Optimizing High-Dimensional Text Embeddings in Emotion Identification: A Sliding Window Approach

Hande Aka Uymaz[a] and Senem Kumova Metin[b]
İzmir University of Economics, Department of Software Engineering, İzmir, Turkey
{*hande.aka, senem.kumova*}*@ieu.edu.tr*

Keywords: Natural Language Processing, Emotion, Large Language Models, Vector Space Models.

Abstract: Natural language processing (NLP) is an interdisciplinary field that enables machines to understand and generate human language. One of the crucial steps in several NLP tasks, such as emotion and sentiment analysis, text similarity, summarization, and classification, is transforming textual data sources into numerical form, a process called vectorization. This process can be grouped into traditional, semantic, and contextual vectorization methods. Despite their advantages, these high-dimensional vectors pose memory and computational challenges. To address these issues, we employed a sliding window technique to partition high-dimensional vectors, aiming not only to enhance computational efficiency but also to detect emotional information within specific vector dimensions. Our experiments utilized emotion lexicon words and emotionally labeled sentences in both English and Turkish. By systematically analyzing the vectors, we identified consistent patterns with emotional clues. Our findings suggest that focusing on specific sub-vectors rather than entire high-dimensional BERT vectors can capture emotional information effectively, without performance loss. With this approach, we examined an increase in pairwise cosine similarity scores within emotion categories when using only sub-vectors. The results highlight the potential of the use of sub-vector techniques, offering insights into the nuanced integration of emotions in language and the applicability of these methods across different languages.

1 INTRODUCTION

Natural language processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics that aims to enable machines to understand and generate human language. In text-based natural language processing, the first step is to convert the given textual content into a numerical format that computers can process. These numerical representations are expected to reflect the complex elements of language, including grammatical rules, vocabulary, and various linguistic components. In the field, the process of converting textual data into numerical representations is commonly referred to as vectorization. The combined representation of documents within a common vector space is known as the vector space model (Manning et al., 2008). This model, which is grounded in linear algebra, allows for vector-based operations like addition, subtraction, and similarity calculations.

We can examine vectorization methods in three groups: traditional (i.e., one-hot encoding, TF, IDF), semantic (i.e., Word2Vec (Mikolov et al., 2013) and GloVe (Global Vectors for Word Representation) (Pennington et al., 2014)), and contextual (i.e., BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), GPT (Generative pre-trained transformers) (OpenAI, 2023), ELECTRA (Clark et al., 2020)) methods. Traditional methods represent words as discrete, sparse vectors without capturing semantic meaning. Semantic methods generate dense vectors that are designed to capture semantics but fail to account for word polysemy. Contextual methods create vectors that vary with context, capturing deeper semantics and polysemy information. Considering the problems with traditional methods, such as the increased computational demand as the number of existing words increases and the lack of semantic information, or in semantic vectors, the neglect of polysemy information and having a single vector for each word independent of its context in a sentence, recently, contextual vectors are more frequently used in NLP problems and achieve better success.

[a] https://orcid.org/0000-0002-3535-3696
[b] https://orcid.org/0000-0002-9606-3625

Aka Uymaz, H. and Kumova Metin, S.
Optimizing High-Dimensional Text Embeddings in Emotion Identification: A Sliding Window Approach.
DOI: 10.5220/0012899300003838

In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 1: KDIR*, pages 258-266
ISBN: 978-989-758-716-0; ISSN: 2184-3228

Unlike static word embeddings, models such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2018), and DistilBERT (Sanh et al., 2019) produce embeddings that consider the word sense and polysemy by adapting to the specific context in which a word is used. ELMO employs a bi-directional long short-term memory architecture to create multiple vectors for words in different contexts, enhancing tasks such as question answering and sentiment detection. BERT, introduced by Google, utilizes a multi-layer bidirectional transformer encoder and a masked language model approach, showing performance in various NLP applications through transfer learning. BERT's significant potential and performance have led to the development of efficient variants such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and DistilBERT (Sanh et al., 2019). Beyond BERT-based models, approaches like ULMFit and XLNet have also shown promising results in tasks like sentiment and emotion analysis, further diversifying the landscape of contextual embeddings in NLP.

