Deep Learning-based Sentence Embeddings using BERT for Textual Entailment

Mohammed Alsuhaibani
Department of Computer Science, College of Computer,
Qassim University, Buraydah 52571, Saudi Arabia

***Abstract*—This study directly and thoroughly investigates the practicalities of utilizing sentence embeddings, derived from the foundations of deep learning, for textual entailment recognition, with a specific emphasis on the robust BERT model. As a cornerstone of our research, we incorporated the Stanford Natural Language Inference (SNLI) dataset. Our study emphasizes a meticulous analysis of BERT's variable layers to ascertain the optimal layer for generating sentence embeddings that can effectively identify entailment. Our approach deviates from traditional methodologies, as we base our evaluation of entailment on the direct and simple comparison of sentence norms, subsequently highlighting the geometrical attributes of the embeddings. Experimental results revealed that the $L2$ norm of sentence embeddings, drawn specifically from BERT's 7th layer, emerged superior in entailment detection compared to other setups.**

***Keywords*—Textual entailment; deep learning; entailment detection; BERT; text processing; natural language processing systems**

I. Introduction

Textual entailment (TE), an essential notion within natural language processing (NLP), is expressed as a binary correlation between two segments of text [1]. Text T is stated to entail another text H if the comprehension gathered from T would compel a reader to deduce that H is most probable [2]. For example, the sentence "The dog is playing in the park" entails that "There is a dog at the park". This unfolds as a unidirectional correlation, where TE serves as a fundamental pillar within NLP, supporting numerous applications in various disciplines.

TE's multifaceted applications extend across diverse tasks, including question answering (QA) [3], where the precise extraction of responses from intricate texts hinges significantly on the correct discernment of entailment. It also impacts the effectiveness of information retrieval (IR) [4] tasks and the success of information extraction processes. TE is also an essential ingredient in the creation of text summarization [5, 6] mechanisms. The vast reach of these applications accentuates the crucial nature of textual entailment and the importance of its accurate identification.

Nonetheless, TE introduces a notable challenge, especially in terms of understanding the semantic relationships between sentences [7, 8, 9]. To tackle this, sentence embeddings have garnered significant attention lately. At their core, sentence embeddings are condensed vector depictions of sentences created to encode their semantic meanings within a fixed-dimensional vector [10]. The deployment of sentence embeddings enables swift and effective comparison and assessment of different sentences, acting as an important instrument in a range of NLP tasks, including TE.

In the domain of sentence embeddings generation, deep learning has led the advancements. The hierarchical learning aptitudes of deep learning models enable them to produce semantically rich sentence embeddings, encompassing the intricate syntactic and semantic attributes of sentences. Notably, these models have demonstrated remarkable proficiency in discerning nuanced relationships, like entailment, among sentences [11, 12].

In recent advancements of deep learning for NLP, Transformer-based models, with particular emphasis on BERT (Bidirectional Encoder Representations from Transformers) [13], have signified noteworthy progress. The ability of BERT to consider the complete context of a sentence bi-directionally (left and right) permits the creation of superior-quality sentence embeddings. This unique capability has earned BERT widespread recognition and usage in the NLP community, particularly for tasks such as TE [14, 15, 16].

The assessment of various methods and models in TE rests on numerous specific datasets. The Stanford Natural Language Inference (SNLI) [1] dataset is one such resource, offering a large collection of sentence pairs annotated for entailment, contradiction, and neutrality. Resources like SNLI enable consistent and comparable evaluation of different TE techniques, encouraging advancement in the field.

Despite the remarkable progress in TE, current methods, especially those founded on deep learning, still exhibit shortcomings. These include an intense dependence on complex architectural designs and extensive computational resources. In addition, a majority of these models primarily concentrate on the syntactic features of sentences, frequently neglecting the geometric attributes of sentence embeddings.

To address these issues, our study delves into the detailed examination of the use of sentence embeddings for TE. Utilizing, directly, the strength of the BERT model, we scrutinize the effects of employing varying layers for the extraction of sentence embeddings. Our study departs from traditional methods by assessing entailment through the comparison of sentence norms, thereby focusing on the geometric characteristics of the embeddings, a less explored yet potentially beneficial aspect.

