2024 4th Int. Conf. on Innovations in Science, Engineering and Technology (ICISET)
26-27 October 2024, Chittagong, Bangladesh

Consistency of Contextual Embedding in Literary Texts

Md Minhazul Islam*, Md Tanbeer Jubaer†, Mohammad Harun Or Rashid‡
*†*Department of Computer Science & Engineering*
‡*Department of Humanities*
Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
E-mails: minannu.2001@gmail.com, tanbeerjubaer@gmail.com,
harunmr9@gmail.com

***Abstract*—Despite the fact that computational analysis has opened a wider window for literature and its pertinent text-based studies, it is not well-proven whether the contextual embedding technique works better with seemingly complex literary texts. In this regard, this research has investigated the consistency of contextual embedding in a literary corpus. Based on a small corpus consisting of two novels of Ian Fleming and two short stories of Guy De Maupassant, this study employed the formidable capabilities of BERT to dissect the consistency of contextual embedding in the literary texts. By the combination of a fine-tuned BERT model and the extraction of contextual embeddings, this study performed five correlated comparative studies among two shorter and two longer literary texts. A series of comparative studies among the assigned literary texts uncovered the striking contrasts within each narrative. this ensured the comprehensive examination of the semantic relationships and thematic associations between words. This study confirmed that contextual embedding worked significantly better with literary texts and produced consistent outcomes while preserving the thematic accuracy of the narratives.**

***Index Terms*—Contextual Embedding, BERT, Literary Analysis, Short Stories, Novels**

I. Introduction

Contextual embedding has already been used in many disciplines thanks to its sense-level representation and capabilities of showing context-based overt and covert word relationships. A handful of models appeared to be consistent in producing results that have confirmed the usability of this technique. As a result, it offers significant benefits not only for various natural language processing tasks but also for diversified fields to find context-based word patterns. [1] showed that the contextual embeddings learned by particular models "capture subtle statistical dependencies reflecting syntactic, semantic, and pragmatic relationships among words." They also opined that this technology is "capable of generating novel sentences with human-like, context-sensitive linguistic structure." Consequently, it has been used in diversified fields with varied applications. However, it is not clear how this method is capable of addressing the seemingly complex literary texts that are specially composed based on both imaginary and real-time experiences. To the best of our knowledge, contextual embedding has neither been applied using the literary dataset nor been used to investigate the literary patterns of the literary texts. In order to address this gap, this paper is going to analyze how consistent result this contextual embedding produces when literary texts are analyzed by applying this technique. In order to materialize this objective, this paper is going to work on a tiny corpus consisting of two short stories and two novels.

II. Literature Review

This research has used the contextual embedding technique, one of the robust computational text analysis tools, to show the patterns found in the seemingly large set of the literary corpus. [1] opined that it is different from other traditional embeddings since it moves beyond word-level semantics. They further said that this context-based technique can "capture many syntactic and semantic properties of words under diverse linguistic contexts." BERT [2] (Bidirectional Encoder Representations from Trans-formers) which is a transformer-based deep learning model developed by Google for NLP tasks has been used here to understand the context of the text. Till today, BERT has already proven its superiority in terms of the performance of contextual embeddings which has already been addressed in many studies[3]. Its popularity sustained over many domains of NLP fields and applications due to its reliable performance. [4] reviewed forty-nine articles from different domains such as medicine, Psychology, Engineering, etc., and found that the most popular contextualized embedding model in the reviewed papers is BERT which covers thirty-seven papers of the reviewed studies.[5] showed that a BERT-based model pre-trained on clinical corpora achieved superior performance across all concept extraction. [6] showed how contextualized word embeddings capture human-like distinctions when it is analyzed with English word senses. In this regard, they show how humans' judgments are correlated with distances between senses in the BERT embedding space, which confirms the superiority of this technique. [7] examined the popular contextual encoders and found that BERT generally does stronger performance on downstream tasks, suggesting that BERT carries a greater capacity to handle longer distances. A handful of studies have indeed been focused on word embedding-based literary analysis. For example, by using word embedding techniques, [8] measured the gender bias in

