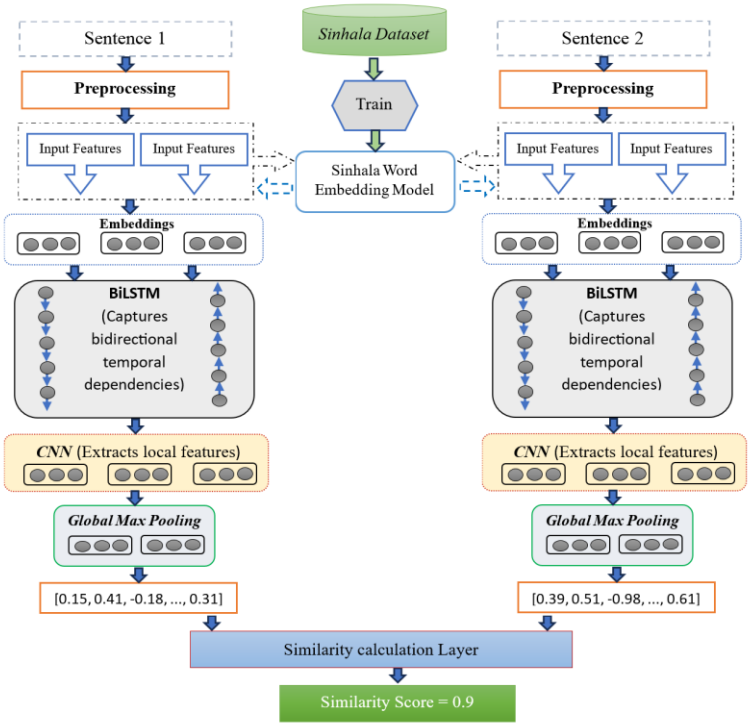
Siamese Hybrid Network Approach for Sentence Similarity

D.A.A. Deepal[1,2], A.M.R.R. Bandara[1] and P.R.S. De Silva[1]

[1]*Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura, Gangodawila, Sri Lanka.*

[2]*Faculty of Graduate Studies, University of Sri Jayewardenepura, Gangodawila, Sri Lanka.*

Date Received: 04-09-2024 Date Accepted: 25-12-2024



Abstract

This paper presents a novel Siamese Hybrid Network approach, namely Siamese Bidirectional Long Short Memory with Convolutional Neural Network (SiBiLConv), for evaluating the similarity in natural language. The model integrates a Siamese neural network architecture with similarity metrics, including Manhattan Distance and Cosine Similarity, to improve the accuracy of semantic relationships measurement between sentences. Evaluations were performed on Sinhala, a complex and under-resourced language spoken in Sri Lanka, which poses unique challenges due to its morphological richness and syntactic variability. The SiBiLConv model achieved an accuracy of 89.80%, an F1 score of 0.9041, and a mean squared error (MSE) of 0.0281 with the Cosine Distance metric outperforming baseline models such as MaLSTM, which achieved an accuracy of 78.99% and an F1 score of 0.7797. While existing methods for sentence similarity primarily focus on resource-rich languages, this work addresses the pressing need for tailored approaches in low-resource language contexts, where pre-trained models and annotated datasets are often limited. The novelty lies in SiBiLConv's hybrid architecture and metric integration, specifically designed to overcome the syntactic and semantic complexities of Sinhala. This research not only bridges a critical gap in the application of sentence similarity models for low-resource languages but also establishes a framework adaptable to other morphologically rich languages, advancing the broader scope of natural language processing.

Keywords: *Siamese Hybrid Network, Sentences similarity, Sinhala sentence similarity, Morphologically Rich Language Processing*

**Correspondence: ravi@sjp.ac.lk*

1. Introduction

Sinhala, an official language of Sri Lanka, is known for its linguistic complexity, characterized by rich morphology, high inflection, and nuanced word order. These features make natural language processing (NLP) tasks in Sinhala particularly challenging. A notable example is the diverse manifestations of named entities within sentences, such as the name 'නදීක' (Nadeeka) which can appear as variations like නදීකට (Nadeeka'ta'), නදීකව (Nadeeka'va'), නදීකම (Nadeeka'ma'), නදීකටම (Nadeeka'tama'), නදීකවම (Nadeeka'vama'), නදීකගෙ (Nadeeka'ge'), නදීකගේ (Nadeeka'gē'), නදීකගෙම (Nadeeka'gema'), නදීකගේම (Nadeeka'gēma') depending on contextual usage. Such variations, while not entirely unique to Sinhala, occur more frequently and with greater complexity than in languages like English, emphasizing the need for tailored NLP approaches.

Sinhala exhibits a high degree of inflection, leading to various forms of verb conjugation. For example, the verb 'go' in the past tense can take forms such as 'ගියේය' (giyēya), 'ගියෝය' (giyōya), and 'ගියෙමි' (giyemi), depending on the subject type within the sentence (Lakmal et al., 2020). These morphological nuances significantly complicate semantic and syntactic analysis. Furthermore, the overlap of proper nouns with common nouns, such as "නුවණ්" (denoting both "eyes" and a person's name), introduces additional ambiguity, complicating disambiguation tasks (Manamini et al., 2016).

Existing methods for sentence similarity assessment, often reliant on word embeddings, encounter significant limitations. Word embeddings typically represent sentences as a combination of their constituent word vectors, which can lead to challenges in capturing subtle semantic differences. For example, they often fail to consider the importance of word order, resulting in identical representations for semantically distinct sentences. As noted by Mueller et al. (Mueller and Thyagarajan, 2016) and Gao et al. (Gao et al., 2021), embedding-based models may struggle with generalization in tasks requiring nuanced semantic understanding. For instance: Sentence 1: නරියෙක් (a fox) බල්ලෙක්(a dog) පසුපස(after) ලුහුබඳී(chase). (A fox chases after a dog.) and Sentence 2: බල්ලෙක්(a dog) නරියෙක්(a fox) පසුපස(after) ලුහුබඳී(chase). (A dog chases after a fox.) While both sentences share the same words, their semantic meanings are entirely different due to word order.

Conversely, in cases such as: Sentence 1: බල්ලන් සහ පූසන් යනු හොඳ මිතුරන් නොවේ. (Dogs and cats are not good friends.) and Sentence 2: පූසන් සහ බල්ලන් යනු හොඳ මිතුරන් නොවේ. (Cats and dogs are not good friends.), the word order is less critical, and the sentences retain equivalent meanings. This variability highlights the need for approaches capable of capturing both order-sensitive and order-invariant relationships.

Despite the importance of sentence similarity in NLP, significant research gaps remain as listed below.

Morphological and Syntactic Richness: Existing models struggle to capture the intricate relationships arising from word inflections, syntactic variations, and the contextual use of named entities in Sinhala.

High computational requirements: State-of-the-art transformer-based models like BERT are computationally intensive, making them unsuitable for low-resource environments or applications requiring offline functionality in disconnected settings.

Real-World Applicability: There is a lack of lightweight and efficient models tailored for real-world scenarios, such as embedded systems, chatbots, and robots operating without internet connectivity or access to APIs.

This study addresses these gaps through a novel Siamese hybrid neural network, Siamese Bidirectional Long Short-Term Memory with Convolutional Neural Network (SiBiLConv). The proposed model integrates computationally efficient metrics Manhattan Distance and Cosine Similarity within a hybrid architecture to capture both semantic and syntactic nuances while minimizing resource requirements. Compared to traditional corpus-based or embedding-based approaches, SiBiLConv achieves superior performance in handling the complexities of Sinhala and similar languages.

*Correspondence: ravi@sjp.ac.lk

The key contributions of this study include an architecture that combines linguistic adaptability and computational efficiency, making it suitable for deployment in low-resource environments or offline systems. Benchmarking against baseline methods demonstrates the model's superior performance, establishing its relevance for advancing NLP in Sinhala and other low-resource languages. By addressing morphological and syntactic challenges and ensuring real-world applicability, this work lays a foundation for NLP research in morphologically complex and resource-constrained settings.