The vectors created to represent any text unit are high-dimensional vectors (e.g., the vectors produced from BERT-base and BERT-large models have dimensions of 768 and 1024, respectively). When performing classification, measuring similarity, and/or running other procedures employing these high-dimensional vectors, they can lead to significant memory and computational costs, especially when working with large datasets. Furthermore, feature engineering holds great importance in classification problems. Although high-dimensional vectors carry detailed information, not all dimensions may be necessary in the solution of a specific problem. Eliminating irrelevant or low-information features can improve the model's performance and prevent overfitting. Additionally, feature selection can reduce the computational costs and memory requirements of the model, providing a significant advantage. In this context, we investigated the following 3 research questions (RQ) for this study:

RQ1. How can we enhance the effectiveness of vector representations by optimizing computational efficiency?

Our goal was to tackle the computational challenges associated with high-dimensional vectors, particularly when handling large datasets. By employing a sliding window method, we systematically examined recurring patterns within these vectors to enhance computational efficiency.

RQ2. Can we have insights into the nuanced integration of emotions within language representations of text units?

As detailed in Section 3, we investigated whether the method we applied to BERT vectors of words/sentences labeled with different emotions could detect emotional information in specific parts of the vector representations.

RQ3. What are the differences or similarities between the application of an optimization approach on vectors in the English and Turkish languages?

In the literature, while many methods used in the field of NLP on texts demonstrate success in the English language, it is observed that the same method may not yield the same success or effects when applied to different languages. Therefore, both for this reason and to make comparisons, we conducted experiments for the proposed method in both English and Turkish languages. The reason for choosing Turkish as a second language is that it differs significantly from English in terms of grammar. Among the general features of Turkish, its agglutinative structure, vowel harmony, and frequent usage of idioms and proverbs can be counted. For example, the 22-letter Turkish word "Anlamlandıramadıklarım." can be expressed in English as the 6-word sentence "What I couldn't make sense of.".

In summary, we examined whether certain dimensions within the representations of text units might include concealed information, such as emotions. This led us to explore the possibility of detecting emotional cues through a detailed analysis of these dimensions. To achieve this goal, we employed a sliding window approach to partition vectors and identify consistent patterns, aiming to enhance computational efficiency and gain a deeper understanding of the integration of emotions within these vectors. Our experiments involve emotion lexicon words and emotionally labeled sentences, and we also utilized BERT as an embedding model. Ultimately, this approach, which offers a new perspective on emotional representation, can be applied to any text unit, any embedding model, and any hidden information that can be detected. The contributions of the study can be listed as follows:

1. A dimensionality reduction technique through a sliding window approach is introduced to partition high-dimensional vector representations of texts into smaller sub-vectors, improving computational efficiency while maintaining or enhancing the effectiveness of representations.
2. Specific sub-vectors within BERT vectors that contain emotional information have been identified, suggesting that emotional clues are localized within certain dimensions of the vectors.
3. Experiments utilizing only sub-vectors are conducted in both English and Turkish, demonstrating the effectiveness of the proposed method for

two languages with different grammatical structures.

In the subsequent sections of the paper, Section 2 provides a literature review, Section 3 details the proposed method, Section 4 presents the experiments and results, and Section 5 concludes with the findings and implications.

2 LITERATURE REVIEW

Vector space models refer to the numerical representation of text units (like words or phrases) in a vector space. As can be seen in Figure 1, the models can be considered in two different groups: context-free and contextual models.

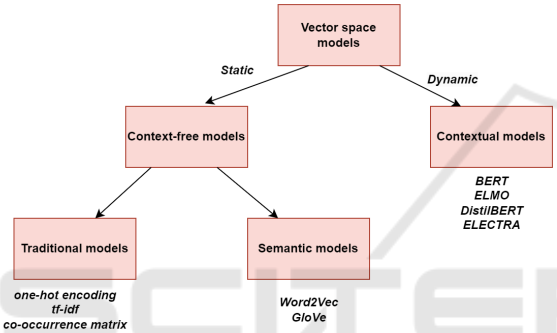

Figure 1: Vector space models.

From the context-free models, traditional models like one-hot encoding, tf-idf, and co-occurrence matrix representation lack semantic understanding. For instance, co-occurrence matrix representation counts word occurrences but fails to capture the nuances of word meanings and their semantic associations. Thus, these models struggle to comprehend the deeper meaning and context of language, which brings a drawback in tasks requiring semantic understanding, such as sentiment analysis and language translation. Semantic embeddings like Word2Vec and GloVe provide the representation of words with similar meanings close together in vector space. Capturing semantic relationships between words helps these models manage tasks like semantic similarity and word analogy. Although they have been a significant innovation in the field of NLP for containing semantic information, these models generate only a single static vector for each word. In other words, these models that produce context-free vectors do not consider polysemy and content.