Our hands-on findings underline the good performance of the L2 norm of sentence embeddings, specifically those extracted from the 7th layer of BERT. These findings offer a fresh perspective on the TE. Our results particularly emphasise the importance of layer selection in the extraction of sentence

embeddings as well as the consideration of the geometric properties of sentence embeddings in addressing TE.

The remainder of this paper unfolds as follows. We will first dive into the related work in Section II, where we discuss the key literature on textual entailment and sentence embeddings. In Section III we will share our proposed method which utilizes the BERT model. Next, in Section IV, we will discuss SNLI dataset that we used for our experiments. We then move to the experiments and results in Section V, where we lay out the outcomes of the experiments and interpret our results and speak on any limitations we have come across. And lastly, in the conclusion, Section VI, we will bring everything together by summarizing our findings, reaffirming what our study brings to the field, and pondering over potential areas for future research.

II. Background and Related Work

Textual Entailment (TE), also known as Natural Language Inference (NLI), entails determining the relationship between two sentences, specifically, if one sentence (the hypothesis) implies, contradicts, or remains neutral to the other (the premise) [17]. This is a demanding task as it necessitates understanding the essence of both sentences and their interplay.

One method to accomplish TE employs sentence embeddings, which are vector representations encapsulating the semantic significance of sentences [18]. These embeddings can be used to train a model to anticipate the relationship dynamics between a pair of sentences.

There exists a plethora of techniques to generate sentence embeddings. A prevalent approach involves deploying a word embedding model to create word embeddings [19, 20, 11], which are then amalgamated to craft a sentence embedding. An alternative strategy employs a deep learning model specifically trained for generating sentence embeddings [21, 22].

BERT [13] has gained popularity as a deep learning model for sentence embeddings. As a transformer-based model, BERT is pre-trained on an extensive corpus of text, enabling it to effectively learn and represent word and sentence meanings. This capability is useful for a wide spectrum of NLP tasks, including TE. There has been a growing body of research on using BERT for TE. In fact, when Devlin et al. introduced BERT itself, it was trained using next-word prediction and missing-word prediction, allowing it to acquire meaningful word and sentence representations and has proven useful for several NLP tasks, including TE.

Moreover, Lin and Su [15] examine BERT's proficiency in handling TE tasks, particularly its capability to bypass any latent biases in the dataset. To simplify the investigation, they design a straightforward entailment judgment scenario using only binary predicates in clear English. The results suggest that BERT's learning curve is somewhat slower than expected. However, they found that incorporating task-specific features significantly improved the learning efficiency, leading to a data reduction by a factor of 1,500. This key discovery highlights the importance of domain knowledge in effectively utilizing neural networks for TE tasks.

Similarly, Gajbhiye et al. [23] introduce a new model for TE, dubbed External Knowledge Enhanced BERT (ExBERT). It improves BERT's language understanding and reasoning capabilities by integrating commonsense knowledge from external sources into the existing contextual representation. The model uses BERT-derived contextual word representations to pull and encode relevant knowledge from knowledge graphs. It's designed to seamlessly blend this external knowledge into the reasoning process.

Pang et al. [24] have developed a method for integrating syntax into TE models. Their approach uses contextual token-level vector representations derived from a pre-trained dependency parser. This technique, similar to other contextual embedders, can be applied to a wide range of neural models. They tested this method with some established TE models, such as BERT. The findings showed an increase in accuracy across the benchmark datasets.

Cabezudo et al. [25] investigate various methods to enhance inference recognition in the ASSIN [26] dataset, a dataset specifically designed for entailment recognition in Portuguese. They also study the effects of adding external data, such as multilingual data or an automatically translated corpus, to improve model training. They use the multilingual pre-trained BERT model in their experiment and their results show an improvement in the ASSIN. Interestingly, their findings suggest that using external data does not significantly improve the performance of the model.

Wehnert et al. [27] have introduced three distinct methods for the classification of entailment. The first approach harmonizes Sentence-BERT embeddings with a graph neural network, while the second strategy leans on the specific LEGAL-BERT model, which undergoes additional training on the competition's retrieval task and is fine-tuned specifically for entailment classification. Their third method ingeniously employs the KERMIT encoder to embed syntactic parse trees and integrates this with a BERT model. Their study delves into the potential of this third tactic and provides insights into why the LEGAL-BERT submissions, among all entries, might have managed to edge out the graph-based method in performance.

