Project#3: Semantic Textual Relatedness

Rajarshi Dutta[1], **Udvas Basak**[2], **Shivam Pandey**[3]
[1]200762, [2]201056, [3]200938
[1]MSE, [2]AE, [3]ME
rajarshi20@iitk.ac.in, udvasb20@iitk.ac.in, shivamp20@iitk.ac.in

Semantic Textual Relatedness is important for deciphering meaning and its applications lie on various NLP tasks like **Information Retrieval**, **Question Answering**, **Text Summarization**, **Machine Translation**, **Paraphrase Detection** etc. The challenge is mainly focussed on automatically detecting the degree of relatedness between pairs of sentences, for 14 languages, which include high-resource as well as low-resource languages, specially Asian and African languages. The major challenge lies in the efficient development of a metric to facilitate the calculation of the **relatedness** score between the sentence pairs, and harnessing the structure of multiple languages to create an efficient model. The data is trained on pairs of sentences with manually determined relatedness scores between 0(completely unrelated), and 1(maximally related).

All files relating to our solution can be found at: https://github.com/Rajarshi1001/CS779AProject. list

1 Introduction

Semantic relatedness measures the proximity in meaning between units of language, such as words, sentences, phrases, or n-grams (Mohammad, 2008). Semantically similar sentences exhibit paraphrasal or entailment relations, while relatedness encompasses broader commonalities, including topic alignment, viewpoint, temporal context, elaboration, and more (e.g., two sentences about the same event). Most NLP work focuses on Semantic Textual Similarity in English, emphasizing the importance of the scoring metric. Relatedness algorithms like Normalized Google Distance (Lopes and Moura, 2019), which do not rely on transformers, show promise in calculating relatedness scores.
This paper documents our efforts in addressing this problem, particularly in combining relatedness scores intelligently to achieve high accuracy.

2 Problem Definition

The task presented in SemEval 2024 has three tracks. The tasks are presented verbatim as follows:

Supervised: Participants are to submit systems that have been **trained using the labeled training datasets provided**. Participating teams are allowed to use any publicly available datasets (e.g., other relatedness and similarity datasets or datasets in any other languages)

Unsupervised: Participants are to submit systems that have been developed without the use of any labeled datasets pertaining to semantic relatedness or semantic similarity between units of text more than two words long in any language. **The use of unigram or bigram relatedness datasets (from any language) is permitted**.

Cross-Lingual: Participants are to submit systems that have been developed without the use of any labeled semantic similarity or semantic relatedness datasets in the target language and with **the use of labeled dataset(s) from at least one other language**.

3 Related Work

Key points:

1. (Reimers and Gurevych, 2019) Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation - Nils Reimers, Iryna Gurevych Ubiquitous
   This study introduces a novel approach to extend monolingual sentence embedding models to new languages. It uses multilingual knowledge distillation, where a student model, $\hat{M}$, learns from a

teacher model, M, on source language sentences (s_i) and their translations (t_i) by minimizing the mean-squared loss function.
$\text{Loss} = \frac{1}{|B|} \sum_j \left(M(s_j) - \hat{M}(s_j)\right)^2 + \left(M(s_j) - \hat{M}(t_j)\right)^2$
This method significantly boosts retrieval performance across languages, especially for low-resource languages, surpassing LASER, mUSE, and LaBSE in multilingual semantic textual similarity (STS) tasks. It successfully aligns vector spaces across languages, addressing limitations in LASER and LaBSE, but also reveals a "curse of ity," where adding more languages can harm model performance.
Future research may explore additional languages and domains to assess generalizability. Mitigating language bias in multilingual models is another promising direction.