Contextual models like BERT and ELMO produce different embeddings based on the context in which they are used, even for the same words with different meanings. These contextual models are designed to capture nuanced information in language and represent the complex relationships between words in various contexts. The representations are based on high-dimensional embeddings, typically ranging from 512 to 1024 dimensions. For instance, BERT has two versions: BERT-base with 768 dimensions and BERT-large with 1024 dimensions. Similarly, ELMO embeddings have 1024 dimensions. Two embedding models from GPT, *text-embedding-3-small*, and *text-embedding-3-large*, produce vectors with lengths of 1536 and 3072, respectively. While these high-dimensional embeddings capture rich and detailed linguistic information, they have challenges such as increased computational complexity and memory requirements. In the literature, dimensionality reduction techniques, such as PCA (Principal Component Analysis) and t-SNE (t- Stochastic Neighbor Embedding), are often used to address these issues while preserving the performance in several tasks (Raunak et al., 2019; Ayesha et al., 2020; George and Sumathy, 2022; Álvaro Huertas-García et al., 2022; Zhang et al., 2024). For example, (Zhang et al., 2024) study investigates the effects of reducing the dimensionality of high-dimensional sentence embeddings. The research assesses various unsupervised dimensionality reduction techniques, such as PCA, SVD (truncated Singular Value Decomposition), KPCA (Kernel PCA), GRP (Gaussian Random Projections), and autoencoders, to compress these embeddings. The aim is to cut down on storage and computational expenses while preserving performance in different downstream NLP tasks. Their findings indicate that PCA is the most efficient method, achieving a 50% reduction in dimensionality with only a 1% performance loss. Notably, for some sentence encoders, reducing dimensionality even enhanced accuracy. In the research conducted by (Su et al., 2021), they utilize a technique referred to as “whitening”, which is based on PCA (Principal Component Analysis), to process BERT sentence representations. This method reduces the embedding size to 256 and 384, aiming to address the issue of anisotropy and diminish dimensionality. Experimental results on seven benchmark datasets demonstrate that their method substantially enhances performance and reduces vector size, optimizing memory storage and accelerating retrieval speed.

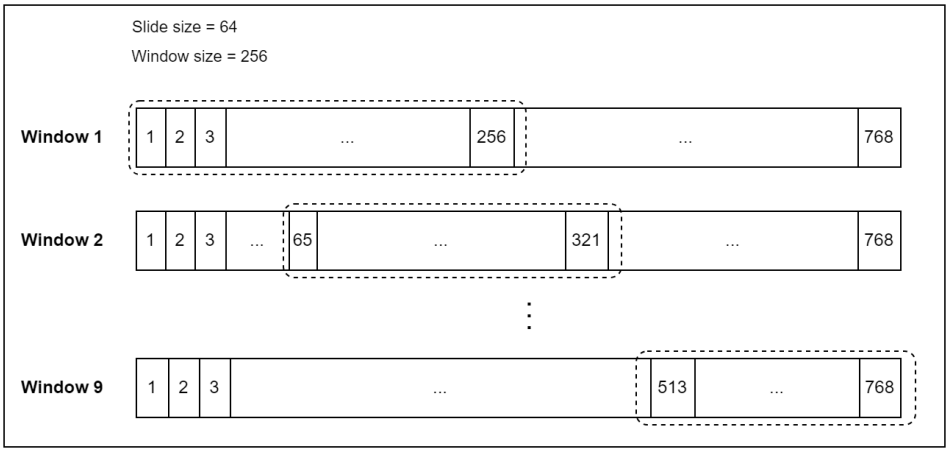

Figure 2: Framework for vector partitioning with sliding window technique.

3 METHOD: DIMENSIONALITY REDUCTION

Although contextual embeddings effectively capture both semantic and contextual knowledge, their high-dimensional vectors can be both space-consuming and computationally expensive, especially with large datasets. Additionally, specific dimensions or segments of these vectors might capture information related to specific features of language or properties of the text unit they represent. In this study, we proposed an alternative approach that emphasizes identifying patterns within vectors of any text unit, thereby reducing the complexity of the analysis. This approach is adaptable to any vectorization model.