Shajalal et al. [28] develop a new method for identifying the textual entailment relationship between a text and its hypothesis. They introduce a new semantic feature that uses empirical threshold-based semantic text representation. This approach makes use of an element-wise Manhattan distance vector-based feature, designed to understand the semantic entailment relationship within a text-hypothesis pair. They tested their method using several experiments on the benchmark entailment classification dataset, SICK-RTE [29], with a variety of machine learning algorithms. Their empirical sentence representation technique improved the semantic understanding of the texts and hypotheses.

Jiang and de Marneffe [30] have taken on the task of addressing an issue prevalent in TE datasets. They have come up with a strategy, redefining the use of the CommitmentBank for TE. Their idea is to adjust the emphasis on how committed a speaker is to the complements of clause-embedding verbs in a range of contexts that cancel entailment. This move leads to the creation of hypotheses that are free from artefacts and naturally intertwined with the premises. Even though their fresh approach lets a BERT-based model hit a good result with BERT, they stated that the model is not yet fully grasping

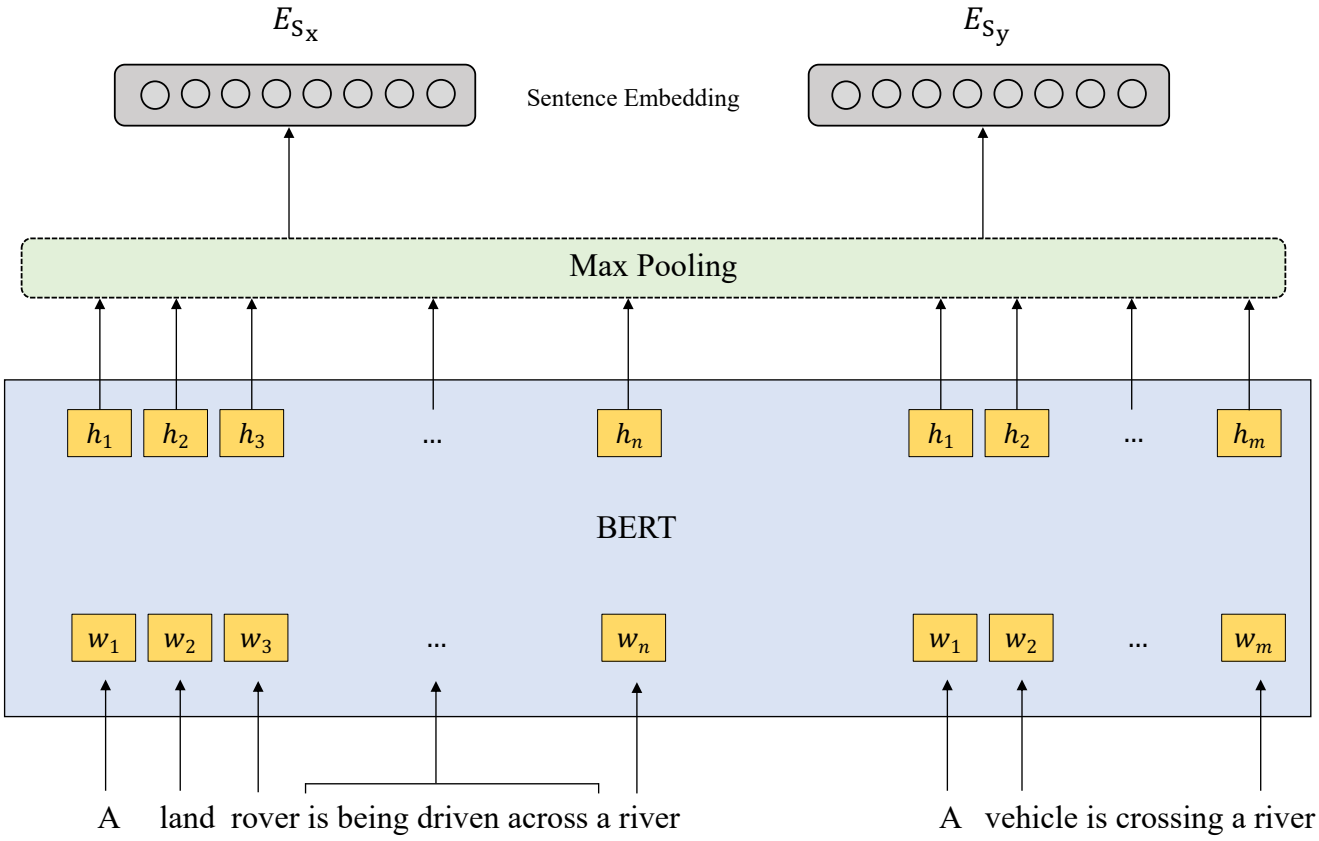

Fig. 1. Extraction of sentence embeddings from BERT with a *max* pooling strategy from the token-level embeddings [13].

the nuances of pragmatic reasoning and certain linguistic generalizations.