2. What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study - Mohamed Abdalla, Krishnapriya Vishnubhotla, Saif M. Mohammad
(Abdalla et al., 2023) This study investigates factors influencing semantic relatedness between English sentence pairs, introducing the STR-2022 dataset annotated with diverse domain-relatedness scores. The authors employ a comparative annotation approach, ensuring high reliability (Spearman correlation: 0.84) in human judgments. Four key factors affecting sentence relatedness are considered: Lexical Overlap (Q1), Related Words (Q2), Related Words of the Same Part of Speech (Q3), and Related Subjects and Objects (Q4). Notably, related nouns, including proper nouns, and related objects are influential.
In empirical evaluation, unsupervised and supervised sentence representation models are explored. Unsupervised models offer marginal improvements over a lexical overlap baseline. Supervised models, especially fine-tuned BERT-based ones, excel in capturing semantic relatedness. BERT-base (mean) and RoBERTa-base (mean) achieve high Spearman correlations of 0.82 and 0.83, emphasizing the effectiveness of supervised approaches.
Future research may investigate the impact of additional linguistic features on relatedness and explore novel embedding tech

3. The Google Similarity Distance - Rudi L. Cilibrasi, Paul M.B. Vitanyi
(Lopes and Moura, 2019) Cilibrasi and Vitanyi introduced the concept of Normalized Google Distance (NGD) as a novel metric to measure semantic similarity between words or concepts. The NGD is calculated using Google search result counts, enabling a data-driven approach to quantify the relatedness of terms. Their formula for NGD is given by:
$\text{NGD}(w_1, w_2) = \frac{\max(\log(N(w_1)), \log(N(w_2))) - \log(N(w_1, w_2))}{\log(N) - \min(\log(N(w_1)), \log(N(w_2)))}$
In this formula, N represents the total number of web pages indexed by Google, N($w1$) and N($w2$) are the counts of pages containing words $w1$ and $w2$, and N($w1, w2$) is the count of pages containing both w1 and w2. The NGD formula normalizes these counts by considering the overall corpus size and the co-occurrence of terms in web pages.
Their research confirmed NGD's effectiveness in word sense disambiguation, concept classification, and natural language translation, with consistently high accuracy. Comparing NGD-based semantic categories with WordNet showed strong agreement, highlighting its potential for advancing computational semantics and natural language understanding through web-scale data.
Future research should refine NGD, address biases, extend to sentence analysis, develop scalable algorithms, explore multimodal semantics, and adapt NGD for industry-specific applications.

4. (Al Sulaiman M, 2022) Semantic textual similarity for modern standard and dialectal Arabic using transfer learning - Al Sulaiman et al
In the study under review, the authors focused on addressing the challenge of Semantic Textual Similarity (STS) in the Arabic language and its dialectical variants, specifically Egyptian and Saudi Arabic. They proposed three main approaches to improve the accuracy of STS models. They utilized automatic machine translation to translate English STS data into Arabic, interleaved English STS data with Arabic BERT models, and employed knowledge distillation-based models to fine-tune them using a translated dataset. Their dataset was manually

translated into Modern Standard Arabic (MSA), Egyptian Arabic, and Saudi Arabic. The main contribution of the paper was the development of STS models that significantly improved accuracy, achieving an absolute 2% gain over the state-of-the-art level for MSA.
While this work shows promising results for Arabic STS, there is still a considerable gap between the accuracy of Arabic models and English models. The techniques and methodologies presented in this paper hold the potential for addressing cross-lingual semantic textual relatedness task (track C), making them applicable to a wider range of Asian and African languages and dialects.

5. (Samruddhi Deode, 2023) L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT - Samruddhi Deode et al
The paper presents a novel approach to address the scarcity of high-quality language models for low-resource Indian languages. The authors introduce a range of SBERT models for ten popular Indian languages, trained using a synthetic corpus. This dataset is significant as it provides a valuable resource for languages with limited linguistic resources. The key innovation lies in the two-step training method, where models are fine-tuned using the NLI dataset followed by STSb, which results in substantial improvements in embedding similarity scores and cross-lingual performance. The paper demonstrates that the monolingual SBERT models outperform vanilla BERT models in terms of embedding similarity. Furthermore, the multilingual IndicSBERT exhibits strong cross-lingual performance, outperforming existing multilingual models like LaBSE.