We conducted an experimental study to find sub-vectors containing emotion information within BERT vectors of sentences and words labeled with different emotion categories (anger, fear, sadness, and joy) and measured the performance of word and sentence representations using only these sub-vectors. To perform a comparative study and observe the method's effectiveness in different languages, we conducted the experiments in both English and Turkish. Our proposed methodology is summarized as follows:

1. A sliding window technique is employed to examine and extract meaningful patterns from BERT vectors. This method divides the vectors into smaller, fixed-size parts (windows), enabling us to obtain local contextual information.
2. Cosine similarity between words (both for English and Turkish) labeled with the same emotion category is measured using only certain windows of BERT vectors for word representations. Here, an increase in cosine similarity values is expected if there is emotion-specific information in certain windows of the vectors.

To determine the window size for the sliding window technique we referred to the study of (Su et al., 2021). They proposed another dimensionality reduction technique to decrease BERT vectors to lengths of 256 and 384. Thus, in our study, the window size is selected as 256. Initially, BERT word vectors, labeled by 4 different emotion categories and having a length of 768, are divided into sub-vectors with a window size of 256. The slide size is determined to be 64 to cover every dimension of the BERT vectors. For example, the first sub-vector (window) starts at dimension 1 and ends at dimension 256, and the second one spans from dimension 65 to 321 as can be seen detailly in Figure 2. To sum up, employing the sliding window technique, we segmented the 768-dimensional word BERT vectors into nine subvectors.

4 EXPERIMENTS

In this study, we utilized the NRC English emotion lexicon (Mohammad and Turney, 2013) words and the Turkish-translated NRC emotion lexicon (TT-NRC) (Aka Uymaz and Kumova Metin, 2023). Both lexicons are annotated by Plutchick's (Plutchik, 1980) emotion categories. In the experimental study, we considered the lexicon words labeled by four emotion categories, namely anger, fear, sadness, and joy, for both languages. The initial step was obtaining BERT vectors of each lexicon word. Because BERT constructs vectors for words based on their surrounding context, the words and the sentences constituting the words should be given as parameters to BERT. We

Table 1: Pairwise in-category cosine similarity results of *English words* while using only one window.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
In-category cosine similarity	**Anger-Anger**	0.249	0.597	0.628	0.633	0.630	0.361	0.256	0.244	0.233
	Fear-Fear	0.220	0.607	0.634	0.640	0.637	0.340	0.236	0.226	0.215
	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

Table 2: Pairwise in-category cosine similarity results of *Turkish words* while using only one window.

		Windows								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
In-category cosine similarity	**Anger-Anger**	0.288	0.330	0.300	0.324	0.312	0.767	0.766	0.768	0.775
	Fear-Fear	0.276	0.318	0.292	0.321	0.306	0.760	0.760	0.761	0.768
	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

followed the same technique as (Aka Uymaz and Kumova Metin, 2023) for deriving BERT vectors utilizing the collection of three sentence datasets labeled by emotion four emotion categories (anger, fear, sadness, joy): TEI (Mohammad and Bravo-Marquez, 2017), TEC (Mohammad, 2012), and TREMO (Tocoglu and Alpkocak, 2018). After applying our proposed sliding window technique, we divided each BERT vector of lexicon words into 9 sub-vectors. Then, utilizing these sub-vectors individually to represent each word vector, we measured the pairwise cosine similarity score between each word belonging to emotion categories (in-category cosine similarity). Cosine similarity takes values between 0 and 1. 0 indicates that two vectors are completely different, while 1 means they are identical. In this study, a high cosine similarity score may indicate that certain sub-vectors are better at capturing that emotion category. For instance, when assessing cosine similarity between two words labeled with *joy*, we utilized only the subvectors spanning dimensions 1 to 256 and computed the cosine similarity. This procedure was repeated for other windows, resulting in nine cosine similarity experiments for each word represented by a single subvector. The outcomes were shown as heat maps in Tables 1 and 2 for English and Turkish lexicon words, respectively.

The heat maps reveal that certain dimensions within BERT vectors contain emotional clues. Consequently, employing specific subsets of these vectors in cosine similarity assessments yields higher similarity compared to others. This implies that focusing on subsets can be sufficient instead of utilizing all 768-dimensional vectors. Specifically, our examination of English word vectors identified emotional data within windows 2, 3, 4, and 5, while in Turkish, emotional intensity may also found within windows 6, 7, 8, and 9.