979-8-3503-5549-9/24/$31.00 ©2024 IEEE

TABLE I: DATASET AT A GLANCE

Group	Book Name	Genre	No of Lines
Group 1	The Necklace	Realistic Fiction	2879
	The Jewelry		2840
Group 2	Diamonds are forever	Detective	70965
	From Russia With Love		80576

nineteenth-century fiction [9]. Another work has been done on temporal embedding analysis where they used transformer models for narrative text understanding. Using the embedding technique, they attempted to find the correlation between whether two characters belong to the same family [10]. Using a similar fashion and technique, another study performed sentiment analysis of Chinese literary text based on deep learning[11]. However, no studies have so far been addressed using contextual embedding techniques. That's why this research is going to address this issue which is expected to be one of the first of its kind that analysed literary text using contextual embedding model. In order to make a sharp contrast with the very simple statistical embedding model, this research also made a comparative study between the TF-IDF based embedding model and the BERT-based context embedding model.

III. Materials and Methods

A. Dataset Description

For this analysis, we selected four popular literary texts: two short ones and two slightly longer ones. The first group consists of short stories in the realistic fiction genre, while the second group includes detective novels. The classic short stories 'The Necklace' and 'The Jewelry' by Guy De Maupassant were chosen for the first group, while Arthur Conan Doyle's popular fiction 'Diamonds are Forever' and 'From Russia with Love' were chosen for the second group. Table I describes the dataset which includes the total line numbers of each text.

B. Preprocessing

The first stage of the data preprocessing pipeline begun with the extraction of the content from files stored in the .txt format. After that, a list was compiled that includes all sentences found within the text files. Next, several cleansing operations were performed to address the inconsistencies within the text corpus. This process involved eliminating unnecessary whitespace characters, abbreviations, and any unidentified Unicode values. Moreover, all texts were converted to lowercase to ensure consistency in the word level. After removing any null or empty sentences from the dataset, the texts were segmented into individual sentences using punctuation marks like full stops, question marks, and exclamation marks. This thorough preprocessing procedure guaranteed the refinement and standardization of the textual data, which helped to prepare for further analysis and model training. Figure 1 describes the workflow of this study.

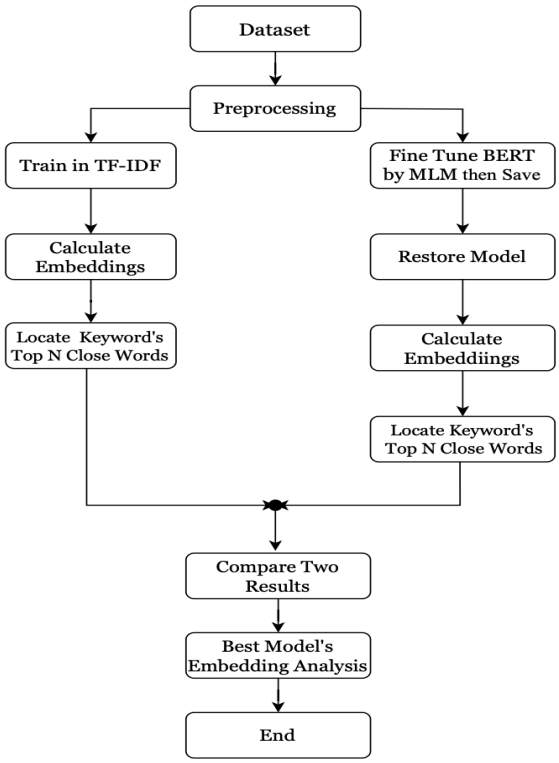

Fig. 1: Workflow

C. Vectorization Approach

In this section, we took all the processed sentences and made their embeddings. This research used two models relating to word and contextual embeddings which also require its own vectorization process. This step is described below separately.

