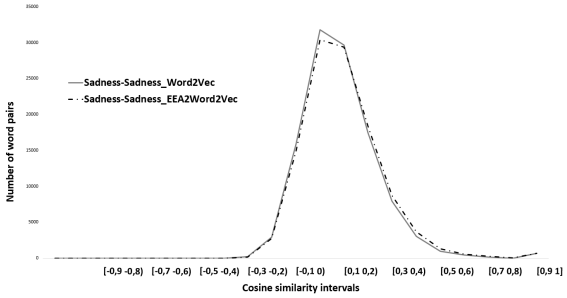
(a) Anger



(b) Fear



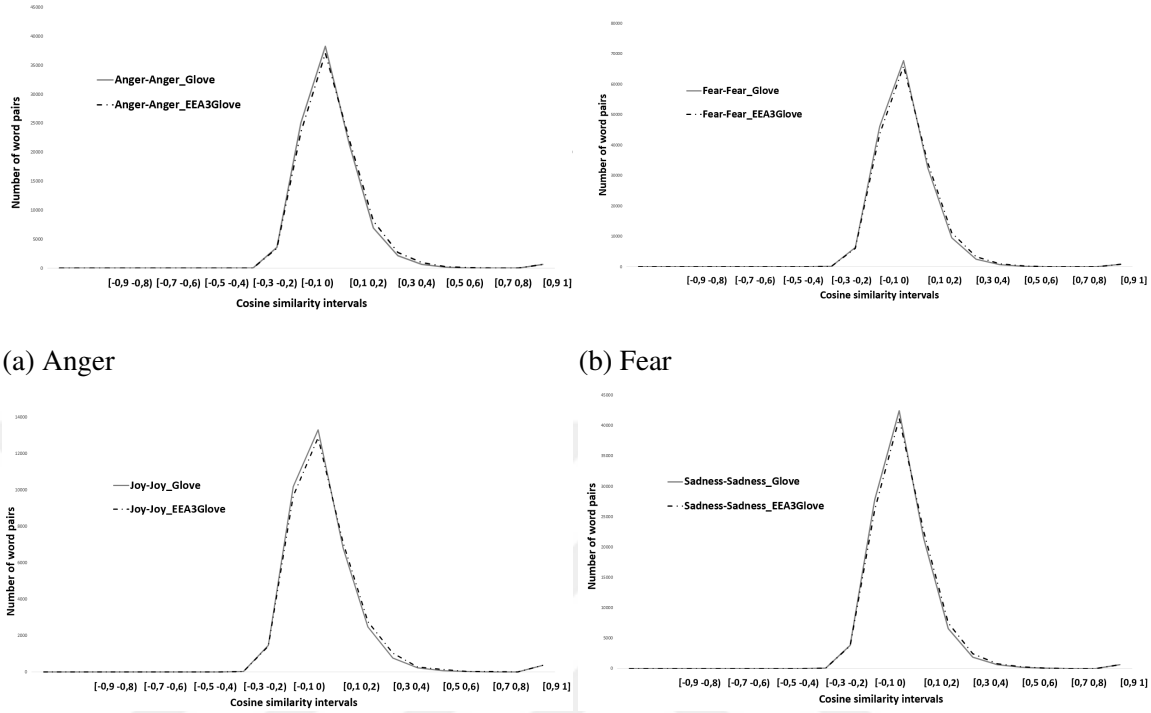
(c) Joy



(d) Sadness

Figure 16. Pairwise CS histograms (Word2Vec and EEA2_Word2Vec).

Appendix D - Pairwise cosine similarity histograms (GloVe and EEA3_GloVe)



(a) Anger

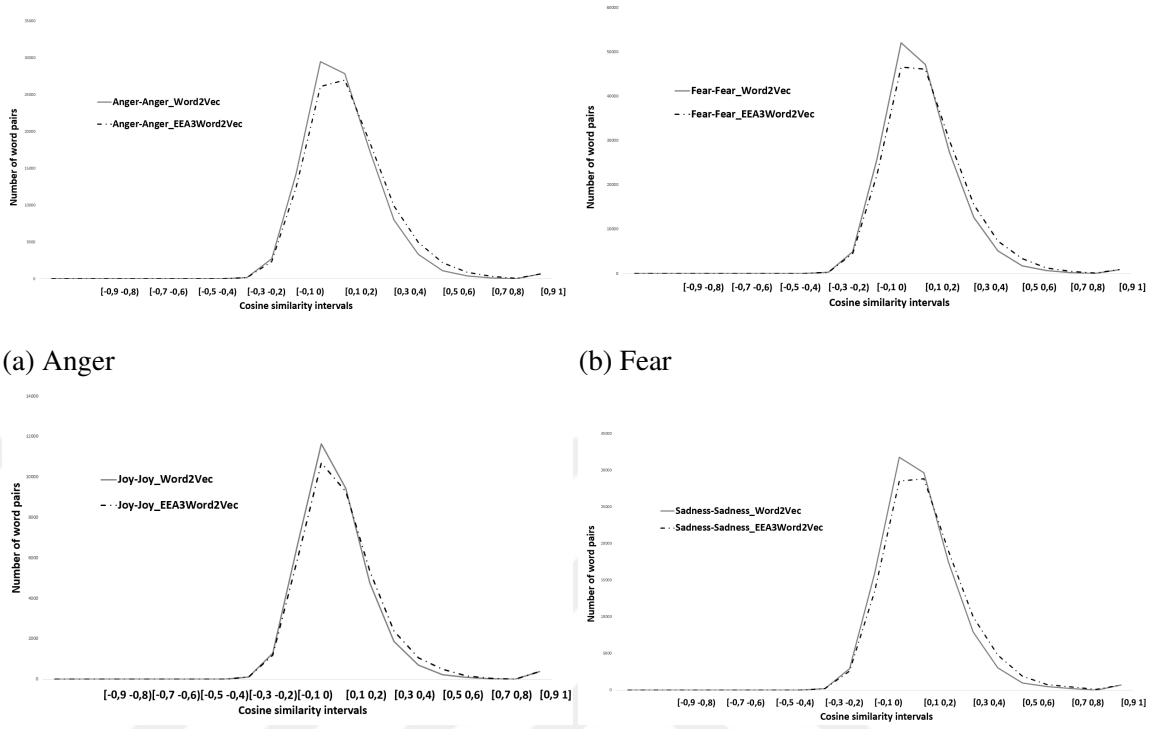
(b) Fear

(c) Joy

(d) Sadness

Figure 17. Pairwise CS histograms (GloVe and EEA3_GloVe).

Appendix E - Pairwise cosine similarity histograms (Word2Vec and EEA3_Word2Vec)



(a) Anger (b) Fear

(c) Joy (d) Sadness

Figure 18. Pairwise CS histograms (Word2Vec and EEA3_Word2Vec).