Following analyzing the in-category cosine similarity among lexicon words represented by a window-based vector, we applied these findings to a specific process in emotion identification: emotion enrichment of text units. The experimental study on emotion enrichment consists of two phases: sentence sub-vector construction and emotion enrichment on sentence vectors.

In this phase of the experimental study, we utilize the TEI (Mohammad and Bravo-Marquez, 2017), TEC (Mohammad, 2012), and TREMO (Tocoglu and Alpkocak, 2018) datasets. Among these, TREMO is a Turkish dataset, while the others are English datasets. To enable experiments with both English and Turkish, we translated the English datasets into Turkish and the Turkish dataset into English. Subsequently, we selected 500 sentences from each emotion category (anger, fear, sadness, joy) randomly, to construct the Emotion Sentence Dataset (ESD) used in the sentence-based experiments. In order to construct sentence sub-vectors, firstly, as an alternative to using the 768-dimensional BERT vectors for sentence representations, we utilized the sub-parts identified as having emotional information prior to word-based experiments for both English and Turkish as can be seen in detail in Figure 3. We combined the sub-parts that yielded the best results in each language. For instance, it was found that the English BERT vectors had more emotive information in sub-vectors 2, 3, 4, and 5. These sub-parts were concatenated to create a vector that spans from the start of the second window's dimensions to the end of the fifth window's dimensions. The process for combining these sub-vectors is illustrated in Figure 4.

Later, we observed the success of BERT vectors and sub-vectors from sentences in both languages in the emotion enrichment process (EEP). In studies on emotion classification or detection, emotion/sentiment enrichment is a frequently researched process in the literature (Agrawal et al., 2018; Wongpatikaseree et al., 2021; Matsumoto et al., 2022). It

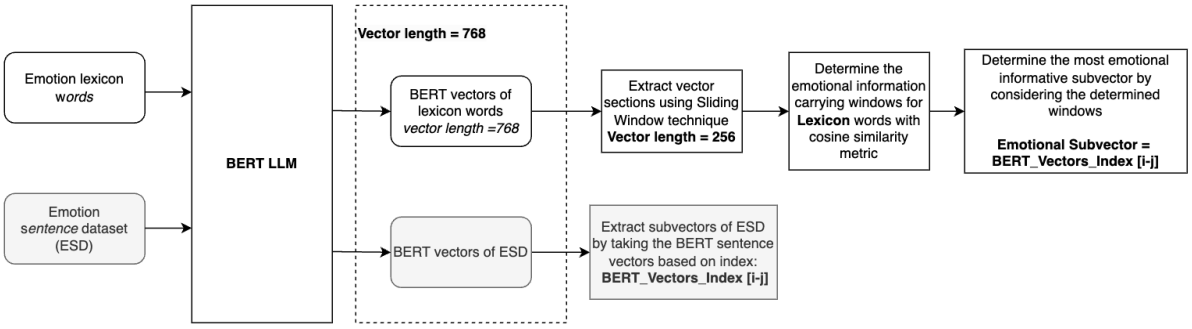

Figure 3: Flowchart for dimensionality reduction for *word* and *sentence* vectors.

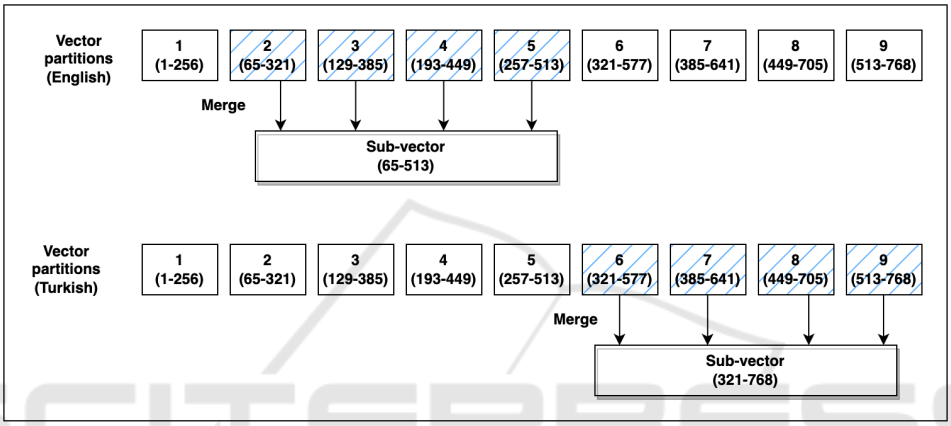

Figure 4: Framework for extracting sub-vectors.