The historical significance and linguistic richness of Sinhala present an intriguing challenge that calls for innovative solutions. However, research focusing on Sinhala language processing, particularly in the field of string similarity, remains sparse. This scarcity has led to difficulties in designing and developing robots and chatbots based on the Sinhala language.

This paper ventures into the domain of hybrid score functions, specifically tailored for assessing sentence similarity in Sinhala. Our approach integrates modern word embedding techniques with well-known metrics such as the Cosine similarity, Manhattan Distance, and Siamese Hybrid network Model for measuring Sinhala Sentence Similarity. We named the proposed model as Siamese Bidirectional Long Short Term with Convolutional Neural Network (SiBiLConv), which will be used to refer to this model hereafter. Through the fusion of these methods, we hope to achieve a high level of accuracy in measuring the similarity between Sinhala sentences, potentially opening new avenues in language processing.

2. Related work

The field of sentence similarity has evolved significantly with advances in both traditional methods and deep learning techniques. Although early approaches employed statistical and rule-based models, recent developments have seen a shift towards more sophisticated methods, including neural network architectures and hybrid models. This section discusses the key approaches, their strengths and limitations, and their relevance to low-resource languages like Sinhala.

2.1. Traditional and Statistical Methods

Traditional methods for sentence similarity primarily relied on statistical and rule-based approaches. Kadupitiya et al. (Kadupitiya et al., 2016) proposed a hybrid methodology combining corpus-based and knowledge-based similarity measures. This approach captured both semantic and syntactic relationships but faced scalability issues, particularly for morphologically rich languages like Sinhala. Meng et al. (Meng et al., 2021) enhanced statistical models by integrating a domain-specific dictionary with TF-IDF and cosine similarity, showing improvements in specialized domains. However, these techniques lacked the ability to capture deep semantic meaning, making them less effective for diverse and complex linguistic tasks.

2.2. Word and Sentence Embedding Techniques

Word embeddings have played a pivotal role in natural language understanding by encoding words as dense vectors. Mikolov et al. (Mikolov et al., 2013) introduced Word2Vec, an efficient model with CBOW and Skip-gram architectures, laying the groundwork for modern embeddings. Building on this, FastText (Bojanowski et al., 2017) incorporated subword information, improving performance for morphologically complex languages and addressing limitations of traditional word embeddings. The Universal Sentence Encoder (Cer et al., 2018) introduced transformer-based and deep averaging network models, demonstrating exceptional versatility in transfer learning tasks across diverse NLP applications. Liyanage et al. (Ranathunga and Liyanage, 2021) applied the Word2Vec Skip-gram model to Sinhala sentiment classification, achieving better results than traditional Bag-of-Words approaches. More recently, Weeraprameshwara et al. (Weeraprameshwara et al., 2022) developed a two-tiered sentence embedding model specifically for Sinhala, showing improved performance in sentiment analysis tasks compared to traditional one-tier word embeddings. However, the applicability

*Correspondence: ravi@sjp.ac.lk

of these embedding models for the specific application, which is the highlight of the current research, has not been well explored. For Sinhala, Lakmal et al. (2020) demonstrated that FastText outperformed other models in intrinsic (word analogy) and extrinsic (sentiment analysis) evaluations, further validating its efficacy for morphologically rich languages. Hence, the proposed model utilized FastText as one of its representation components.

2.3. Deep Learning Models (Non-Transformer)

Deep learning has introduced more advanced techniques for sentence similarity, with Siamese networks emerging as a key architecture. He et al. (He et al., 2015) demonstrated the efficacy of CNN-based Siamese architectures for capturing lexical, syntactic, and semantic features. Mueller and Thyagarajan (Mueller and Thyagarajan, 2016) proposed MaLSTM, which leveraged LSTM networks and Manhattan distance to encode semantic relationships effectively, laying the groundwork for subsequent neural methods. Ichida et al. (Ichida et al., 2018) tailored Siamese GRU models for semantic similarity using metric learning and GRU for improved contextual understanding. Zhu et al. (Zhu et al., 2018) further advanced the field by integrating BiLSTMs with attention mechanisms in a Siamese network framework to address challenges such as word segmentation and character variations. Edo-Osagie and co-workers (Edo-Osagie and De La Iglesia, 2019) designed Attention-Based RNN models (ABRNN) for short text classification, leveraging attention mechanisms to enhance word importance in applications like Twitter data classification. While these methods have demonstrated exceptional performance, their applicability on under-resourced languages have not been well explored. Hence, the current research proposes a method that utilizes the efficacy of CNN based architecture as well as the recurrent information capturing capabilities of BiLSTM model together to achieve better similarity measurement of sentences in Sinhala, the under-resourced language.

2.4. Transformer-Based Models

Recent advancements in transformer-based models have set a new standard in NLP. Devlin et al. (Devlin et al., 2019) introduced BERT, a bidirectional transformer model pre-trained on large corpora, enabling the extraction of rich contextual embeddings. Building on this, Reimers and Gurevych (Reimers and Gurevych, 2019) developed Sentence-BERT (SBERT), which incorporates a Siamese network to enhance computational efficiency for sentence similarity tasks. In 2021, Thakur et al. (Thakur et al., 2021) proposed BEIR, a benchmark for zero-shot evaluation in information retrieval, which is particularly relevant for cross-lingual applications. Wijaya et al. (Wijaya, 2021) adapted BERT for automatic grading of short answers in Indonesian, demonstrating its adaptability to non-English languages. Gao et al. (Gao et al., 2021) introduced SimCSE, a state-of-the-art contrastive learning framework for sentence embeddings, which employs both supervised and unsupervised techniques. Dhananjaya et al. (Dhananjaya et al., 2022) advanced transformer-based models for Sinhala with SinBERT-large and SinBERT-small, publicly released Sinhala-specific RoBERTa models that contribute to sentence similarity tasks. In 2024, John Snow Labs Inc. ("JohnSnowLabs/spark-nlp," 2024) developed SinhalaWord2Vec, further advancing Sinhala-specific language models. It is a known challenge that transformer-based models are data-hungry and which causes these models not to grasp the right domain when fine-tuned with smaller dataset which is typical for under-resourced languages (Wang et al., 2024).

2.5. Hybrid Architectures

Hybrid models combining different neural network architectures have shown promise in improving sentence similarity and text classification tasks. Liyanage et al. (Ranathunga and Liyanage, 2021) demonstrated the potential of combining Word2Vec embeddings with machine learning classifiers such as SVM, Logistic Regression, and Random Forest for Sinhala sentiment classification, achieving notable results for resource-constrained languages. Wang et al. (Wang et al., 2020) introduced a hybrid model that combines CNNs with semantic expansion techniques, enhancing word

*Correspondence: ravi@sjp.ac.lk

matching precision and improving performance in short text classification. Ranathunga et al. (Ranathunga and Liyanage, 2021) extended this concept for Sinhala sentiment classification, integrating machine learning classifiers with Word2Vec embeddings to boost classification accuracy for Sinhala news comments.

An overview of various techniques, methods, and their key contributions in the field of sentence similarity, including their strengths, limitations, and applicability across different categories such as traditional methods, word embeddings, deep learning models, transformer-based models, and hybrid architectures are presented in Table 1. This table situates the current research within the broader context of NLP advancements.

