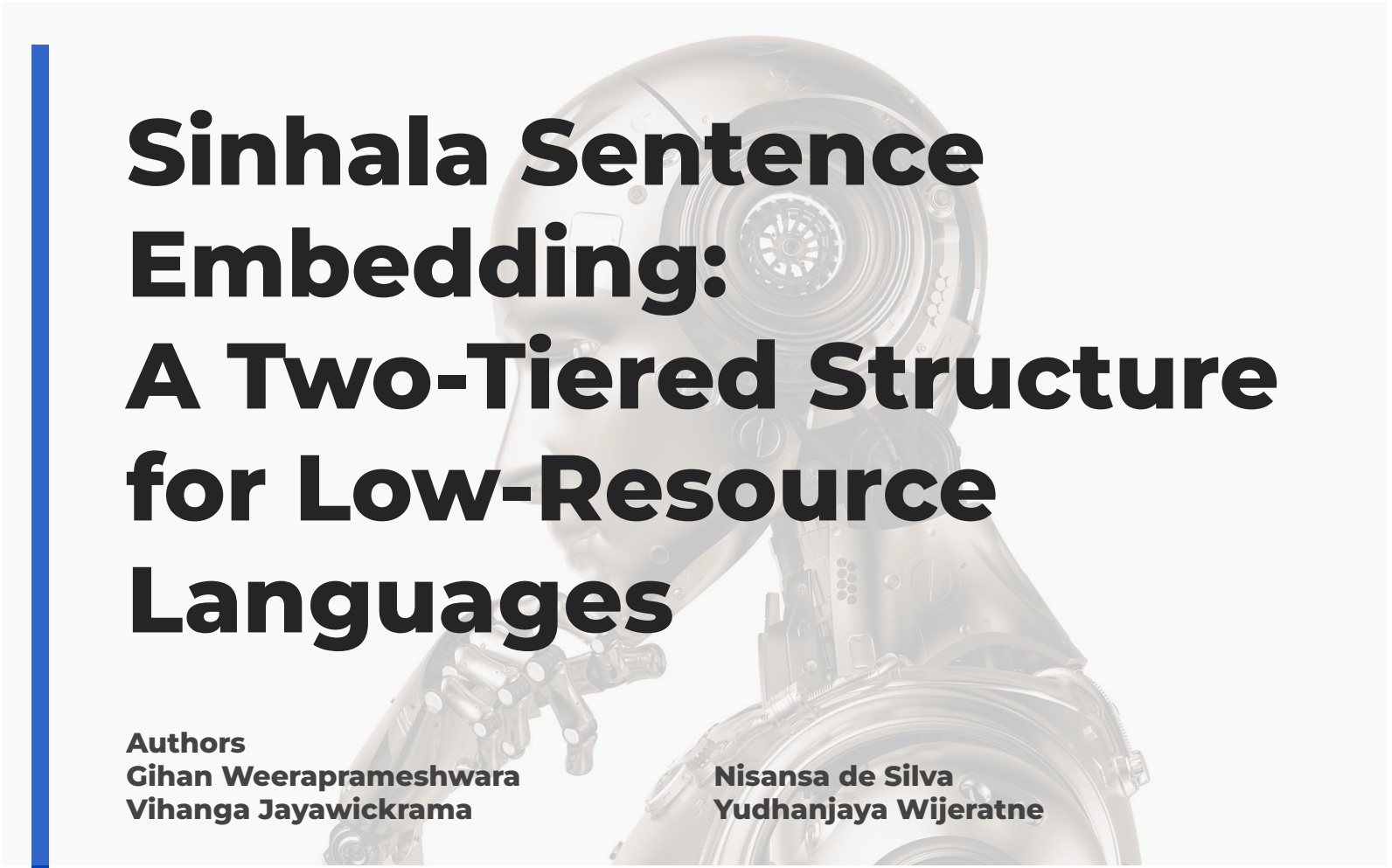
Sinhala Sentence Embedding: A Two-Tiered Structure for Low-Resource Languages