1) BERT Vectorization: In this process, all sentences were tokenized by the BERT tokenizer. After tokenization it returned a token and attention mask. Some words may be broken down into small parts to keep relevant and small vocabulary. This process is called sub word level tokenization. These tokens and attention masks then feed to the model. For example, for the following sentence "One evening her husband returned an elated bearing in his hand a large envelope," after tokenization, tokens are ['[CLS]', 'one', 'evening', 'her', 'husband', 'returned', 'el', '##ated', 'bearing', 'in', 'his', 'hand', 'a', 'large', 'envelope', '[SEP]'], input ids are [101, 2028, 3944, 2014, 3129, 2513, 3449, 4383, 7682, 1999, 2010, 2192, 1037, 2312, 11255,102] and attention mask are [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]. After tokenization, each token in the input sentence is mapped to a high-dimensional vector. BERT employs a vast pre-trained vocabulary, where each token corresponds to an index. The model retrieves embeddings for each token by looking up these indexes in the vocabulary. Notably, BERT takes into account the context of each token within the sentence, meaning that the same word may have different embeddings depending on its context. This

contextual understanding is crucial for capturing the nuances and meaning of the text. The Transformer architecture of BERT comprises multiple layers of Transformer encoders. In the case of the base model, there are 12 encoder layers, while the large model consists of 24 encoder layers. These layers operate on the token embeddings in a bidirectional manner, meaning that each token is influenced by both the tokens before and after it in the sequence. This bidirectional processing allows BERT to capture context from both directions, enhancing its ability to understand the meaning of the text. This context-aware processing is a key innovation of BERT and contributes significantly to its effectiveness in natural language understanding tasks[12].

2) TF-IDF Based Vectorization: Utilizing the TF-IDF (Term Frequency-Inverse Document Frequency) approach for vectorization, we aim to represent textual data in a numerical format conducive to machine learning algorithms. Taking inspiration from our earlier example, suppose we have a corpus containing multiple documents, each comprising of sentences. TF-IDF assigns weights to words based on their frequency within individual sentences and across the entire corpus. Specifically, the term frequency (TF) component quantifies how often a term appears within a sentence, while the inverse document frequency (IDF) component diminishes the weight of terms that occur frequently across the entire corpus. Consequently, rare terms that are discriminative for specific sentences garner higher weights. To illustrate, consider this sentence "One evening her husband returned an elated bearing in his hand a large envelope" within our corpus. Terms like 'envelope' and 'elated' may occur infrequently across the entire dataset but frequently within this particular sentence, thus receiving elevated TF-IDF scores. Conversely, ubiquitous terms such as 'a' or 'in' are likely to receive lower scores due to their high frequency across the entire corpus.TF-IDF-based vectorization process, each sentence is transformed into a numerical representation, facilitating subsequent analysis and machine learning tasks.

3) Fine Tune Model by MLM: We employed the Masked Language Modeling (MLM) technique from the BERT paper to refine the model's parameters [2]. In MLM, special tokens like '[CLS]' and '[SEP]' are added to the sequence, and random tokens are masked, denoted by '[MASK]'. The model predicts the original token, enhancing its contextual understanding. We fine-tuned four models on four texts, using an 85-15 train-validation split. All models were trained for four epochs and saved with specific names. The same settings were applied to all datasets. In Table II, the hyperparameters for fine-tuning are outlined, with 'max position embedding' set to 512 and 'mlm probability' set to 15%. The remaining parameters are provided in the Table II.

4) Refining Process of the Used Model: In this stage, we reapplied pre-trained models customized for our datasets, thus kickstarting the generation of embeddings relevant to our task. We began by tokenizing sentences using established methods. These tokenized sentences are then fed into the model, which produces its output along with additional information. Our extraction process predominantly revolves around utilizing the output from hidden states to generate embeddings. We achieve this by computing the average of outputs from the last four layers, effectively capturing the essence of the sentences being analyzed. Additionally, we segmented the sentences into individual words and generated word embeddings for each word, consolidating them into a designated list. This process yields two crucial components: a collection of tokens and their associated embeddings, enabling thorough analysis and downstream applications within our task framework.