has been observed in studies that although semantic and contextual embeddings demonstrate significant success in representing any text unit, they have some shortcomings in expressing emotional information. Therefore, it has been suggested that these vectors be enhanced by adding emotional information. Studies using cosine similarity-based or classification-based approaches with vectors containing emotional information have shown higher success. Various methods have been proposed in the literature. In this study, we applied the emotion enrichment method proposed by (Aka Uymaz and Kumova Metin, 2023) to our English and Turkish sentence datasets. In summary, this method works by comparing the vectors to be enriched with the vectors of emotion lexicon words. In this comparison, the similarity (cosine similarity) of each word to the emotional words in the lexicon is calculated. The closest emotional words are identified, and their vectors are used to enhance the original word's vector by weighting and averaging them based on their emotional relevance. Finally, a hybrid word representation is constructed by integrating semantic/contextual and emotional embeddings.

In the experiments involving the emotion enrichment process, we used Turkish and English sentences as the text units to be enriched with emotional information. Then, we calculated pairwise in-category cosine similarity scores within every emotion category before and after enrichment. For the vector representation of the sentences, we used 768-dimensional BERT vectors and the BERT sub-vectors obtained in the previous stage. The lexicons we used in the emotion enrichment process were the NRC and TT-NRC lexicons. We followed the same procedure for the vector representation of the lexicon words as we did for the sentences. That is, we first represented the lexicon words with BERT, then subjected the words to the enrichment process as in (Aka Uymaz and Kumova Metin, 2023), and finally obtained their sub-vectors.

Tables 3 and 4 present the emotion enrichment process of English and Turkish sentences by representing them with BERT and BERT sub-vectors. The first row in each table presents the average cosine similarity results within emotion categories for sentences, using BERT vectors of 768 lengths without additional enrichment. We used these values as a baseline and evaluated the outcomes of various enrichment combinations in comparison, showcasing the increments as percentages in the tables. In the second row,

Table 3: English Sentence embeddings enrichment with several combinations. (The best results are shown in bold.).

Sentence embedding	Enrichment process	Enrichment by	In-category similarity (% improvement)									
			Anger		Fear		Joy		Sadness		Average	
BERT	-	-	0,610	-	0,593	-	0,623	-	0,597	-	0,606	-
BERT	✓	**Emotion Lexicon Words (BERT + EEP)**	0,844	38,36%	0,838	41,32%	0,879	41,09%	0,845	41,54%	0,852	40,57%
BERT Subvector	✓	**Emotion Lexicon Words Subvector (BERT + EEP)**	**0,885**	**45,09%**	**0,880**	**48,44%**	**0,905**	**45,28%**	**0,883**	**47,88%**	**0,888**	**46,65%**

Table 4: 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	✓	**Emotion Lexicon Words (BERT + EEP)**	0,922	22,61%	0,931	24,63%	0,943	24,41%	0,927	24,10%	0,931	23,93%
BERT Subvector	✓	**Emotion Lexicon Words Subvector (BERT + EEP)**	**0,953**	**26,67%**	**0,959**	**28,45%**	**0,966**	**27,45%**	**0,956**	**28,03%**	**0,959**	**27,65%**

768-dimensional BERT vectors were subjected to the emotion enrichment process with 768-dimensional lexicon word vectors, while in the third line, alternatively, both sentence and lexicon word sub-vectors were used to represent and subjected to the emotion enrichment process. As can be seen, in both languages, the in-category cosine similarity results of emotionally enriched sentence vectors have yielded the best outcome when subvectors of both sentence and lexicon words' vectors are utilized for all four emotions. The best results in both languages have been observed in the *joy* emotion category with scores of 0,905 for English and 0,956 for Turkish. These results provide promising insights into the effectiveness of using sub-vectors instead of high-dimensional vectors, both for the emotion enrichment process and potentially reducing computational costs due to decreased vector size.

5 CONCLUSION

Natural language processing stands as a bridge between computer science, artificial intelligence, and linguistics, which focuses on machines that can comprehend and generate human language better through extensive analyses in various domains such as sentiment analysis, text summarization, and classification.