The major achievements of the paper include the development of high-quality SBERT models for ten Indian languages, with particularly promising results for extremely low-resource languages like Odia and Punjabi. The multilingual IndicSBERT shows remarkable proficiency in handling both monolingual and multilingual datasets, making it a versatile tool for applications across multiple Indian languages. The robust cross-lingual properties of IndicSBERT, especially when compared to LaBSE, highlight its effectiveness in enhancing cross-lingual information retrieval systems and semantic search engines for diverse language requirements in countries like India.

The techniques presented in this paper offer a practical approach to creating language models for languages with limited resources. These techniques can be applied to other languages with similar constraints, making it possible to develop BERT models tailored to various linguistic needs. This work provides valuable insights and a blueprint for creating high-quality language models for languages with less data and resources, which can be particularly useful for the supervised track of semantic textual relatedness tasks.

4 Corpus/Data Description

4.1 Supervised and Cross-Lingual Tracks

Out of the 14 languages, namely, Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu, data has been released only for English, Marathi, Telugu, Hausa, Moroccan Arabic and Amharic.

The sizes of the datasets are noted in Table (4).

Language	Train	Dev
English	5500	250
Moroccan Arabic	1000	118
Amharic	992	95
Hausa	1736	212
Marathi	1200	300
Telugu	1170	130
Spanish	1562	140
Algerian Arabic	924	71
Afrikaans	1761	97

Table 1: Sizes of Training and Validation Sets for each language

At the preliminary level, the only non-semantic variable in these datasets is dataset length. To assess semantic relatedness, it's crucial to mitigate these biases.

Plots indicate no discernible correlation between sentence lengths and scores, suggesting well-distributed data suitable for training. Correlation coefficients fall in the range $-0.13 < \rho < 0.15$.

4.2 Unsupervised Track

1. In Track B of the research project, an unsupervised learning task was undertaken with a unique constraint – access to labeled datasets was prohibited. The initial focus was on the creation of bigram datasets, utilizing only unigram and bigram datasets. A vast corpus was assembled by scraping Wikipedia articles, serving as the primary data source. A method was devised to track the co-occurrence of words within this corpus.
2. For each bigram identified, a three-element tuple was created, capturing the number of times these words co-occurred in the same sentence, paragraph, and document. Word relationships were quantified, accounting for word repetitions in the same sentence. The goal was to derive word embeddings using these co-occurrence counts.

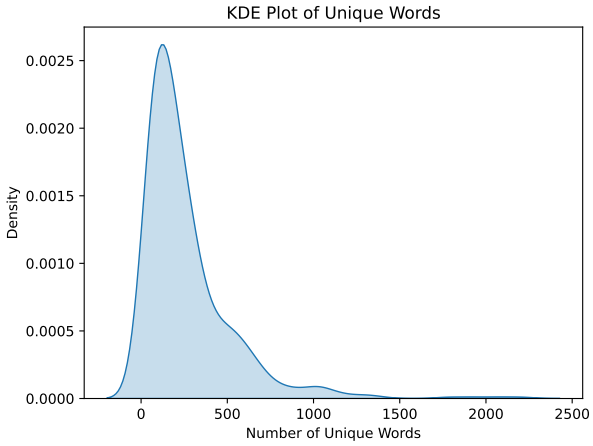


Figure 1: KDE Plot of number of unique Words in scraped Wikipedia page of each iteration

3. The plan included refining the approach by implementing hierarchical clustering to identify similarities and relationships between words. Additionally, a 1:1 **negative sampling strategy** was to be employed. The resulting embeddings would be used to calculate relatedness scores by integrating bigrams constructed from individual sentence pairs and lexical overlap.
4. A shift in direction occurred when clarification was received from the organizers, granting permission to utilize pre-trained models. At this point, the corpus creation had already been completed, yielding significant results.