While the above-mentioned approaches significantly advanced TE, its reliance on intricate designs and significant computational power is notable. To rectify this, our research investigates the application of sentence embeddings in TE. We simply and directly utilize the BERT model's potential, exploring the effects of various layers for sentence embeddings extraction. In contrast to conventional approaches, we utilize a simple and straightforward approach to evaluate entailment by comparing sentence norms, spotlighting the geometric aspects of embeddings, a relatively uncharted but potentially advantageous area.

III. Proposed Approach

Our approach to TE revolves around using BERT to extract sentence embeddings. While loading pre-trained BERT and tokenizer, we set configurations for sub-token pooling, which determines the token piece embeddings used in constructing the token embedding. Options include using the *first* subtoken, the *last* subtoken, both the *first* and *last*, or an *average* overall (mean). Additionally, we specify the layer (layers 1 to 12) from which the embeddings should be extracted. Specifically, as shown in Fig. 1, we generate sentence embeddings for each pair of sentences in the dataset. We will feed the premise and hypothesis into BERT and extract the output of the $[CLS]$ special token, which is a fixed-length representation of the entire input sequence. This will provide us with a pair of sentence embeddings that capture the semantic and syntactic information of the premise and hypothesis.

Given a pair of sentences (x, y) with $x = w_1, ..., w_n$ and $y = w_1, ..., w_m$ forming a tuple, we use the loaded pre-trained BERT model to encode each sentence individually. We employed two possible strategies: default document embeddings and token-based document embeddings.

In default document embeddings, we derive one vector representing the entire sentence as $E_{S_i} = TN(i)$, where $i \in x, y$ and TN denotes a Transformer-based network, BERT. Basically, it extracts one feature as the sentence embedding using a default pooling strategy that simply selects the first token feature $[CLS]$from the standard word-piece tokenzation as proposed in BERT. On the other side, in the token-based document embeddings (Fig. 1), we extract a vector corresponding to each token in a sentence, for example, $S_x = (E_1, ..., E_n)$, where $E_i = TN(w_i) \in \mathbb{R}^D$ (D is the embedding size). To generate a sentence vector, we then compute either a *min*, *max* or *mean* pool across all these token vectors.

$$E_{S_x} = \frac{1}{n} \sum_{i}^{n} E_i \tag{1}$$

When using the *mean*, we calculate an average across all vectors to derive a sentence vector. The sentence embedding of x, E_{S_x}, is calculated using (1), and E_{S_y} for y is computed

similarly. Besides the *mean*, we also test with other pooling strategies like *min* and *max*. *Min* involves sorting the token vectors based on their norm magnitude and using the vector with the least magnitude as the sentence vector. Conversely, *max* employs the vector with the largest norm magnitude as the sentence vector. We will use the *max* pooling in our experiments which empirically gives the best performance as we will detail in Section V.

Lastly, we predict entailment by comparing the norms of the pair of sentences in our input tuple. If the norm of x is greater than or equal to the norm of y, we consider it as entailment; otherwise, it is not (as shown in (2) and (3)). This approach provides a direct and effective way to determine TE.

$$V = ||E_{S_x}||_2 \geqslant ||E_{S_y}||_2 \tag{2}$$

$$Entailment = \begin{cases} True & \text{if } V(x, y), \\ False & \text{otherwise.} \end{cases} \tag{3}$$

IV. Data

As a main dataset, we have leveraged the Stanford Natural Language Inference (SNLI) [1] dataset, a comprehensive collection of sentence pairs instrumental in training TE models. The SNLI is a robust dataset of approximately 570,000 human-authored English sentence pairs, each meticulously annotated to ensure balanced classification across three categories: entailment, contradiction, and neutral. Its wide acceptance and usage for training and testing models in TE have earned it the reputation of a standard benchmark within the field. It is worth noting that the creation of this dataset involved a crowdsourcing approach. Meaning human contributors generated the sentence pairs and assigned the entailment categories. This human involvement ensures the quality and reliability of the data.