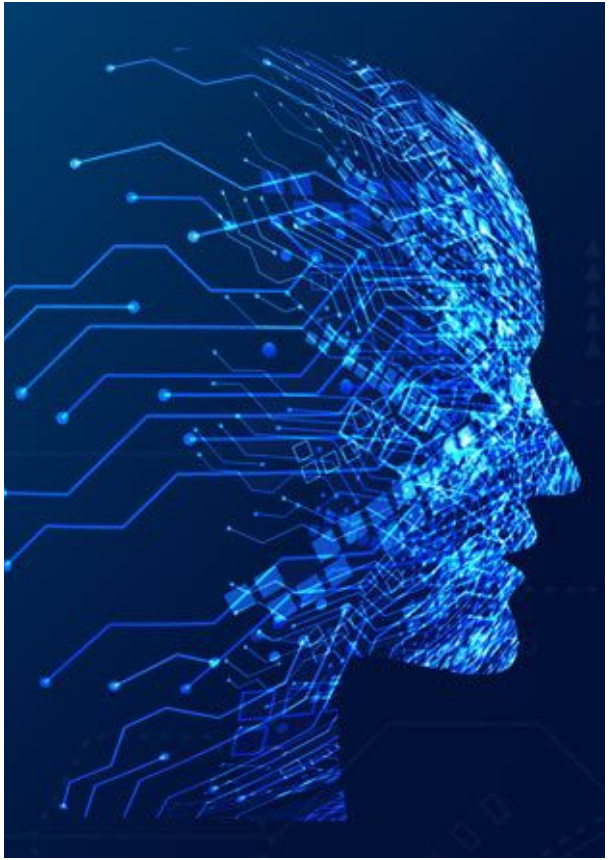


Authors
Gihan Weeraprameshwara
Vihanga Jayawickrama
Nisansa de Silva
Yudhanjaya Wijeratne

Outline

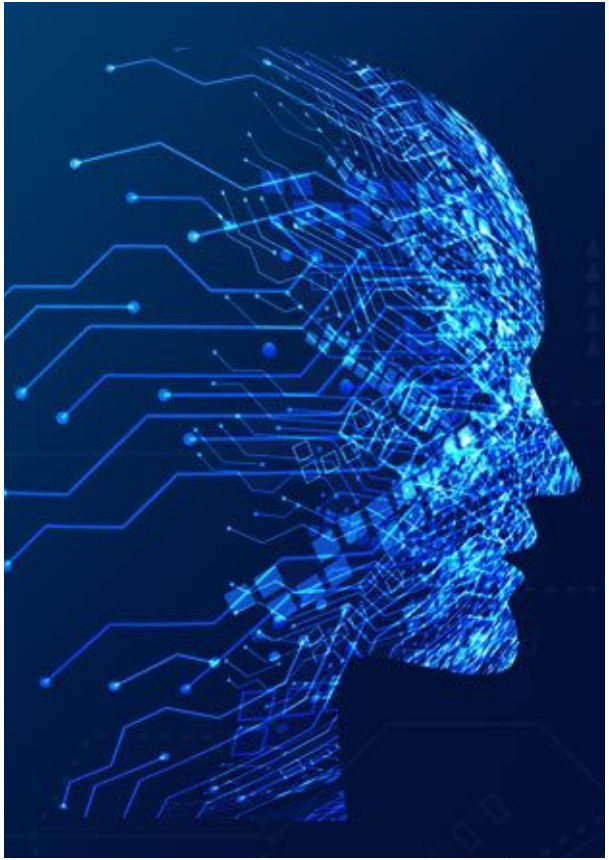
1. Problem
2. Solution
3. Related Work
4. Methodology
5. Results
6. Conclusion
7. Future Work

1. Problem



An efficient embedding structure for Sinhala ?