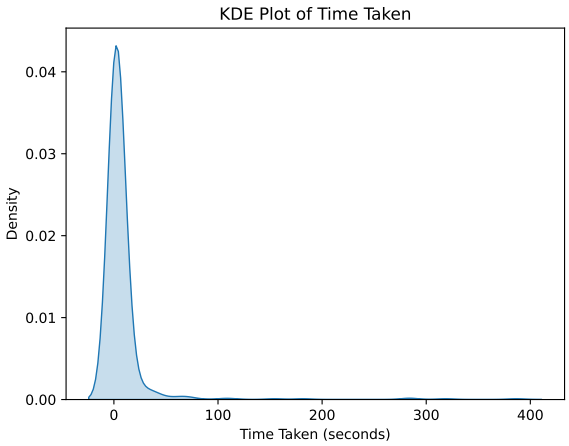


Figure 2: KDE Plot of time taken for scraping and updating vocabulary in each iteration

5. Despite the change in direction, important insights from the initial exploration of co-occurrence were retained. Notably, it was found that **65.22%** of unique words in the English test dataset matched words within the vocabulary derived from Wikipedia scraping. When generating bigrams from the English test dataset and cross-referencing with the co-occurrence dictionary, it was discovered that **53.97%** of bigrams had already occurred in the dataset, demonstrating extensive coverage. An additional **24.21%** of bigrams had individual words present in the vocabulary but had not co-occurred in any document.
6. These were the result of 600 iterations of scraping random Wikipedia articles, illustrating the potential and value of the co-occurrence dictionary. The project's focus later shifted towards leveraging pre-trained models after the announcement by organizers on the slack channel.
7. The organizers' confirmation prompted the use of SBERT to generate sentence embeddings for diverse languages. This involved models such as **all-MiniLM-L6-v2** for Hindi sentences and **paraphrase-xlm-r-multilingual-v1** for other languages like Hindi and Spanish. These models

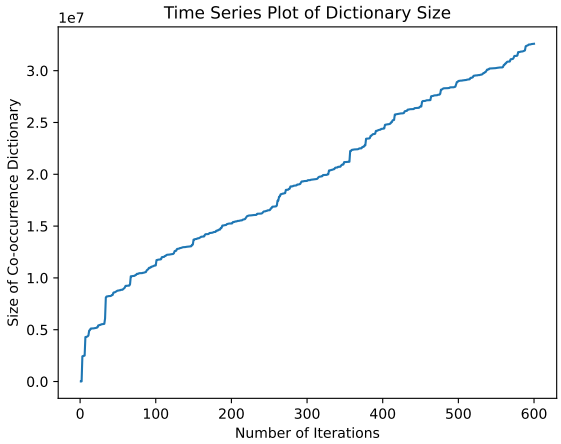


Figure 3: Evolution of the Bigram Co-occurrence Dictionary Size with Iterations

improved the accuracy of comparing and scoring sentences, contributing to a more effective multilingual approach.

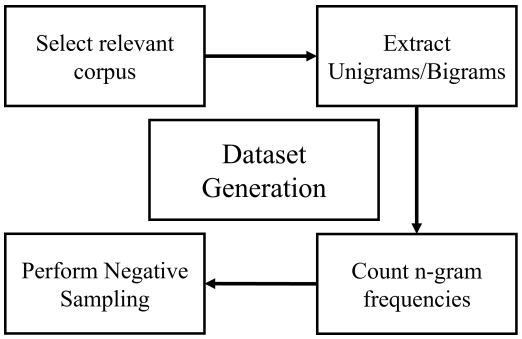


Figure 4: Data Generation Strategy for Unsupervised Task

5 Future Directions

1. Extending NGD for relatedness calculation among sentences

   The algorithm calculates the Normalized Google Distance (NGD) for two input sentences, text1 and text2. It starts by tokenizing and removing stop words from both sentences, followed by part-of-speech tagging. Then, it calculates NGD values for pairs of words with the same part of speech in both sentences. The NGD scores are normalized and averaged to compute the overall NGD score, representing the semantic similarity between the two sentences.