SNLI dataset has been a pivotal element in the evolution of many contemporary NLP models, including transformative models like BERT and their subsequent iterations.

Table I features select examples from the SNLI dataset used in our approach. For ease of comprehension, we've adopted a color-coding scheme: instances of entailment are presented in green-shaded rows, neutral examples have been uncolored, while contradiction cases appear in rows shaded red. This approach to color differentiation offers an intuitive visualization of the varied sentence pairs that the SNLI dataset encompasses.

V. Experiments and Results

A. Experimental Settings

In this section, we provide an outline of the steps we have followed to execute our experiments, covering the specific details of loading data, data preprocessing, and the application of pre-trained models and tokenizers.

Our experimental framework incorporates the use of the Hugging Face API[1] for the purpose of loading BERT pre-trained model and tokenizers. As part of our configuration parameters, we have included a setting for sub-token pooling. This setting dictates the manner in which token piece embeddings are utilized to form the final token embedding.

The data loading process involves drawing sentences from one of two file formats: Excel $(.xlsx)$ or JavaScript Object Notation $(.json)$. Furthermore, we have prepared an alternate method to load data, using the Hugging Face dataset loader object as a substitute for traditional content loading from text or $json$ files.

In the data preprocessing step, we apply a series of operations to refine and structure the data. Initially, we clean each sentence pair in the dataset by eliminating superfluous spaces found at the sentence boundaries. Following this, we organize the cleaned pairs of sentences into tuples, i.e., a sentence pair $(sentence1, sentence2)$, culminating in a list of such tuples. This process ensures that our data is well-organized and conducive to subsequent tasks.

With the aid of the Hugging Face API, we have streamlined the process of loading BERT pre-trained weights for a variety of PyTorch[2] and TensorFlow[3] models. This step is critical in harnessing the capabilities of BERT pre-trained model, which has already acquired useful representations from extensive text corpora, to kickstart our task-specific model.

Subsequent to extracting a vector that corresponds to each token in a sentence, we carry out additional processing on these token vectors to derive a unified sentence vector. As highlighted in Section III, this is achieved by implementing one of the multiple pooling strategies, *min*, *max* or *mean* across all token vectors.

Our initial experimentation revealed that the *max* pooling strategy surpassed the performance offered by the *min* and *mean* strategies. Hence, we chose to incorporate the *max* pooling strategy in all subsequent experiments for generating sentence vectors from token vectors. This choice proved pivotal in boosting the effectiveness of our entailment detection procedure.

Alongside our selected pooling strategy, we also examined the effect of different layers within the BERT model on our results. We extracted embeddings from a range of layers within BERT, extending from layer 1 to layer 12, and studied their influence on the task of TE. This experiment offers insight into the role each layer has in shaping the quality of sentence embeddings. This expansive exploration across all layers of the BERT model enables us to pinpoint the optimal layer for our specific task, a factor in boosting the efficacy of our entailment detection procedure.

In an extension to our experimental setup, we investigated the impact of various norms, $L1$, $L2$, and L-inf on the entailment detection. As norms play a vital role in comparing sentence embeddings in our methodology, experimenting with different norms helped us identify which norm leads to the most precise and reliable entailment predictions. The outcomes of these investigations are reported in our study, shedding light on the influence of each norm on the performance of our entailment detection approach.