2. Solution



A two tiered embedding structure using word and sentence embeddings

Sinhala Colloquial Text

- Sinhala and English codemixed
- Any other language is excluded

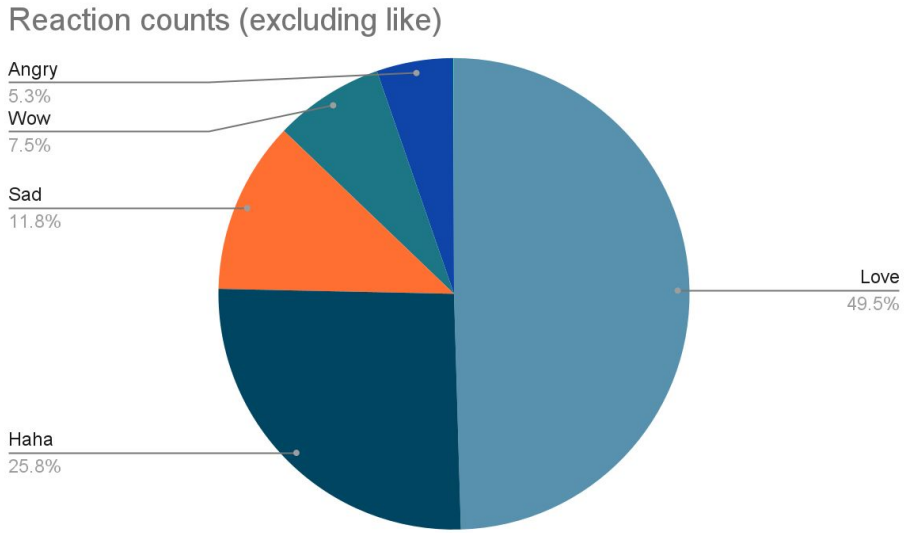
Using a Facebook dataset [1]

Using Facebook reactions as the annotations

[1] Y. Wijeratne and N. de Silva, “Sinhala language corpora and stop words from a decade of sri lankan facebook,”arXiv preprint arXiv:2007.07884, 2020.

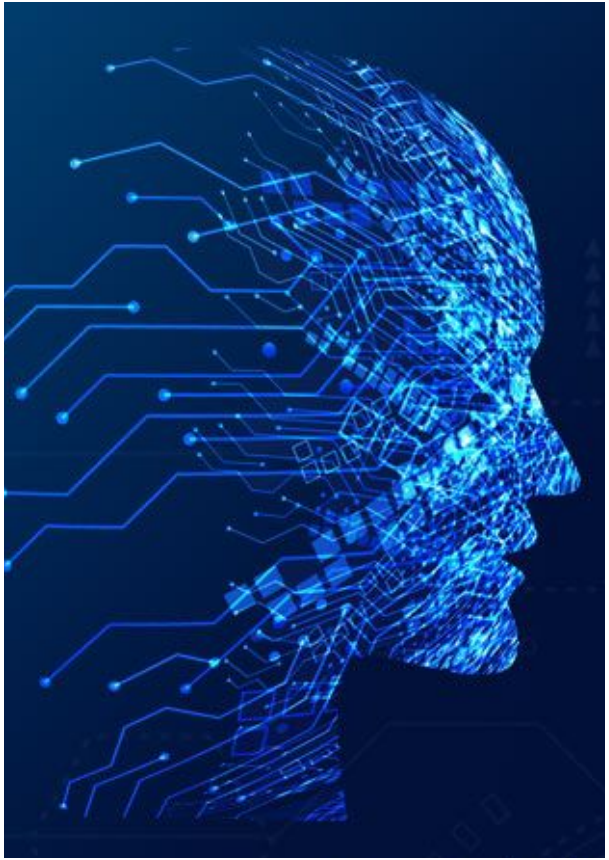
Dataset

- Developed by Yudhanjaya Wijerathne and Nisansa de Silva [1]
- Contains 1.8 million Facebook posts spanning over a decade from different sources.
- Over 540 million user reactions
- 526,732 data rows after preprocessing steps



[1] Y. Wijeratne and N. de Silva, “Sinhala language corpora and stop words from a decade of sri lankan facebook,”arXiv preprint arXiv:2007.07884, 2020.

3. Related Work



- **Word embeddings**
  - FastText embeddings developed by P. Bojanowski et al. [2] and A. Joulin et al [3].
  - Word2Vec embeddings developed by T. Mikolov et al. [4]
  - Glove (Global Vectors) embeddings developed by J. Pennington et al. [5]

- **Sentence embeddings**
  - Seq2Seq model introduced by I. Sutskeve [6]
  - The modified version of the Seq2Seq model by K. Cho et al. [7] with the RNN units

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information,"Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.
[3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification,"arXiv preprint arXiv:1607.01759, 2016.
[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space,"arXiv preprint arXiv:1301.3781, 2013.
[5] 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), pp. 1532–1543, 2014.
[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems,pp. 3104–3112, 2014.
[7] K. Cho, B. Van Merri ̈enboer, C. Gulcehre, D. Bahdanau, F. Bougares,H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation,"arXiv preprint arXiv:1406.1078, 2014.