Table 1. Summary of Key Methods and Contributions in Sentence Similarity Research

Category	Year	Author/s	Technique/Method	Key Contributions
Traditional and Statistical Methods	2016	Kadupitiya et al.	Corpus-based and Knowledge-based Similarity Measures	Combines semantic and syntactic features for better similarity measurement.
	2021	Meng et al.	TF-IDF + Cosine Similarity	Enhances statistical models by integrating a domain-specific dictionary with TF-IDF and cosine similarity, showing improvements in specialized domains.
Word and Sentence Embedding Techniques	2013	Mikolov et al.	Word2Vec (CBOW, Skip-gram)	Introduces dense vectors for efficient word encoding but struggles with OOV words and morphological nuances.
	2017	Bojanowski et al.	FastText	Incorporates subword information, effective for morphologically complex languages.
	2018	Cer et al.	Universal Sentence Encoder	Leverages transformers for sentence embeddings, suitable for transfer learning tasks.
	2020	Lakmal et al.	FastText	Demonstrates superior performance for Sinhala in intrinsic (word analogy) and extrinsic (sentiment analysis) evaluations.
	2021	Ranathunga et al.	Sinhala Word2Vec Skip-gram	Applies embedding techniques for Sinhala sentiment classification, showing improved performance over traditional Bag-of-Words approaches.
	2022	Weeraprameshwara et al.	Two-tiered sentence embedding model	Developed for Sinhala, showing improved performance in sentiment analysis compared to traditional one-tier embeddings.
Deep Learning Models (Non-Transformer)	2015	He et al.	CNN-based Siamese architectures	Captures lexical, syntactic, and semantic features effectively.
	2016	Mueller and Thyagarajan	MaLSTM	Uses LSTM networks and Manhattan distance for effective semantic relationship encoding.
	2018	Ichida et al.	Siamese GRU models	Tailored for semantic similarity using metric learning and GRU for improved contextual understanding.
	2019	Edo-Osagie and De La Iglesia	Attention-Based RNNs	Leverages attention mechanisms to enhance word importance in short texts.

*Correspondence: ravi@sjp.ac.lk

Transformer-Based Models		Devlin et al.	BERT	Pre-trained transformer model for bidirectional sentence embeddings.
		Reimers and Gurevych	Sentence-BERT	Enhances computational efficiency for sentence similarity using a Siamese architecture.
	2021	Thakur et al.	BEIR	A benchmark for zero-shot evaluation in information retrieval, relevant for cross-lingual applications.
		Wijaya	BERT adaptation for Indonesian language	Automatic grading of short answers, demonstrating adaptability to non-English languages.
		Gao et al.	SimCSE	Contrastive learning framework for sentence embeddings, effective in cross-lingual tasks.
	2022	Dhananjaya et al.	SinBERT-large, SinBERT-small	Sinhala-specific RoBERTa models for sentence similarity tasks.
	2024	JohnSnowLabs/spark-nlp	SinhalaWord2Vec	Advances Sinhala-specific language models, addressing data availability challenges.
Hybrid Architectures	2020	Wang et al.	CNNs + semantic expansion	Enhances word matching precision in short text classification tasks.
	2021	Ranathunga et al.	Word2Vec + machine learning classifiers	Combines embeddings with SVM, Logistic Regression, and Random Forest for Sinhala sentiment analysis.
		Ranathunga et al.	CNN + BiLSTM	Combines CNN for local features and BiLSTM for long-range context in Sinhala sentence similarity.
		Yoo et al.	CNNs, RNNs, and BERT	Uses multiple neural network layers with cosine similarity for sentence similarity.
	2022	Liu et al.	CNN + LDA + Word2Vec	Integrates CNN, LDA, and Word2Vec for short text classification.
		Ji and Zhang	HA-RCNN	Combines BiGRU, CNN, and headword attention for improved short text similarity.
	2023	Yang and Zhang	Hybrid architectures	Optimized for sentence similarity in morphologically complex languages like Sinhala.

Yoo et al. (Yoo et al., 2021) employed a combination of CNNs, RNNs, and BERT with cosine similarity, refining sentence similarity tasks and offering promise for low-resource languages like Sinhala. Liu et al. (Liu et al., 2022) further advanced hybrid models by proposing a CNN-based method that integrates Word2Vec with LDA topic modeling, addressing challenges in short text classification. Ji and Zhang (Ji and Zhang, 2022) introduced HA-RCNN, combining BiGRU, CNN, and headword attention for short text similarity. Most recently, Yang and Zhang (2023) explored hybrid architectures optimized for sentence similarity in morphologically complex languages. The existence of a vast majority of the literature regarding the hybrid models being used in under-resourced languages like Sinhala, emphasize that this is a potential direction where the sentence similarity measurement can be improved in Sinhala, hence hybrid approach has inspired the proposed model.

3. Proposed Method

To address the unique challenges in Sinhala sentence similarity, we propose a Siamese Hybrid Network approach named SiBiLConv. This method leverages Siamese Bidirectional Long Short-Term

*Correspondence: ravi@sjp.ac.lk

Memory (BiLSTM) and Convolutional Neural Networks (CNNs) with advanced similarity metrics such as Manhattan Distance and Cosine Similarity. By combining the strengths of BiLSTMs for contextual relationships and CNNs for local feature extraction, SiBiLConv achieves both high accuracy and computational efficiency.

The proposed model is designed to handle the morphological and syntactic richness of Sinhala, capturing both semantic nuances and structural relationships, thereby making it particularly well-suited for low-resource environments and applications requiring offline deployment. The architecture also incorporates a robust preprocessing pipeline and a high-quality annotated dataset to ensure linguistic relevance and diversity.

3.1 Data Collection and Preprocessing

3.1.1 Sinhala Word Embedding Model Dataset

Sinhala sentences can predominantly be categorized into two patterns: spoken and written. While online resources are abundant for written patterns, our research focuses on establishing a robust environment for calculating Sinhala Sentence Similarity based on the Siamese Hybrid Network Approach. To create a robust dataset for Sinhala natural language processing, we employed a rigorous methodology, carefully selecting diverse and reliable sources to ensure linguistic richness and contextual relevance.

For the Sinhala word embedding model, data was gathered from both primary and domain-specific sources. The Common Crawl dataset (de Silva, 2023) served as a foundational resource, offering a broad spectrum of textual content from web pages that captured both formal and informal Sinhala usage. Additionally, Sinhala Wikipedia (de Silva, 2023) provided structured examples of formal language, enriching the dataset's lexical and contextual variety. To supplement these, we incorporated government-published Sinhala medium textbooks, which represented structured, formal language commonly used in educational contexts, and online news articles, which reflected evolving linguistic trends and informal sentence structures.

The preprocessing pipeline was critical to ensuring the quality of the dataset. This involved removing web-related tags, numerical data, non-Sinhala characters, and special symbols to create clean and linguistically accurate text. These steps ensured that the final corpus was tailored for training Sinhala-specific word embeddings, capturing both formal and informal aspects of the language.

3.1.2 Sentence similarity Dataset

The Sinhala sentence similarity dataset consists of 14,200 annotated sentence pairs, carefully curated to represent the richness and complexity of the Sinhala language. A significant portion of this dataset was derived from the Sentences Involving Compositional Knowledge (SICK) dataset (Marelli et al., 2014), a widely used English dataset containing 10,000 sentence pairs designed to evaluate compositional semantic relationships. The SICK dataset was chosen for its comprehensive annotations and diverse semantic relationships, which were ideal for translation and adaptation to Sinhala. While Sinhala-specific datasets could provide native language data, existing resources were limited in scope and lacked the semantic richness and diversity required for this study. The translation of the SICK dataset into Sinhala ensured a strong foundation for developing a robust similarity model, bridging the gap between high-quality semantic annotations and low-resource language needs.

During the adaptation process, expert linguists manually translated selected SICK sentences into Sinhala, emphasizing semantic fidelity and linguistic adaptability. To reduce bias, three linguists discussed each sentence prior to translation, ensuring consensus on the intended meaning and addressing any potential ambiguities. This collaborative approach helped avoid individual biases in translation and ensured that the sentences were accurately represented. Translation bias was further minimized by excluding meaningless or semantically redundant pairs and ensuring that the translated sentences preserved their original intent while incorporating unique linguistic features of Sinhala, such

*Correspondence: ravi@sjp.ac.lk

as morphological inflections and contextual word order. Regular reviews and iterative refinement further reduced the potential for errors or misrepresentation in the dataset.