[1] https://huggingface.co/models

[2] https://pytorch.org/

[3] https://www.tensorflow.org/

TABLE I. RANDOMLY CHOSEN SAMPLES FROM THE SNLI DATASET USED IN THE PROPOSED APPROACH, COLOR-CODED BY ENTAILMENT CATEGORY

Text	Judgments	Hypothesis
A middle-aged man in a gray t-shirt and brown pants sitting on his bed reading a flyer-like paper.	entailment E E E E E	A man is sitting on his bed reading.
A young boy and girl playing baseball in a grassy field.	entailment N E E E E	Kids play baseball.
Numerous people sitting in a dim lit room talking, drinking coffee and using computers.	entailment E E E E E	People are in a dimly lit room drinking coffee.
A white race dog wearing the number eight runs on the track.	entailment E E E E E	A dog is running.
A woman reaching for candy bars that are on a shelf.	neutral N N E N C	The candy bars are above the womans head.
The boy wearing the blue hooded top is holding a baby goat in his arms.	neutral N C N N N	The goat jumped into the boys arms.
A little girl is sitting on a bench in a park.	neutral N N N N N	The little girl is having fun.
A small child playing in a dusty square.	neutral E N N N N	A child is playing with a doll.
Multiple people starting to pack their parachutes after a successful skydive.	contradiction C C C C C	cat chased by tiger.
A swimming dog with a small branch in its mouth.	contradiction C C C C N	A dog is ice skating.
A man with a mustache is playing ice hockey with snow in the background.	contradiction C C C C C	People are swimming in the lake.
A busy street full of shops and people holding hands and walking.	contradiction C C C C C	People sitting in a restaurant.

In Section III, we laid out our strategy for evaluating the proposed method, which, despite its apparent simplicity, yields potent results. The heart of our approach to entailment prediction lies in comparing the norms of the sentence pairs that make up our input tuple. If the norm of x equals or surpasses that of y, we mark it as an entailment instance. In contrast, if it fails to meet this criterion, we label it as non-entailment (refer to Equations (2) and (3) for further clarity). When it comes to gauging performance, we turn to the accuracy metric. This indicator gives us the ratio of successful classifications. By resorting to this measure, we can quantify how adept our model is at correctly categorizing sentence pairs in alignment with their actual entailment status. This simple yet effective measure offers a clear insight into our proposed approach's efficiency in entailment prediction.

B. Results and Discussion

The results reflected in Table II offer a thorough perspective of the outcomes generated through our proposed approach. We have incorporated accuracy percentages that depict the repercussions of diversifying two primary parameters: the BERT model's layers (from 1 to 12) and the types of norms ($L1$, $L2$, and L-inf). Regardless of these alterations, the max pooling strategy remained a constant, thereby offering a consistent benchmark for comparison.

Our findings lead us to two insights. The first is related to the choice of norm type; the $L2$ norm systematically outpaced both $L1$ and L-inf norms regardless of the layer, and $L1$ come second. Whereas, L-inf performs poorly across the layers.

Our second insight arises from the analysis of BERT model's layers. As per the empirical findings, it appears that the 7th layer offers an optimal environment for the extraction of embeddings with as high accuracy as $\%91$. This is important as it aids us in pinpointing the most suitable layer, thereby optimizing the sentence embedding generation process.

To simplify the understanding of the results and make them visually discernible, we have plotted the model's performance. For this, in Fig. 2, we considered the $L2$ norm (proven to offer superior results) and plotted its influence on the

TABLE II. PERFORMANCE ACCURACY OF THE PROPOSED APPROACH WITH BERT LAYER VARIATION AND NORM TYPES WITH MAX POOLING STRATEGY. BOLD INDICATES THE BEST PERFORMANCE FOR EACH NORM

Norm	**Layers**											
	1	2	3	4	5	6	7	8	9	10	11	12
$L2$	0.75	0.83	0.83	0.84	0.82	0.83	**0.91**	0.87	0.77	0.77	0.76	0.83
$L1$	0.73	0.81	0.81	0.81	0.77	0.74	0.80	**0.83**	0.65	0.60	0.57	0.59
L-inf	0.26	0.22	0.17	0.16	0.19	0.24	0.33	0.22	0.47	0.41	0.32	**0.49**