- **Hyperbolic embeddings**
  - The work done by Q. Lu et al. [8] to understand the applications of hyperbolic embeddings by applying the concept in the medical field to improve state-of-the-art models
  - Poincar´e embeddings introduced by M. Nickel et al. [9] for learning hierarchical representations in hyperbolic space
  - Skip gram word embeddings in hyperbolic space introduced by M. Leimeister et al. [10]
  - Reinforcing the methods introduced by M. Nickel by the work of B. Dhingra et al. [11] on embedding text on hyperbolic space

[8] Q. Lu, N. de Silva, S. Kafle, J. Cao, D. Dou, T. H. Nguyen, P. Sen,B. Hailpern, B. Reinwald, and Y. Li, “Learning electronic health records through hyperbolic embedding of medical ontologies,” inProceedings of the10th ACM International Conference on Bioinformatics, Computational Bi-ology and Health Informatics, pp. 338–346, 2019
[9] M. Nickel and D. Kiela, “Poincar´e embeddings for learning hierarchical representations,”Advances in neural information processing systems, vol. 30,pp. 6338–6347, 2017.
[10] M. Leimeister and B. J. Wilson, “Skip-gram word embeddings in hyperbolic space,”arXiv preprint arXiv:1809.01498, 2018.
[11] B. Dhingra, C. J. Shallue, M. Norouzi, A. M. Dai, and G. E. Dahl, “Embedding text in hyperbolic spaces,”arXiv preprint arXiv:1806.04313, 2018.

- **Sinhala NLP research**
  - Survey on publicly available sinhala natural language processing tools and research by N. de Silva [12] to identify the advancements in Sinhala NLP
  - The model collection discussed by Senevirathne et al. [13] on sentiment analysis for sinhala language using deep learning techniques and the news comment dataset consists of 15000 data items
  - The research work of Jayawickrama et al.[14] to understand the Facebook dataset[1] and the sentiment mapping of Facebook reactions
  - The annotation method and the deep learning model used in this paper is taken from the work of G. Weeraprameshwara et al [15]

[12] N. de Silva, "Survey on publicly available sinhala natural language processing tools and research,"arXiv preprint arXiv:1906.02358, 2019.
[13] L. Senevirathne, P. Demotte, B. Karunanayake, U. Munasinghe, andS. Ranathunga, "Sentiment analysis for sinhala language using deep learning techniques," 2020.
[14] Vihanga Jayawickrama, Gihan Weeraprameshwara, Nisansa de Silva, and Yudhanjaya Wijeratne. 2021. Seeking sinhala sentiment: Predicting facebook reactions of sinhala posts. In 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter), pages 177–182. IEEE.
[1] Y. Wijeratne and N. de Silva, "Sinhala language corpora and stopwords from a decade of sri lankan facebook," arXiv preprint arXiv:2007.07884, 2020
[15] Gihan Weeraprameshwara, Vihanga Jayawickrama, Nisansa de Silva, and Yudhanjaya Wijeratne. 2022. Sentiment analysis with deep learning models: a comparative study on a decade of sinhala language facebook data. In 2022 The 3rd International Conference on Artificial Intelligence in Electronics Engineering, pages 16–22.

Reaction Mapping

- Positive reactions
  - Love, Wow
- Negative reactions
  - Sad, Angry
- Neglected reactions
  - Like, Thankful, Haha, Care