In addition to the SICK dataset, other sentence pairs were sourced to enhance linguistic diversity. These included annotated sentence pairs from the NLP Centre at the University of Moratuwa (NLP Centre, University of Moratuwa, 2021), publicly available datasets on GitHub, and Sinhala news articles written by different authors on similar topics. Custom-generated sentence pairs were also included to address underrepresented linguistic phenomena, such as idiomatic expressions, complex syntax, and morphological nuances. This comprehensive approach ensured a dataset that was both balanced and representative of real-world usage.

To refine the dataset, a robust preprocessing pipeline was applied. This included the removal of web-related tags, punctuation, special characters, and numeric figures to eliminate noise. Non-Sinhala characters, such as English letters, were filtered out to ensure linguistic consistency. The resulting clean dataset was split into 80% for training and 20% for validation, with minimal overlap between the two subsets to enable independent evaluation of machine learning models. Assumptions about dataset quality were addressed by prioritizing linguistic diversity and domain coverage, ensuring the dataset reflected both formal and informal usage scenarios.

By combining a high-quality English dataset with localized adaptations and supplementary Sinhala sources, this dataset provides a strong foundation for developing and evaluating sentence similarity models. Its rigorous design and thoughtful curation address the unique challenges of Sinhala NLP while ensuring applicability to practical, real-world tasks.

3.1.3 Data Annotation

To ensure the reliability and granularity of the dataset, a 0–10 annotation scale was adopted for assigning similarity scores to each sentence pair. Three expert annotators evaluated each pair, with 0 indicating complete dissimilarity and 10 representing absolute similarity. To minimize bias and ensure consistency, the annotators discussed any discrepancies in their initial assessments before finalizing the scores. In cases where consensus could not be reached, the final score was determined by averaging the ratings of the annotators. This scale was specifically chosen for its flexibility and granularity, which addressed several limitations observed in other commonly used scales, such as the 1–5 or 7-point Likert scales.

In the original SICK, a 1–5 scale was employed to annotate semantic similarity. However, this scale posed challenges, particularly in cases where sentence pairs were weakly related. For example, a score of 1 denoting strong dissimilarity often failed to account for subtle degrees of relatedness, leading to ambiguity and potential annotator discomfort. By extending the scale to 0–10, we explicitly defined 0 as complete dissimilarity, removing any confusion about the absence of a relationship between sentences. This broader scale facilitated more precise annotation, enabling the annotators to capture nuanced differences in semantic similarity more effectively.

Additionally, the 0–10 scale allows for smoother normalization to a [0, 1] range, which is critical for machine learning applications. The finer granularity also enhances the dataset's suitability for tasks requiring high precision, such as training models for semantic similarity in low-resource languages like Sinhala.

To further enhance the reliability of the dataset, the annotated similarity scores were used to categorize sentence pairs into similar and dissimilar groups. A similarity threshold of 0.7 was applied, where pairs with scores ≥ 0.7 were labeled as similar and those below the threshold as dissimilar. This threshold enabled finer granularity in similarity assessments, supporting the development of robust and accurate models.

The dataset was split into 80% for training and 20% for validation, ensuring minimal overlap between subsets to maintain independent evaluations. By combining diverse sources, meticulous annotation, and thoughtful curation, the resulting dataset captures the linguistic richness of Sinhala, addressing its unique morphology, syntax, and semantics. This carefully curated resource provides a

*Correspondence: ravi@sjp.ac.lk

strong foundation for developing and evaluating Sinhala sentence similarity models, supporting both theoretical research and real-world applications.

3.2 Sinhala Word Embedding Model

In this study, we employ a FastText model based on the Continuous Bag of Words (CBOW) architecture, which was released by Facebook (Bojanowski et al., 2017). This model, characterized by a vocabulary size of approximately 3.95 million words and subword n-grams, generates high-dimensional vectors of size 300 to represent words. The learning rate, set at 0.025, facilitates effective training. Operating under the CBOW paradigm, the model excels in predicting target words by utilizing the context of the words within a sentence. It is noteworthy that the model was trained on a refined dataset described in Section 3.1.1, providing linguistic and cultural insights specific to the Sinhala language. During training, a window size of 5 was employed, enabling the model to consider the context within a span of five words on either side of the target. By harnessing subword information, the model adeptly handles out-of-vocabulary words and captures intricate morphological relationships.

FastText demonstrates proven effectiveness in handling morphologically rich languages like Sinhala. Lakmal et al. (Lakmal et al., 2020) demonstrated that FastText outperformed Word2Vec and GloVe in both intrinsic (word analogy) and extrinsic (sentiment analysis) evaluations for Sinhala. Key advantages of FastText include its ability to incorporate subword information, which is critical for representing the complex morphology of Sinhala, such as prefixes, suffixes, and infixes. This feature enables the model to generate embeddings that are robust to out-of-vocabulary (OOV) words and nuanced linguistic variations.

Furthermore, Word2Vec, which operates at the word level, often struggles with OOV words and cannot effectively capture the rich inflectional nature of Sinhala. GloVe, while efficient in leveraging global co-occurrence statistics, has shown comparatively lower performance for Sinhala due to dataset size limitations (Lakmal et al., 2020). FastText's subword-based approach provides a significant advantage in these scenarios, making it the most suitable embedding method for this study.

3.3 Siamese Hybrid network

The Siamese neural network (Lecun et al., 1998), also known as a twin neural network, is a specialized architecture designed to compute comparable output vectors by processing two distinct input vectors simultaneously using identical weights and structures across its parallel branches. This shared-weight design ensures consistent feature extraction and enables the computation of a similarity metric between input sentences.

The proposed Siamese Hybrid Network incorporates three key components: an Input Layer, a Semantic Embedding Module, and a Similarity Calculation Layer, as illustrated in Figure 1. This architecture is specifically tailored to capture semantic relationships and assess sentence similarity, making it particularly effective for morphologically rich languages like Sinhala. By leveraging its shared-weight structure, the model reduces the number of parameters, ensuring robustness, computational efficiency, and reduced memory usage. These characteristics make the network suitable for resource-constrained environments and offline systems, such as embedded applications in chatbots and robots.

The Siamese Hybrid Network achieves a high level of semantic understanding of sentence pairs, even with limited labeled data. Its design facilitates efficient training and evaluation, providing a robust foundation for developing sentence similarity models and advancing natural language processing applications in Sinhala. By addressing the linguistic challenges of inflections, named entities, and word order, the network ensures accurate sentence similarity modeling in low-resource settings.

a) Input Layers:

The network uses two input layers to receive sentence pairs for comparison. These inputs are processed independently but identically in the Semantic Embedding Module, ensuring consistent feature extraction for both sentences.

*Correspondence: ravi@sjp.ac.lk

b) Semantic Embedding Module

The Semantic Embedding Module begins with a pre-trained embedding layer initialized with FastText embeddings, as detailed in Section 3.2. These embeddings, specifically selected for Sinhala, represent words as dense vectors, capturing subword information critical for handling morphological richness and out-of-vocabulary words. The subsequent layers in the module include:

Pre-Trained Embeddings: FastText embeddings are employed to represent words as dense vectors, leveraging subword information critical for handling Sinhala's morphological richness and out-of-vocabulary words.

Bidirectional Long Short-Term Memory (BiLSTM): Captures bidirectional temporal dependencies, ensuring that context from both preceding and succeeding words influences the representation. This is particularly important in Sinhala, where word order significantly impacts meaning.