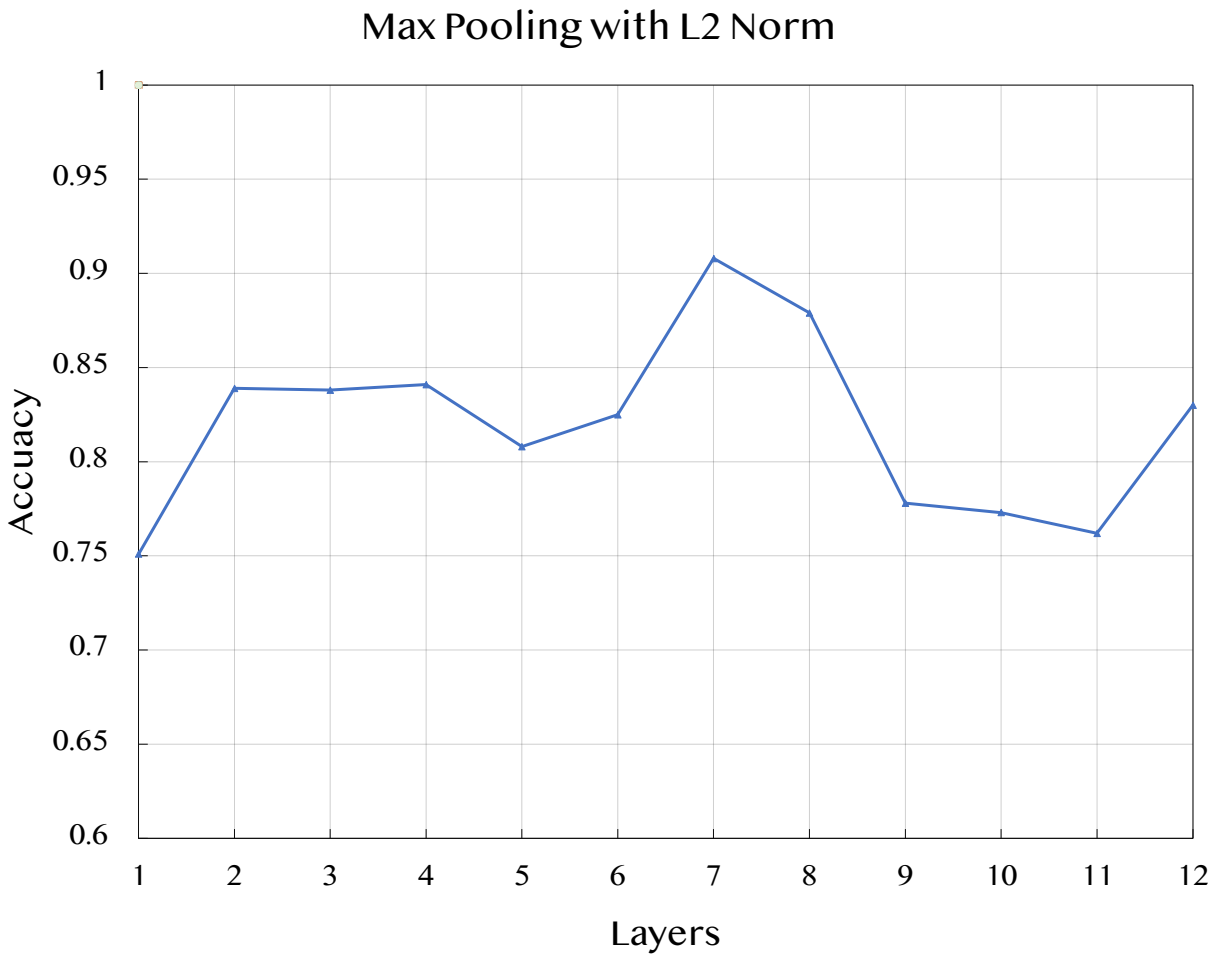

Fig. 2. Proposed approach performance with $L2$ norm and various layers.

varying layers, with the latter serving as the x-axis. Despite the changes in layers, we ensured the max pooling strategy remained unchanged, facilitating a focused study on the layers' influence. The resulting graph offers a straightforward visual comparison of the performance impact due to different layers.

VI. CONCLUSION

In this study, we have delved TE, using the expansive SNLI dataset as our sandbox. Our approach lies in leveraging the strength of pre-existing models, with an emphasis on the BERT model. Our methodology consists of extracting token embeddings and transforming them into sentence vectors. In our quest to streamline these vectors, we experimented with several pooling strategies, *min*, *max*, and *mean*. Our observations consistently pointed towards the *max* pooling strategy as the most effective. We focused on the implications of various layers within the BERT model on the task of entailment detection. Our experiments revealed that the seventh layer of the model stood out as the most impactful for generating potent embeddings for this task.

Norms, too, were given considerable attention in our experimental setup. We tested different norms, namely $L1$, $L2$, and L-inf. Our findings tipped the scales in favor of the $L2$ norm, emphasizing the influential role norms play in determining the quality of entailment detection.