TABLE II: HYPERPARAMETER VALUES FOR MODEL

Model Name	**Hyper Parameter**
bert-base-uncased [2]	Learning rate=2e-5, weigh decay=.01, mlm probability=.15, num_hidden_layer=24 max_positional_embeddings=512, epochs=4

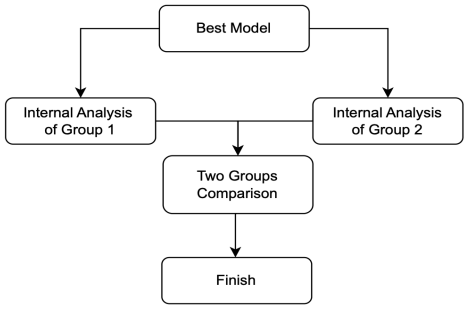

Fig. 2: Comparative Workflow

5) Calculate Distance: After obtaining embeddings and tokenization, the next step involves calculating the distance from our target keyword to all tokens to identify its neighboring elements. Cosine similarities between vectors serve as the metric for quantifying this distance. The formula for cosine similarity between two vectors A and B is given by:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Our approach entails iterating through all tokens, computing distances, sorting, and filtering them to identify the top N neighbors. To maintain the integrity of our findings, prepositions, and pronouns are deliberately excluded from consideration in this process.

IV. Experimental Outcomes

Several studies have already proved that contextual embedding works with general texts. In order to prove the contextual embedding consistency over literary texts, this paper performs a series of comparative studies between the selected texts employing selected keywords to understand the context of the words.

A. Comparative Study Between BERT and TF-IDF

In the first comparative study, we analyzed how contextual embeddings work better than the very traditional statistical method (TF-IDF) when it is used for shorter and longer literary

texts. Here, the top twenty nearest words were generated by BERT embeddings from two distinct text types: 'The Necklace' and 'Diamonds Are Forever'. The analysis of a shorter text 'The Necklace' in Table III revealed the BERT embeddings exhibited closer proximity to the word 'money' with associated terms such as 'note', 'franc', and 'thousand'. Furthermore, upon closer inspection, the highlighted words (i.e. interest, debt, poverty) appeared to align with the narrative context of this short story providing insights into the story's thematic elements. Conversely, the TF-IDF output yielded limited discernible or contextually meaningful results. Though this method produced a number of relevant words, it also produced a number of seemingly non-contextual words such as take, like, and women which confirmed the infectivity of this approach. Likewise, in the much longer narrative 'Diamonds Are Forever', BERT embeddings exposed associations with violent language and financial themes, mirroring the storyline. Notably, Table III showcases highlighted words such as 'blood', 'gun', and 'gambling', reflecting the overarching context of money within the novel. However, TF-IDF yields several irreverent words that largely fail to reveal any significant correlations with the contexts. So, this comparative study suggests that BERT embeddings provide more nuanced contextual comprehension compared to the TF-IDF method.

Consequently, our research focus shifts to comparing BERT embeddings across various narrative contexts. So for our study the best model now is BERT and we will discuss this pertinent issue according to Fig. 2.

B. Analysis of Different Groups

The second comparative study has dealt with how internal comparison produces similar and accurate results when it is analyzed through contextual embeddings. Table IV is the reflection of a comparative study between two shorter texts: 'The Necklace' and 'The Jewelry'. Upon examining the embeddings related to the word 'life', 'The Necklace' produced much accurate thematic integrity centers on depicting a burdensome and distressing life. The occurrences of 'life' within this narrative context unveil significant revelations connecting keywords like horrible, poverty, distress, etc. Likewise, the second short story, 'The Jewelry' reflects how the protagonist's life becomes full of misery due to either the early demise of the protagonist's first wife or the presence of a hot-tempered second wife. Common words from the two stories such as home, wife, woman, etc. suggest how the lives of the characters circulated crisis in the home settings, confirming the thematic consistency of the stories.

However, the comparative study between longer texts shows a contrasting result showing the thematic uniformity of each fiction. In the third comparative study, for example, contextual embedding analysis for the word 'crime' unveils a distinct theme for each fiction. The Table V shows that both novels deal with crime, law, and order; however, they have their thematic distinctions. The novel 'Diamonds Are Forever' is intricately linked with the portrayal of gangsterism, gambling, and other illicit activities, aligning closely with the overarching narrative of the novel. Conversely, in the novel 'From Russia with Love', the depiction of 'crime' diverges from external gambling motifs, which is perhaps related to the intricacies of inter-country crimes. This comparative analysis underscores the model's capability to recognize the nuanced thematic variations within novels of the same genre. By identifying distinct associations of the inner thematic elements for the keyword 'crime', this model facilitates a deeper understanding of the underlying narrative structures and thematic motifs inherent to each literary work.