***Convolutional Neural Network (CNN)*:** Extracts local features, such as n-grams, which are essential for sentence similarity tasks. CNNs are effective at detecting short word sequences and syntactic patterns that complement the long-range context captured by BiLSTM. This hybrid approach enables the model to capture both local patterns (via CNN) and long-range dependencies (via BiLSTM), enhancing its ability to understand complex relationships in sentences, especially in languages like Sinhala. He et al. (2015) and Zhu et al. (2018) showed that CNNs, when combined with LSTMs or BiLSTMs, improve performance in semantic similarity tasks by capturing both local and global features. Additionally, CNNs offer computational efficiency, enabling faster training and inference.

Global Max Pooling: Reduces the dimensionality of feature maps while retaining the most salient features, ensuring computational efficiency.

c) Similarity Calculation Layer

The Similarity Calculation Layer plays a crucial role in SiBiLConv model, measuring the similarity between input sentences. This layer employs two complementary metrics: the Manhattan Distance layer and the Scaled Cosine Similarity layer.

i. Manhattan Distance layer

Manhattan distance has been shown to outperform other metrics by facilitating the creation of a structured semantic space, effectively capturing sentence similarity (Mueller and Thyagarajan, 2016). This layer calculates the Manhattan distance (L1 distance) by subtracting the corresponding elements of two input vectors, taking the absolute values, summing these differences, and exponentiating the negative of the result. This approach ensures that the computed dissimilarity values fall within the range of 0 to 1. By integrating this formulation, the layer enhances the model's ability to detect subtle semantic variations in Sinhala sentences. The exponential transformation, adopted from Mueller and Thyagarajan (Mueller and Thyagarajan, 2016), further ensures smooth scaling of dissimilarity scores..

ii. Scaled Cosine Similarity layer

Cosine similarity has been shown to be highly effective in evaluating word and sentence embeddings, particularly for tasks involving semantic comparisons (Lakmal et al., 2020), (Meng et al., 2021). Its scale-invariant nature ensures consistent results regardless of the magnitudes of the embedding vectors. The Scaled Cosine Similarity layer computes the cosine distance between two input vectors by first calculating their dot product and then normalizing each vector's magnitude. Cosine similarity is widely used in natural language processing for assessing the similarity between vectors. To ensure numerical stability, the layer incorporates a small epsilon value to prevent division by zero. The final output represents the complement of the cosine similarity, providing the cosine distance between the input vectors.

To scale the cosine similarity values between 0 and 1, the layer employs the following equation:

$$\text{Scaled Cosine Similarity} = \frac{\text{Cosine Similarity}+1}{2} \quad (1)$$

This layer plays a pivotal role in enhancing the network's ability to discern semantic relationships and capture nuanced patterns in the data. Consequently, it contributes significantly to the success of our research in Sinhala sentence similarity estimation.

*Correspondence: ravi@sjp.ac.lk

3.4 Manhattan Distance Model

In the Manhattan Distance Model, the Siamese Neural Network architecture, constructed upon the foundation of the initial network, is employed. This model integrates a Manhattan Distance layer, efficiently calculating the Manhattan distance between the outputs of the shared model for two input sentences. This distance measure serves as a pivotal indicator of their similarity.

3.5 Cosine Distance Mode

In the Cosine Distance Mode, the Siamese Neural Network architecture builds upon the foundation of the initial network. This model incorporates a Scaled Cosine Distance layer, proficiently calculating the Cosine distance between the outputs of the shared model for two input sentences. This distance measure similarly serves as a pivotal indicator of their similarity.

3.6 Cosine Similarity Word Embedding Model

In our study, we adopted a widely recognized method to measure sentence similarity based on word embeddings, utilizing a pre-trained Sinhala Word Embedding Model as described in Section 3.2, featuring high-dimensional vectors of size 300. The word embedding model, which represents words in a continuous vector space, effectively captures semantic relationships between words. To construct sentence vectors, we employed a standard approach: extracting word vectors for each term in the sentence and then computing the average vector. This process yields a dense representation that encapsulates the semantic content of the entire sentence. Specifically, we utilized the Scaled cosine similarity metric to quantify the resemblance between pairs of sentence vectors. This metric measures the cosine of the angle between two vectors, yielding a similarity score that ranges from 0 (completely dissimilar) to 1 (identical).

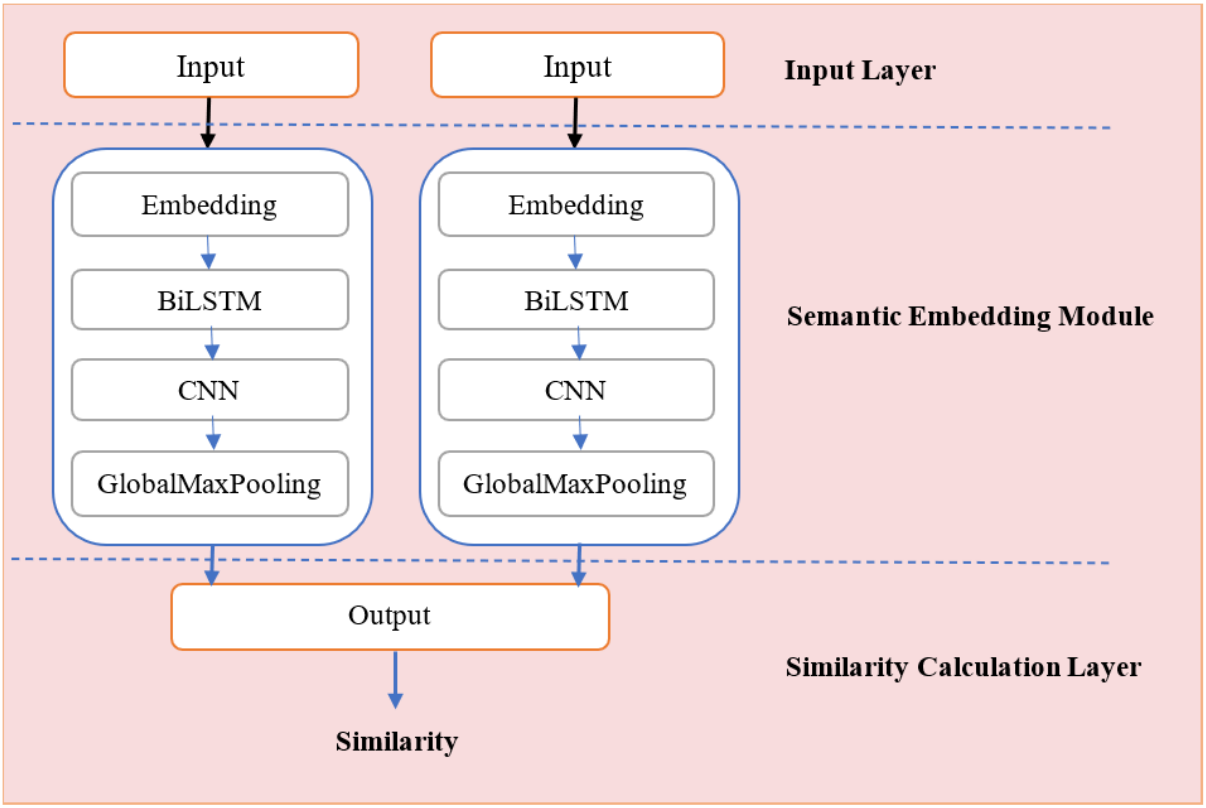


Figure 1: Siamese Hybrid Network Architecture (SiBiLConv). Consists of a pre-trained FastText embedding layer, followed by a Bidirectional Long Short-Term Memory (BiLSTM) layer to capture long-range dependencies, a Convolutional Neural Network (CNN) for local feature extraction, and a final similarity calculation layer.