One of the most important processes in NLP studies is vectorization, which is simply the transformation of textual data into numerical representations, for any computational analysis. Unlike traditional methods like TF-IDF, newer techniques such as Word2Vec and BERT gained popularity because of having semantic and contextual knowledge, respectively, enriching the depth of linguistic representation. Especially, contextual models like BERT and its derivatives not only capture word semantics but also adapt to the nuanced contextual usage and polysemy, thereby addressing the limitations of traditional and semantic approaches.

However, the use of high-dimensional vectors poses computational challenges, particularly in large datasets. Feature selection and computational efficiency enhancements emerged as considerations as optimization strategies. In this context, we identified three research questions for our study:

RQ1. How can we enhance the effectiveness of vector representations by optimizing computational efficiency?

RQ2. Can we have insights into the nuanced integration of emotions within language representations of text units?

RQ3. What are the differences or similarities between the application of an optimization approach on vectors in the English and Turkish languages?

Firstly, related to *RQ1*, we proposed a sliding window technique to partition vectors into smaller, fixed-size parts, enabling the extraction of local contextual information. This method was evaluated through pairwise cosine similarity metric among emotion lexicon words which were annotated by four emotion categories, using both English and Turkish for addressing *RQ3*.

Our experimental findings as an answer to *RQ2* revealed that utilizing BERT vectors demonstrated that certain dimensions are more informative regarding emotional content. This suggests that using subvectors may effectively capture emotional clues and nuances in the languages, potentially reducing the need to utilize entire high-dimensional vector representations.

In the subsequent phase, we applied our findings to sentence vectors, constructing sentence sub-vectors based on the identified emotional dimensions (according to determined windows) from the word-based ex-

periments. Then, to test our hypothesis on the effectiveness of using specific vector segments, we conducted experiments with these subvectors in comparison to using the original vectors in a case study related to the emotion enrichment process on vectors. This process simply incorporates vectors with additional emotional information. The comparative analysis between English and Turkish highlighted the adaptability of our method to different languages, acknowledging the grammatical and structural differences of Turkish.

When we examined the experimental results, we found that using specific sub-vectors instead of the original BERT vectors was both sufficient and could improve performance in cosine similarity calculations within emotion categories at both the word and sentence levels. As far as we know, this perspective and method have not been previously studied in terms of their applicability to any text unit represented by any vectorization method. Additionally, this approach is might be effective in capturing different types of information in vector representations and adapting to different problems.

In future studies, similar experiments can be conducted on other large language models (e.g., GPT models (OpenAI, 2023), RoBERTa (Liu et al., 2019), ELMO (Peters et al., 2018)) that have shown successful results in the literature. This approach may enable the investigation of different sub-vectors containing emotional information in these models and to get new perspectives. In our study, we carried out comparative analyses on English, a language rich in resources, and Turkish, an agglutinative language with fewer resources and a different grammatical structure. This study can be expanded to include languages from different language families and with various features. Additionally, vectors can be reanalyzed for different problems or information searches and the effectiveness of the approach in various scenarios can be examined.

REFERENCES

Agrawal, A., An, A., and Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aka Uymaz, H. and Kumova Metin, S. (2023). Emotion-enriched word embeddings for Turkish. *Expert Systems with Applications*, 225:120011.

Ayesha, S., Hanif, M. K., and Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators.

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

George, L. and Sumathy, P. (2022). An integrated clustering and bert framework for improved topic modeling.

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.

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.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

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

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

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

Mohammad, S. and Bravo-Marquez, F. (2017). Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

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

OpenAI (2023). Gpt-large language model.

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

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *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 Plutchik, R. and Kellerman, H., editors, *Theories of Emotion*, pages 3–33. Academic Press.

Raunak, V., Gupta, V., and Metze, F. (2019). Effective dimensionality reduction for word embeddings. In Augenstein, I., Gella, S., Ruder, S., Kann, K., Can, B., Welbl, J., Conneau, A., Ren, X., and Rei, M., editors, *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, Florence, Italy. Association for Computational Linguistics.

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

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

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

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*, 25(7):73–82.

Zhang, G., Zhou, Y., and Bollegala, D. (2024). Evaluating unsupervised dimensionality reduction methods for pretrained sentence embeddings.

Álvaro Huertas-García, Martín, A., Huertas-Tato, J., and Camacho, D. (2022). Exploring dimensionality reduction techniques in multilingual transformers.