The fourth comparative study illustrates the contextual embedding output between two short stories and two pieces of fiction. This inter-group comparative study employed the keyword 'diamond' to see the consistency of contextual embedding among the four literary texts; the output of the analysis is presented in Table VI. Upon analysis of the first group which consists of two much shorter short stories, the keyword 'diamond' is associated with notions of adornment and opulence, typically about jewelry and luxury items. As a result, the output from this group includes ornament-related words such as necklace, and bracelet, and the associated materials such as gold, pearl, silver, etc. Conversely, within the context of the second group which consists of two longer novels, the connotation of the word 'diamond' shifts towards the depiction of diamonds as significant assets, potentially leading to conflicts or confrontations. These observations afford us a nuanced understanding of the semantic nuances surrounding the keyword. Interestingly, the first group doesn't show any noticeable differences, with no aggressive language linked to the keyword 'diamonds.' The portrayal of diamonds here predominantly aligns with concepts of jewelry and luxury. In contrast, the second group features a significant presence of intense or aggressive vocabulary such as smuggle, casino, gambling associated with diamonds. This contrast is central to this investigation which highlighted the divergence in contextual associations across genres.

C. Get Insights from Network Graph

The fifth comparative study which is presented in a network graph shows the distinct comparison from different groups—'The Necklace' from the first group and 'Diamonds Are Forever' from the second group. Just like the same keyword of the fourth study, this visualization also took the keyword 'diamond' since its contextual significance prevails across varied narratives. This visualization aims to illustrate the connections and differences within the thematic structures of these narratives.

In figure 3, common words are depicted by connections between both nodes, while dissimilar nodes exhibit solitary connections. Upon closer examination of dissimilar nodes, it becomes evident that a clear distinction exists between the representations of 'Diamond.' One instance portrays it within the realm of ornamental adornments, whereas the other underscores its significance as a precious mineral. This visualization 3 serves to elucidate the nuanced differences in thematic portrayals across narrative boundaries, thereby

TABLE III: Comparison of BERT and TF-IDF

Story	Model	Keyword	Top 20 Neighbor	Common
The Necklace	BERT	money	money, note, **interest**, sum, hundred, much, **pay**, lend, cover, **debt**, franc, save, rest, thousand, rich, **poverty**	money, note
	TF-IDF	money	money, race, note, people, whole, miserable, make, take, like, woman, and, go, to, air, be, answer	
Diamonds Are Forever	BERT	money	stuff, cheap, buck, help, **blood**, work, **gun**, money, change, payment, $, **gambling**, grand, job, pay, clothe, man, **trouble**, dollar, million, car	money
	TF-IDF	money	money, ve, put, more, how, good, hot, get, and, make, didn, to, tell, ll, down, if	

TABLE IV: Group 1 Internal Comparison

Group	Story	Key Word	Top 20 Neighbor	Common Words
Group-1	The Necklace	life	life, work, **horrible**, world, story, home, existence, care, **poverty**, possible, **misery**, household, education, woman, **distress**, know, wife, be, last , **suffer**	home, be, wife, misery, woman, life, last
	The Jewelry	life	**die**, last, **dead, misery**, s, life, day, eye, consciousness, name, woman, person, sleep, live, career, be, **wife**, simply, heart, grief, home	