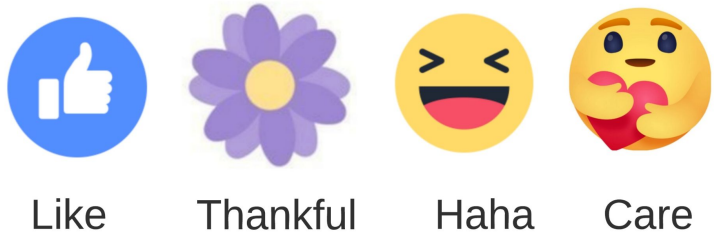
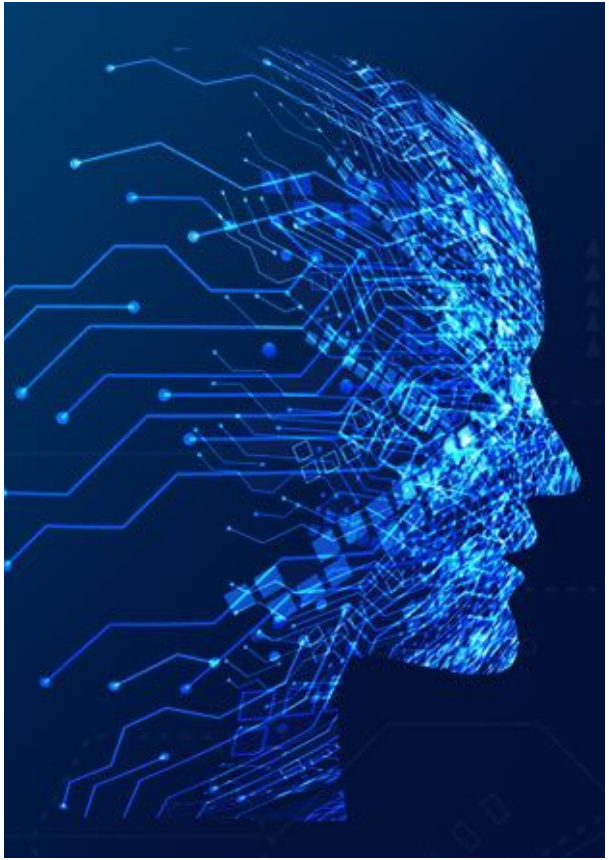
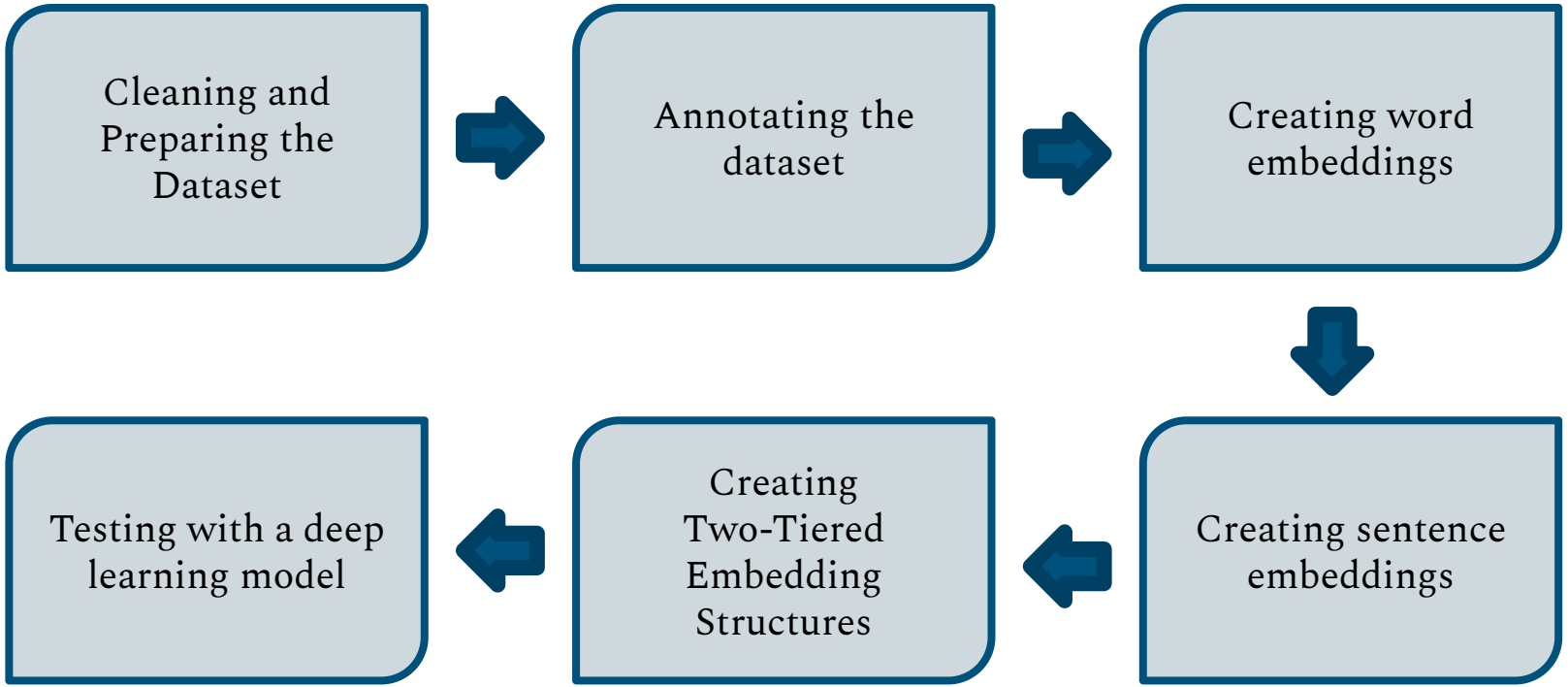