2. Our approach includes utilizing cardinality based metrics like **Dice coefficient** and **Jaccard Similarity**, and similarity and distance metrics like **Cosine similarity**, **Euclidean distance** and **Mahalanobis distance** on preprocessed tokens generated from the sentence pairs as well as devising a **transformer** based architecture for obtaining relatedness scores from the sentence embeddings produced.

3. sBERT based implementation

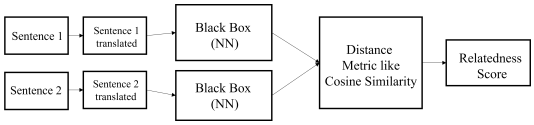


Figure 5: Sentence BERT based approach

4. Siamese Architecture based implementation

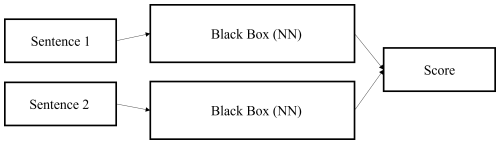


Figure 6: Siamese architecture

5. Initially each of the sentences are passed through a translation system to handle the multilingual aspect of the problem statement. The translated sentences are passed through as **sentenceBERT** to generate sentence embeddings. As for the next part, one approach includes efficiently implementing a **feed-forward neural network** which outputs the relatedness scores in the range of 0 and 1 from the two sentence embeddings. The other approach relies on implementation of an end-to-end **Siamese architecture** trained on the directly outputting the relatedness score without using metrics.

6. Searching for a good relatedness metric as of now or improving the performance of the system by tuning the parameters of the feed forward model to imitate a good relatedness metric system since **relatedness** metric is harder ot design as compared to a **semantic similarity** metric

7. The **supervised** task relies on:

   - Transformer based implementation (described above) for being trained on the la-

belled data provided in 14 different languages.

The **unsupervised** task relies on:

- Extraction of bigrams and training a Siamese architecture on those bigrams along with some negative samples created using **negative sampling** to generate bi gram embeddings
- These trained bigram embeddings are passed through an evaluation metric which are further clustered using methods like **DBSCAN**, **Hierarchical clustering** to calculate the relatedness score based on a threshold value.

The **Cross lingual** task is still under development.

8. Our plan of action currently includes tweaking the architecture parameters and choosing some good models to improve the relatedness score above a baseline which is **0.83**
9. The expected timeline for the future tasks in the project is mentioned below:

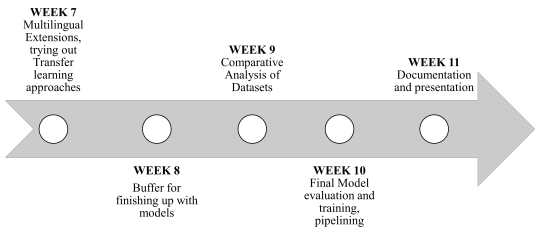


Figure 7: Expected Timeline

5.1 Experiments

(Wang et al., 2020) In order to extend the multilingual diversity of BERT, we have tried training **multilingual BERT** and extending the vocabulary to allocate the tokens of various low resource languages like **Amharic**, Hausa, **Algerian Arabic**, **Afrikaans**, **Indonesian** etc. The approach includes generation of the vocabulary of each some of the language and then calling the pretrained **Multilingual Bert** model and tokenizer. The pretrained tokenizer is extended to include the new tokens generated from the vocab of the corresponding low resource languages generated. We have used a trainable linear layer with added dropout. The loss metric used is **MSE loss** on both the training and validation data and the **Spearman Rank Correlation** is calculated at the end of each validation epoch.

The hyperparameters of the model are listed:

Hyperparameters	Value
Learning Rate	2e-5
Dropout Rate	0.1
Weight Decay	0.01
Number of linear layers	2
Activation	GELU
Max length	512

Table 2: Hyperparameters for extended mBERT

The loss and correlation values are mentioned in Table (4).