*Correspondence: ravi@sjp.ac.lk

4. Experiments

4.1 Experimental Setup

The proposed model is evaluated experimentally comparing with MaLSTM (Mueller and Thyagarajan, 2016) model and the GRU (Mikolov et al., 2013) (Ranasinghe et al., 2019) model which is the most related and competing model found in the literature, using the self-collect Sinhala similarity Dataset. These models were evaluated using the Manhattan distance layer and the scaled cosine similarity layer. As an initial step, the models were used with the negative exponential Manhattan distance function, which was adopted from Mueller and Thyagarajan in MaLSTM. Next, experiments were carried out, replacing the negative exponential Manhattan distance function with the scaled cosine similarity function.

I. Manhattan Distance Model

In the development of our Manhattan Distance Model for Sinhala sentence similarity estimation, we employed a comprehensive experimental setup. Generally, vocabulary indices were commonly used to convert sentences into numeric representations. We utilized the word index of our trained word embedding model to convert sentences into numeric representations. The word embedding layer is initialized with 300-dimensional vectors, contributing to the preservation of semantic information. To optimize SiBiLConv model's performance, we chose the ADAM optimizer for parameter updates, leveraging its efficiency in minimizing mean squared error loss functions. The backpropagation algorithm is employed to compute gradients for all parameters during the training process, enhancing the model's ability to learn and adapt. Key hyperparameters were fine-tuned to ensure the model's effectiveness. The word embeddings are fixed at 300 dimensions, providing a robust representation of word semantics. The learning rate is set to 0.001, optimizing the convergence of the model during training. A dropout rate of 0.3 is used to prevent overfitting, and a kernel size of 5 is chosen for each of the 100 convolutional filters. This setup, combined with the mean squared error as the loss function, is tailored to achieve optimal performance in capturing nuanced patterns and relationships within Sinhala sentences.

II. Cosine Distance Model

In the development of our Cosine Distance Model for Sinhala sentence similarity estimation, we seamlessly applied the same robust training strategy employed in the Manhattan Distance Model.

5. Results and Discussion

In this study, six metrics, namely accuracy, mean squared error (MSE), Pearson correlation coefficient, precision, recall, and F1 Score were used to evaluate the proposed SiBiLConv models. Additionally, our self-trained Sinhala word embedding model with cosine similarity, the MaLSTM (Mueller and Thyagarajan, 2016) model and, the GRU (Mikolov et al., 2013) (Ranasinghe et al., 2019) model was utilized to assess these two models. Table 2 summarizes the performance metrics for all methods, while the key trends and observations are detailed.

5.1 Performance of the Word Embedding Model

Table 2 shows that the performance of Word Embedding model delivered the lowest across all metrics, with an accuracy of 68.75%, an F1 score of 0.7220, and a high MSE of 0.095. Despite achieving a relatively strong Precision of 0.8660, the model exhibited a low Recall of 0.6190, indicating its inability to effectively handle sentence similarity tasks.

The reliance of this model on averaging word embeddings results in oversimplified sentence representations. As an example, sentences with the same words but in different orders are treated as equivalent, despite their vastly different meanings. The morphological richness of Sinhala exacerbates this limitation, as word embeddings fail to capture nuances like inflections, prefixes, or suffixes. This simplistic representation limits the model's ability to generalize in complex semantic scenarios.

*Correspondence: ravi@sjp.ac.lk

Table 2. Performance Evaluation of SiBiLConv Models on Validation Dataset*

Method	Accuracy	MSE	Pearson	Precision	Recall	F1
SiBiLConv with Manhattan Distance Model	0.8727	0.0831	**0.7174**	0.9210	0.8428	0.8798
SiBiLConv with Cosine Distance Model	**0.8980**	**0.0281**	0.7122	**0.9407**	**0.924**	**0.9041**
Word Embedding Model	0.6875	0.0950	--	0.8660	0.6190	0.7220
MaLSTM (Manhattan Distance)	0.7899	0.0831	0.6392	0.9296	0.6728	0.7797
LSTM + Cosine Distance	0.8727	0.0554	0.7125	0.9288	0.8336	0.8766

*Performance metrics for the SiBiLConv model using Manhattan Distance and Cosine Similarity layers, compared with baseline models like MaLSTM and Word Embedding models. Metrics include accuracy, mean squared error (MSE), Pearson correlation, precision, recall, and F1 score

5.2 Performance of the SiBiLConv and Baseline Models

The SiBiLConv models demonstrated superior performance over all baseline methods. Among them, the Cosine Distance-based SiBiLConv achieved the highest accuracy of 89.80% and an F1 score of 0.9041, followed closely by the Manhattan Distance variant, which achieved 87.27% accuracy and an F1 score of 0.8798 (Table 2). These results highlight the advantages of the hybrid architecture in capturing sentence-level semantics.

5.2.1 SiBiLConv vs. MaLSTM

The MaLSTM model, employing LSTM layers with Manhattan Distance, achieved moderate performance (accuracy: 78.99%, F1: 0.7797, Table 2). MaLSTM relies solely on LSTM layers, which capture long-range dependencies but fail to effectively extract localized features such as n-grams or short-term dependencies. This limitation is particularly pronounced in morphologically rich languages like Sinhala, where local features significantly influence sentence semantics. In contrast, SiBiLConv's hybrid architecture integrates BiLSTM for global context representation and CNN layers for local feature extraction. This enables a more nuanced understanding of sentence structure, improving performance across all metrics. Additionally, SiBiLConv leverages hybrid similarity metrics (Manhattan and Cosine Distance), offering complementary perspectives on sentence relationships, whereas MaLSTM relies solely on Manhattan Distance.

5.2.2 SiBiLConv vs. LSTM + Cosine Distance

The LSTM + Cosine Distance model performed better than MaLSTM, achieving an accuracy of 87.27% and an F1 score of 0.8766 (Table 2). However, it fell short of SiBiLConv's performance. While the LSTM + Cosine Distance model benefits from the robustness of the Cosine metric, it lacks the hybrid design of SiBiLConv. The incorporation of CNN layers in SiBiLConv complements the BiLSTM, enabling it to capture local linguistic features, such as morphological patterns, which are crucial for Sinhala. This advantage is evident in the significantly lower MSE of the SiBiLConv with Cosine Distance model (0.0281) compared to LSTM + Cosine Distance (0.0554). The lower MSE indicates better alignment with human-annotated similarity scores, highlighting SiBiLConv's ability to generalize effectively.

5.3 SiBiLConv: Manhattan Distance vs. Cosine Distance

Between the two SiBiLConv configurations, the Cosine Distance-based model consistently outperformed its Manhattan counterpart, achieving higher accuracy (89.80% vs. 87.27%) and a better F1 score (0.9041 vs. 0.8798).

*Correspondence: ravi@sjp.ac.lk

- The Cosine Distance metric focuses on vector orientation, making it more robust in high-dimensional spaces and less sensitive to vector magnitude. This property enhances its ability to capture semantic similarity, particularly in complex datasets such as Sinhala sentences.
- Manhattan Distance, while effective, tends to overestimate dissimilarity in high-dimensional spaces, leading to slightly lower performance. Nonetheless, its effectiveness in capturing sentence relationships is evident from its strong Pearson correlation coefficient of 0.7174 (Table 2).
- **Figures 2 and 3** further illustrate the models' convergence behaviors, with the Cosine Distance model demonstrating better alignment between training and validation accuracy, indicative of superior generalization.
- **Figures 4 and 5** provide a detailed visualization of the classification performance of the SiBiLConv models, with the Manhattan Distance model showing robust performance and the Cosine Distance model demonstrating a more balanced prediction distribution.