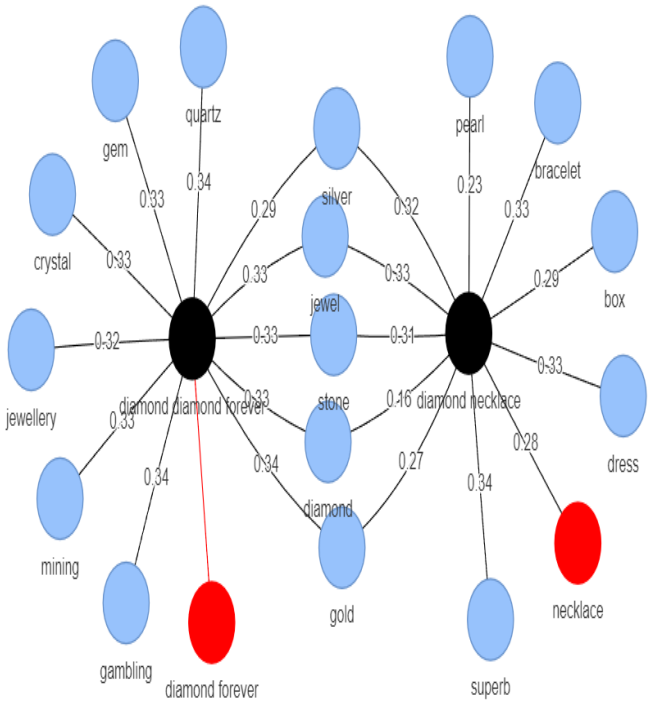

Fig. 3: The placement of the word ”diamond” in The Necklace and The Diamond Forever(top-10)

TABLE V: Group 2 Internal Comparison

Group	Story	Key Word	Top 20 Neighbor	Common Words
Group-2	Diamonds Are Forever	crime	crime, cop, **gambling**, kill, murder, **mafia**, crook, **criminal**, law, danger, **gang**, police, business, operation, thief, illegal, case, **gangster**, mob, **smuggle**, ring	case, law business, crime, police
	From Russia With Love.	crime	crime, offence, guilty, law, case, prison, **mystery, psycho**, killer, man, **officer**, bomb, police, property, decision, business, policeman, game, code, machine, **government**	

enriching our comprehension of the semantic landscape within literary works.

D. Plot embedding in a graph using PCA

The subsequent visualization depicted in Figure 4 presents an alternative perspective on nuanced findings by employing a different plotting method. Utilizing the same narrative and keyword, this technique encompasses the top 20 elements for enhanced granularity. Given these elements originated from distinct models, their weights may vary, potentially resulting in clustering phenomena. Despite this, this part aims to underscore the thematic contrast between the two texts, introducing a novel dimension of analysis. As such, this graph designated two distinct colors to highlight divergence within the plot. Leveraging principal component analysis (PCA), a 2D graph is plotted wherein each word's position reflects its distance from others. The distinct position of red and blue colored words shows a sharp thematic contrast. This approach facilitates a comprehensive examination of the semantic relationships and thematic associations between words, thereby enriching the understanding of the underlying narrative structures.

V. Conclusion

This research has yielded that contextual embedding works substantially better when it works with literary texts. Our result shows that context-based word relationships have been successfully portrayed in the tabular data and visual representations. The BERT-based embedding produced consistent results and themes across the four literary texts; these techniques are perfectly consistent in both individual and group-wise analysis. Now, the findings of this research prove that the pertinent model can capture the thematic contexts perfectly. However, this research has two major limitations. First, the models and approaches used in this research can only work or show single-word relationships, which nonetheless can not work on two or subsequent words or phrases. Second, this corpus is very small

TABLE VI: Inter Comparison Between Two Groups

Group	Story	Keyword	Top 20 Neighbor	Common Words
Group-1	The Necklace	diamond	diamond, pearl, gold, stone, silver, **jewel**, bracelet, superb, shine, wear, new, **necklace**, collar, satin, box, cross, dress, buy, ball, one, rich	jewel, necklace, diamond, stone, pearl, gold, bracelet
	The Jewelry	diamond	diamond, crystal, emerald, sapphire, **earring**, pearl, pebble, gold, **jewelry**, precious, **bracelet, ring**, stone, piece, big, jewel, adorn, **necklace**, beauty, set, value	
Group-2	Diamonds Are Forever	diamond	**mining**, chocolate, diamond, gem, sterling, kimberley, gold, **opium**, golden, pearl, **gambling**, stone, **smuggle**, jewel, silver, jewellery, crystal, **casino**, quartz, turquoise, vegas	jewel, stone, casino, golden, diamond, gold, silver, chocolate, smuggle
	From Russia With Love.	diamond	treasure, jewel, **smuggle**, oil, diamond, coin, **casino**, gift, glitter, stone, **sword**, chocolate, golden, **knife**, price, spice, silver, ring, **money**, gold, tie	