The second approach includes utilizing different distance based metrics like **Cosine Similarity**, **Mahalanobis Distance**, **Euclidean** and **Manhattan** distances as well as other lexical overlap based metrics like **Jaccard** and **Dice** coefficients. We have thereby constructed a vector which incorporates different orders of these scores ranging from 1 to 10. The dataset has columns named as $metric : n$ where **n** represents the metric calculated from the nþpower of sentence embeddings. Each row of the data essentially is a **42-element** sized vector covering higher orders of these metrics. Our attempt to create a new relatedness measure from the combination of all these metrics is accomplished by designing a simple **3-layered ANN** based model which outputs a score in the range of 0 and 1. The model is trained separately for some of the languages and the validation rank correlation values are mentioned in the given table (4)

The third approach involves around the idea of **Neural Machine Translation** system for low-resource languages for their conversion to english such that we can convert the low-resource languages back to english for facilitating the evaluation of the semantic relatedness score between the sentences. Machine translation can be used to check the consistency of semantic relatedness judgments across languages. For instance, if two sentences are deemed closely related in one language, their translations should also be closely related in another language. Machine translations can also be used for data augmentation via back translations process.

Language	Val loss	Val Correlation
Hausa	0.3133	0.0511
Spanish	0.1310	0.5627
Amharic	0.1902	0.4057
English	0.1315	0.7010
Marathi	0.0834	0.0483
Telugu	0.2213	0.3578

Table 3: Validation stats for some low resource languages

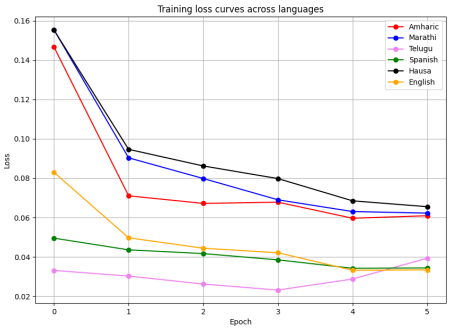


Figure 8: Train loss curves for some languages

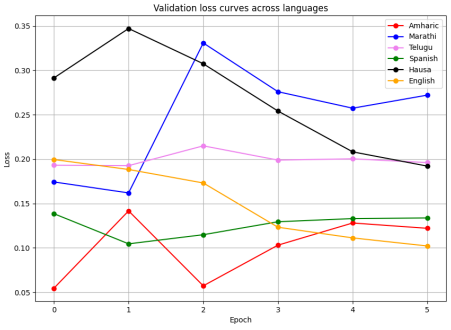


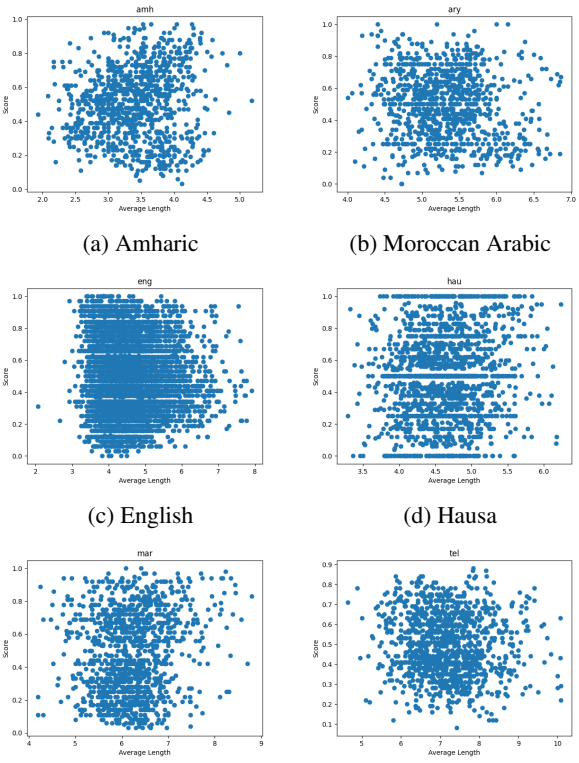
Figure 9: Validation loss curves for some languages

It can be used to generate sentence pairs in different languages to account for low training data in several languages.