To sum it up, our research presents a direct and simple approach for effective entailment detection by utilizing BERT. It underscores the importance of which layers to select for extracting embeddings, the pooling strategies to implement, and the norms to use. Future exploration could include testing our approach on other pre-trained models and entailment datasets to enhance its generalizability.

REFERENCES

[1] S. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 632–642.

[2] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced lstm for natural language inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017.

[3] X. Wang, P. Kapanipathi, R. Musa, M. Yu, K. Talamadupula, I. Abdelaziz, M. Chang, A. Fokoue, B. Makni, N. Mattei *et al.*, "Improving natural language inference using external knowledge in the science questions domain," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7208–7215.

[4] K. Zhou, Q. Qiao, Y. Li, and Q. Li, "Improving distantly supervised relation extraction by natural language inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 14 047–14 055.

[5] H. Song, W.-N. Zhang, J. Hu, and T. Liu, "Generating persona consistent dialogues by exploiting natural language inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8878–8885.

[6] H. Chouikhi, M. Alsuhaibani, and F. Jarray, "Bert-based joint model for aspect term extraction and aspect polarity detection in arabic text," *Electronics*, vol. 12, no. 3, p. 515, 2023.

[7] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of bert and albert sentence embedding performance on downstream nlp tasks," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 5482–5487.

[8] Z. Chen, Q. Gao, and L. S. Moss, "Neurallog: Natural language inference with joint neural and logical reasoning," in *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, 2021, pp. 78–88.

[9] A. Talman, A. Yli-Jyrä, and J. Tiedemann, "Sentence embeddings in nli with iterative refinement encoders," *Natural Language Engineering*, vol. 25, no. 4, pp. 467–482, 2019.

[10] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[11] J. Maillard, S. Clark, and D. Yogatama, "Jointly learning sentence embeddings and syntax with unsupervised tree-lstms," *Natural Language Engineering*, vol. 25, no. 4, pp. 433–449, 2019.

[12] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL), 2021, pp. 6894–6910.

[13] 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.

[14] Y. Song, J. Wang, Z. Liang, Z. Liu, and T. Jiang, "Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference," *arXiv preprint arXiv:2002.04815*, 2020.

[15] Y.-C. Lin and K.-Y. Su, "How fast can bert learn simple natural language inference?" in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 626–633.

[16] Q. He, H. Wang, and Y. Zhang, "Enhancing generalization in natural language inference by syntax," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4973–4978.

[17] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research*, vol. 38, pp. 135–187, 2010.

[18] J. Yu and J. Jiang, "Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 236–246.

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] M. Alsuhaibani, D. Bollegala, T. Maehara, and K.-i. Kawarabayashi, "Jointly learning word embeddings using a corpus and a knowledge base," *PloS one*, vol. 13, no. 3, p. e0193094, 2018.

[21] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

[22] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *5th International Conference on Learning Representations, ICLR 2017*, 2019.

[23] A. Gajbhiye, N. A. Moubayed, and S. Bradley, "Exbert: An external knowledge enhanced bert for natural language inference," in *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*. Springer, 2021, pp. 460–472.

[24] D. Pang, L. H. Lin, and N. A. Smith, "Improving nat-

ural language inference with a pretrained parser," *arXiv preprint arXiv:1909.08217*, 2019.

[25] M. A. S. Cabezudo, M. Inácio, A. C. Rodrigues, E. Casanova, and R. F. de Sousa, "Natural language inference for portuguese using bert and multilingual information," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2020, pp. 346–356.

[26] E. Fonseca, L. Santos, M. Criscuolo, and S. Aluisio, "Assin: Avaliacao de similaridade semantica e inferencia textual," in *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, 2016, pp. 13–15.

[27] S. Wehnert, S. Dureja, L. Kutty, V. Sudhi, and E. W. De Luca, "Applying bert embeddings to predict legal textual entailment," *The Review of Socionetwork Strategies*, vol. 16, no. 1, pp. 197–219, 2022.

[28] M. Shajalal, M. Atabuzzaman, M. B. Baby, M. R. Karim, and A. Boden, "Textual entailment recognition with semantic features from empirical text representation," in *International Conference on Speech and Language Technologies for Low-resource Languages*. Springer, 2022, pp. 183–195.

[29] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 2014, pp. 1–8.

[30] N. Jiang and M.-C. de Marneffe, "Evaluating bert for natural language inference: A case study on the commitmentbank," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 6086–6091.