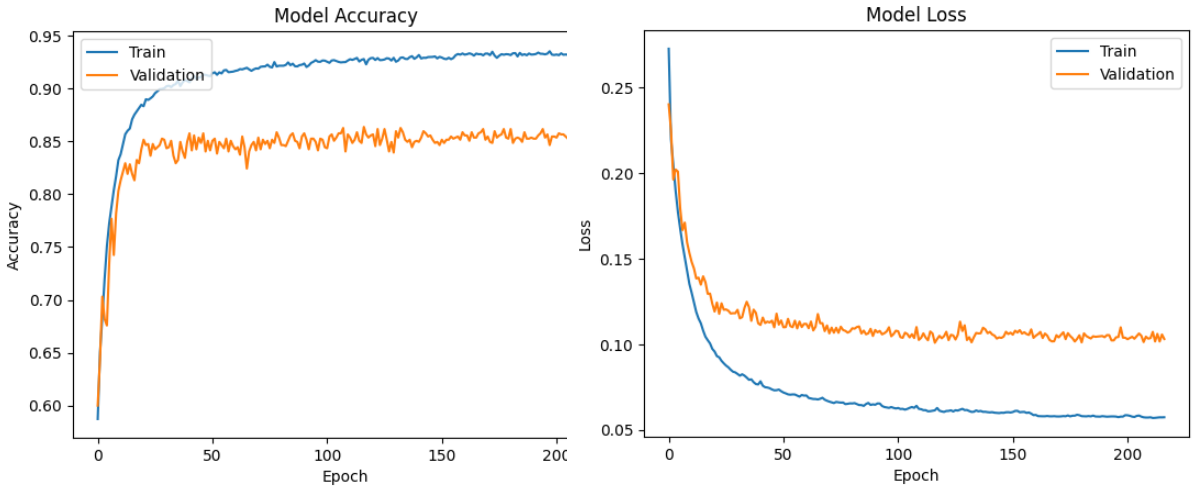


Figure 2: Accuracy and Loss Function of SiBiLConv with Manhattan Distance. Training and validation accuracy and loss for the SiBiLConv model using the Manhattan Distance layer. The model demonstrates rapid convergence, with training accuracy increasing steadily and validation accuracy closely following.

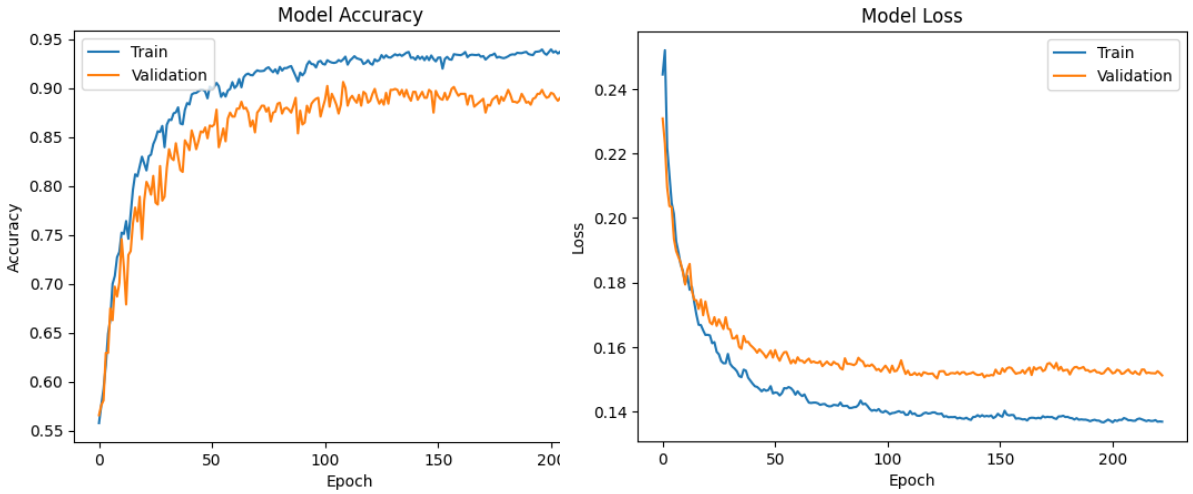


Figure 3: Accuracy and Loss Function of SiBiLConv with Cosine Distance. Training and validation accuracy and loss for the SiBiLConv model using the Cosine Similarity layer. The model exhibits robust performance, with both training and validation accuracy closely aligned, suggesting good generalization.

*Correspondence: ravi@sjp.ac.lk

5.4 Implications for Low-Resource Languages

The strong performance of SiBiLConv underscores its potential for low-resource languages like Sinhala. Its hybrid architecture effectively addresses challenges posed by Sinhala's rich morphology, complex syntax, and high inflection. The CNN layers capture localized patterns and morphological nuances, while BiLSTM layers model long-range dependencies, enabling a comprehensive understanding of sentence semantics.

This study highlights the potential of hybrid architectures like SiBiLConv to advance natural language processing in low-resource settings. By combining local and global feature extraction with advanced similarity metrics, SiBiLConv provides a robust framework for sentence similarity tasks. While the results are promising, the performance of SiBiLConv may vary in other low-resource languages with unique linguistic challenges. Further studies are needed to validate its generalizability across diverse languages.

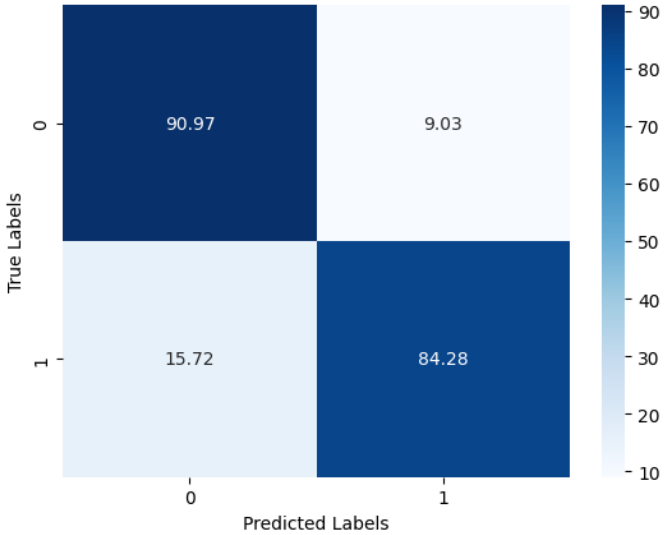


Figure 4: Confusion Matrix for SiBiLConv with Manhattan Distance. Confusion matrix showing the classification performance of the SiBiLConv model with Manhattan Distance for sentence similarity. The matrix indicates the distribution of true positive, true negative, false positive, and false negative predictions.

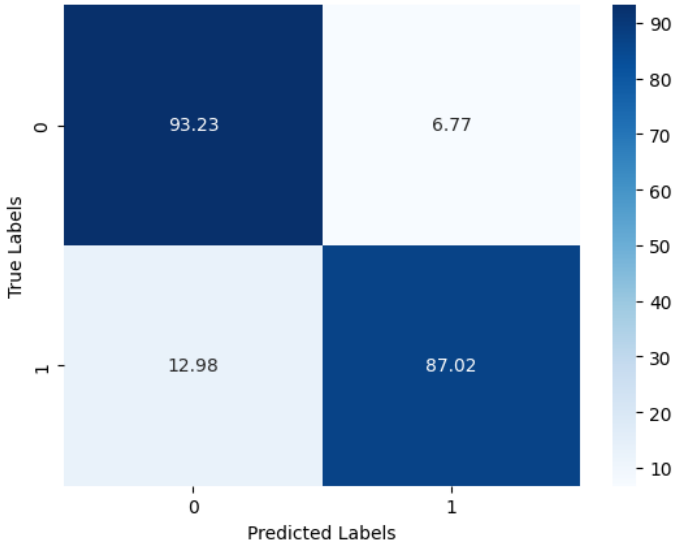


Figure 5: Confusion Matrix for SiBiLConv with Cosine Distance. Confusion matrix showing the classification performance of the SiBiLConv model with Cosine Distance for sentence similarity. The matrix reveals the model's ability to distinguish between similar and dissimilar sentences, with a balanced prediction distribution.