Language	Val loss	Val Correlation
English	0.01769	0.8025

Table 4: Validation stats for metrics based model

Since the English dataset is large, we can do some more preliminary analysis on that data. Particularly, we are interested to see if there is any relation of the current SOTA models [sBERT(Reimers and Gurevych, 2019)] with the length of the sentences. The plot for that can be seen in Fig. 9. Visibly, there is no correlation between them.



(a) Amharic (b) Moroccan Arabic

(c) English (d) Hausa

(e) Marathi (f) Telugu

Figure 10: Plots of Length vs Scores on the training datasets. Some points(outliers) have been removed.

6 Individual Contribution

Rajarshi Dutta - Implementation of the multilingual BERT based approach to calculate the relatedness score using a trainable, feedforward neural network by extending existing vocabulary of BERT, also trying out other implementations like multilingual sBERT, Siamese networks etc. for **track-A**. Performing a comparative study of the correlation scores obtained by these models.

Shivam Pandey - Searching and navigating through several research papers for good relatedness metrics and models. Developed heuristics for entirely unsupervised STR task, creating algorithm from scratch, utilized various pre-trained models like sBERT and it's suitable multilingual variants for producing final results for **track-B**

Udvas Basak - Trying out and improving on tradi-

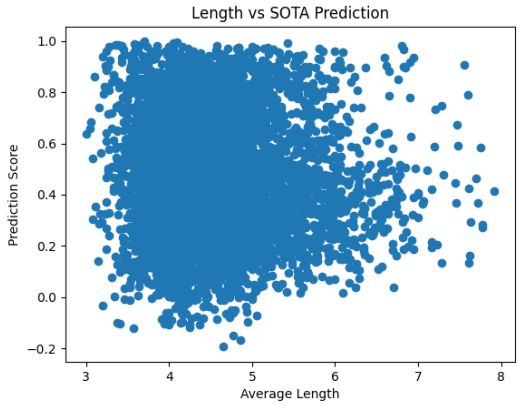


Figure 11: Average Length vs Prediction Score using sBERT. $\rho = -0.0468$

tional lexical overlap and distance based metrics on the datasets to achieve a baseline score. Devising a regression model to evaluate a relatedness metric from combinations of all higher orders of distance and lexical overlap scores which could be integrated to BERT or other contextual embeddings for **track-A**

7 Conclusion

In the evolving landscape of **Natural Language Processing**, the distinction and relationship between **semantic similarity** and **semantic relatedness** are paramount. This report underscored the nuanced differences between the two, highlighting the broader spectrum of relatedness in comparison to similarity. With the majority of previous research primarily concentrating on Semantic Textual Similarity and English language, the **SemEval 2024 challenge** presents a timely opportunity to delve deeper into the relatively less charted waters of semantic relatedness across multiple dimensions: **supervised**, **unsupervised**, and **cross-lingual**. Our objectives, as elucidated, steer towards developing comprehensive methods for computing semantic relatedness scores, harnessing both established and novel datasets, and adapting these methodologies to a cross-lingual setting thereby enhancing the richness of machine comprehension in diverse linguistic landscapes.

8 Presentation Feedback (Max 2 columns)

1. Questions include how do we plan to use the transformer and non-transformer based approaches and how do we exactly tackle the unsupervised portion of the problem statement?

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Abdou S Elgibreen H Faisal M Al Sulaiman M, Moussa AM. 2022. Semantic textual similarity for modern standard and dialectal arabic using transfer learning. *Plos One*.

Carla Teixeira Lopes and Diogo Moura. 2019. Normalized google distance in the identification and characterization of health queries. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Aditi Kajale Ananya Joshi Raviraj Joshi Samruddhi Deode, Janhavi Gadre. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. *CoRR*, abs/2004.13640.