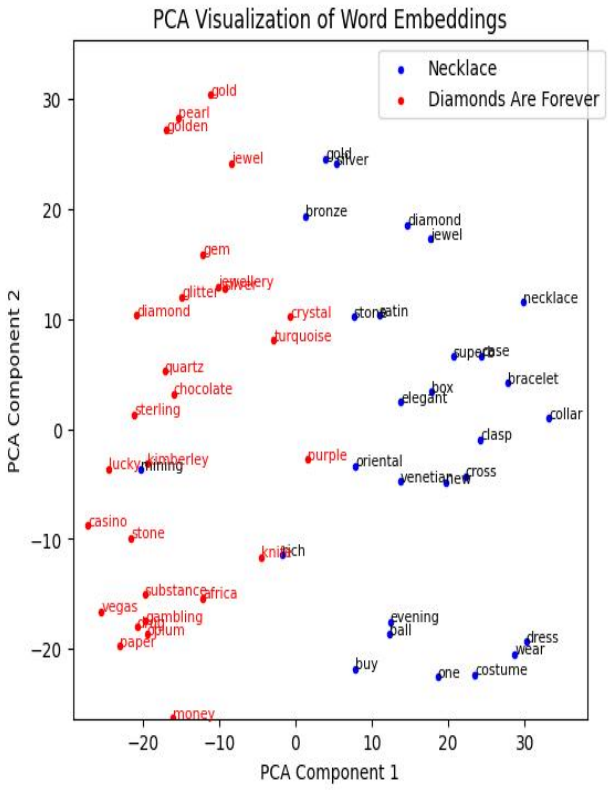

Fig. 4: The placement of the word "diamond" in Necklace and The Diamond Forever(top-20)

in nature and is not a representative corpus by any means. It is also true that these limitations create a newer scope for further research for both contextual embedding and literary text analysis.

REFERENCES

[1] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] S. Arora, A. May, J. Zhang, and C. Ré, "Contextual embeddings: When are they worth it?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 2650–2663, Association for Computational Linguistics, July 2020.

[4] I. Yunianto, A. E. Permanasari, and W. Widyawan, "Domain-specific contextualized embedding: A systematic literature review," in *2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE)*, IEEE, Oct. 2020.

[5] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *Journal of the American Medical Informatics Association*, vol. 26, p. 1297–1304, July 2019.

[6] S. Nair, M. Srinivasan, and S. Meylan, "Contextualized word embeddings encode aspects of human-like word sense knowledge," in *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon* (M. Zock, E. Chersoni, A. Lenci, and E. Santus, eds.), (Online), pp. 129–141, Association for Computational Linguistics, Dec. 2020.

[7] J. Klafka and A. Ettinger, "Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4801–4811, Association for Computational Linguistics, July 2020.

[8] C. Zhang and B. Wu, "Characterizing gender stereotypes in popular fiction: A machine learning approach," *Online Journal of Communication and Media Technologies*, vol. 13, p. e202349, Oct. 2023.

[9] S. Grayson, M. Mulvany, K. Wade, G. Meaney, and D. Greene, *Exploring the Role of Gender in 19th Century Fiction Through the Lens of Word Embeddings*, p. 358–364. Springer International Publishing, 2017.

[10] V. Kanjirangat, S. Mellace, and A. Antonucci, "Temporal embeddings and transformer models for narrative text understanding," *CoRR*, vol. abs/2003.08811, 2020.

[11] X. Shen, "Sentiment analysis of modern chinese literature based on deep learning," *Journal of Electrical Systems*, vol. 20, p. 1565–1574, Apr. 2024.

[12] J. Turton, R. E. Smith, and D. Vinson, "Deriving contextualised semantic features from BERT (and other transformer model) embeddings," in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)* (A. Rogers, I. Calixto, I. Vulić, N. Saphra, N. Kassner, O.-M. Camburu, T. Bansal, and V. Shwartz, eds.), (Online), pp. 248–262, Association for Computational Linguistics, Aug. 2021.