*Correspondence: ravi@sjp.ac.lk

6. Conclusions

This study introduced SiBiLConv, a Siamese hybrid network designed to improve sentence similarity measurements in the Sinhala language. By integrating BiLSTM and CNN layers with advanced similarity metrics (Manhattan and Cosine Distance), the model effectively captures both local linguistic features and global semantic relationships. The SiBiLConv models consistently demonstrated strong performance, surpassing baseline methods, and showing promise for addressing the challenges of morphologically rich, low-resource languages.

However, the study has certain limitations. The evaluation was restricted to a single language, raising questions about the model's adaptability to other languages or multilingual scenarios. The dataset used also presented challenges, particularly in its reliance on human-annotated similarity scores, which can be subjective and inconsistent, yet inevitable. Annotators may evaluate similarity based on individual words rather than holistic sentence meanings or may interpret sentences differently depending on word order and phrasing, potentially impacting the reliability of the ground truth. Additionally, the inherent characteristics of the similarity metrics used such as the overestimation tendencies of Manhattan Distance and the potential sensitivity of Cosine Distance to noise highlight areas for further exploration to optimize their application.

The broader implications of this work emphasize the potential of hybrid architectures like SiBiLConv for advancing natural language processing in low-resource settings. One notable strength of SiBiLConv is its suitability for low-computational-resource environments, where transformer-based models may be infeasible due to their higher demand of computational power. However, within the context of computational power being readily available, integrating transformer-based architectures may further enhance SiBiLConv's scalability and contextual understanding. By addressing linguistic complexities and incorporating both local and global semantic features, SiBiLConv offers a promising framework for other related tasks such as machine translation, question answering, and semantic search.

Future research could focus on refining annotation processes to reduce subjectivity, developing more diverse and representative datasets, and extending the model to cross-lingual and multilingual tasks. Additionally, further exploration of hybrid architectures with transformer-based enhancements could unlock new possibilities for complex NLP applications. SiBiLConv serves as an important step toward bridging the gap in NLP resources for underrepresented languages, providing a foundation for broader applications and further advancements in the field.

References

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching Word Vectors with Subword Information.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., Kurzweil, R., 2018. Universal Sentence Encoder. https://doi.org/10.48550/arXiv.1803.11175

de Silva, N., 2023. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. https://doi.org/10.48550/arXiv.1906.02358

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Dhananjaya, V., Demotte, P., Ranathunga, S., Jayasena, S., 2022. BERTifying Sinhala - A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification, in: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference. Presented at the LREC 2022, European Language Resources Association, Marseille, France, pp. 7377–7385.

Edo-Osagie, O., De La Iglesia, B., 2019. Attention-Based Recurrent Neural Networks (RNNs) for Short Text Classification: An Application in Public Health Monitoring: International Work-

*Correspondence: ravi@sjp.ac.lk

Conference on Artificial Neural Networks, in: Lake, I., Edeghere, O. (Eds.), . https://doi.org/10.1007/978-3-030-20521-8_73

Gao, T., Yao, X., Chen, D., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2021, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

He, H., Gimpel, K., Lin, J., 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks, in: Màrquez, L., Callison-Burch, C., Su, J. (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2015, Association for Computational Linguistics, Lisbon, Portugal, pp. 1576–1586. https://doi.org/10.18653/v1/D15-1181

Ichida, A.Y., Meneguzzi, F., Ruiz, D.D., 2018. Measuring Semantic Similarity Between Sentences Using A Siamese Neural Network, in: 2018 International Joint Conference on Neural Networks (IJCNN). Presented at the 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, Rio de Janeiro, pp. 1–7. https://doi.org/10.1109/IJCNN.2018.8489433

Ji, M., Zhang, X., 2022. A Short Text Similarity Calculation Method Combining Semantic and Headword Attention Mechanism. Scientific Programming 2022, e8252492. https://doi.org/10.1155/2022/8252492

John Snow Labs, 2024. spark-nlp. Available at: https://github.com/JohnSnowLabs/spark-nlp [Accessed on 09-10-2024].

Kadupitiya, J., Ranathunga, S., Dias, G., 2016. Sinhala short sentence similarity calculation using corpus-based and knowledge-based similarity measures.

Lakmal, D., Ranathunga, S., Peramuna, S., Herath, I., 2020. Word Embedding Evaluation for Sinhala, in: Proceedings of the Twelfth Language Resources and Evaluation Conference. Presented at the LREC 2020, European Language Resources Association, Marseille, France, pp. 1874–1881.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791

Liu, J., Ma, H., Xie, X., Cheng, J., 2022. Short Text Classification for Faults Information of Secondary Equipment Based on Convolutional Neural Networks. Energies 15, 2400. https://doi.org/10.3390/en15072400

Manamini, S.A.P.M., Ahamed, A.F., Rajapakshe, R.A.E.C., Reemal, A., Jayasena, S., Dias, G., Ranathunga, S., 2016. Ananya - a Named-Entity-Recognition (NER) system for Sinhala language. https://doi.org/10.1109/MERCon.2016.7480111

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., 2014. A SICK cure for the evaluation of compositional distributional semantic models, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Presented at the LREC 2014, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 216–223.

Meng, F., Wang, W., Wang, J., 2021. Research on Short Text Similarity Calculation Method for Power Intelligent Question Answering, in: 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN). Presented at the 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 91–95. https://doi.org/10.1109/CICN51697.2021.9574692

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

Mueller, J., Thyagarajan, A., 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. AAAI 30. https://doi.org/10.1609/aaai.v30i1.10350

NLP Centre, University of Moratuwa, 2021. Tamil-Sinhala short sentence similarity deep learning.

*Correspondence: ravi@sjp.ac.lk

Ranasinghe, T., Orasan, C., Mitkov, R., 2019. Semantic Textual Similarity with Siamese Neural Networks, in: Mitkov, R., Angelova, G. (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). Presented at the RANLP 2019, INCOMA Ltd., Varna, Bulgaria, pp. 1004–1011. https://doi.org/10.26615/978-954-452-056-4_116

Ranathunga, S., Liyanage, I.U., 2021. Sentiment Analysis of Sinhala News Comments. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20, 59:1-59:23. https://doi.org/10.1145/3445035

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I., 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. https://doi.org/10.48550/arXiv.2104.08663

Wang, H., Tian, K., Wu, Z., Wang, L., 2020. A Short Text Classification Method Based on Convolutional Neural Network and Semantic Extension. International Journal of Computational Intelligence Systems 14, 367–375. https://doi.org/10.2991/ijcis.d.201207.001

Wang, S.-H., Chen, Z.-C., Shi, J., Chuang, M.-T., Lin, G.-T., Huang, K.-P., Harwath, D., Li, S.-W., Lee, H., 2024. How to Learn a New Language? An Efficient Solution for Self-Supervised Learning Models Unseen Languages Adaption in Low-Resource Scenario. https://doi.org/10.48550/arXiv.2411.18217

Weeraprameshwara, G., Jayawickrama, V., de Silva, N., Wijeratne, Y., 2022. Sinhala Sentence Embedding: A Two-Tiered Structure for Low-Resource Languages, in: Dita, S., Trillanes, A., Lucas, R.I. (Eds.), Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation. Presented at the PACLIC 2022, Association for Computational Linguistics, Manila, Philippines, pp. 325–336.

Wijaya, M.C., 2021. Automatic short answer grading system in Indonesian language using BERT machine learning. Revue d'Intelligence Artificielle 35, 503–509. https://doi.org/10.18280/ria.350609

Yoo, Y., Heo, T.-S., Park, Y., Kim, K., 2021. A Novel Hybrid Methodology of Measuring Sentence Similarity. Symmetry 13, 1442. https://doi.org/10.3390/sym13081442

Zhu, Z., He, Z., Tang, Z., Wang, B., Chen, W., 2018. A Semantic Similarity Computing Model based on Siamese Network for Duplicate Questions Identification., in: CCKS Tasks. pp. 44–51.

**Correspondence: ravi@sjp.ac.lk*