Figure: Reaction Annotation

4. Methodology



Walkthrough



Developing word embeddings for Facebook posts:

- FastText [16-17]
- Glove [18]
- Word2Vec [19]
- Poincar´e [20]

[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,”Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.
[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,”arXiv preprint arXiv:1607.01759, 2016.
[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,”arXiv preprint arXiv:1301.3781, 2013.
[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), pp. 1532–1543, 2014.
[20] M. Nickel and D. Kiela, “Poincar´e embeddings for learning hierarchical representations,”Advances in neural information processing systems, vol. 30,pp. 6338–6347, 2017.

Developing sentence embeddings for Facebook posts:

- MaxPooling
- MinPooling
- AveragePooling
- Seq2Seq model [21]
  - with GRU [22] or LSTM [23] units
  - with or without an attention layer [24]

[21] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Advances in neural information processing systems,pp. 3104–3112, 2014.
[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,”arXiv preprint arXiv:1412.3555, 2014.
[23] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,”IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, pp. 5998–6008, 2017

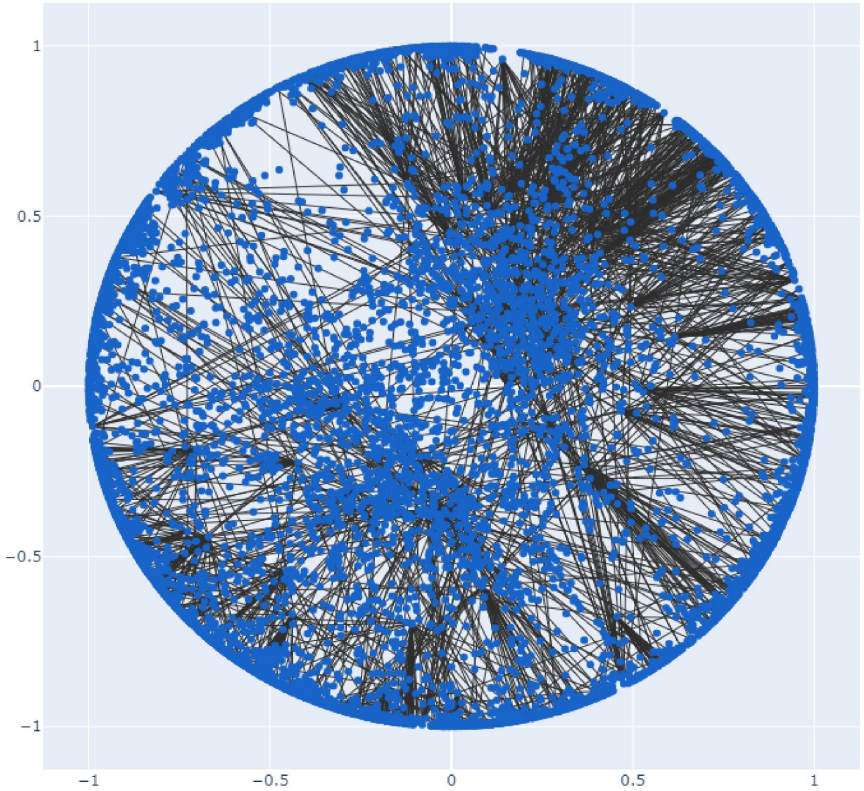


Figure: Poincar´e embeddings illustrations

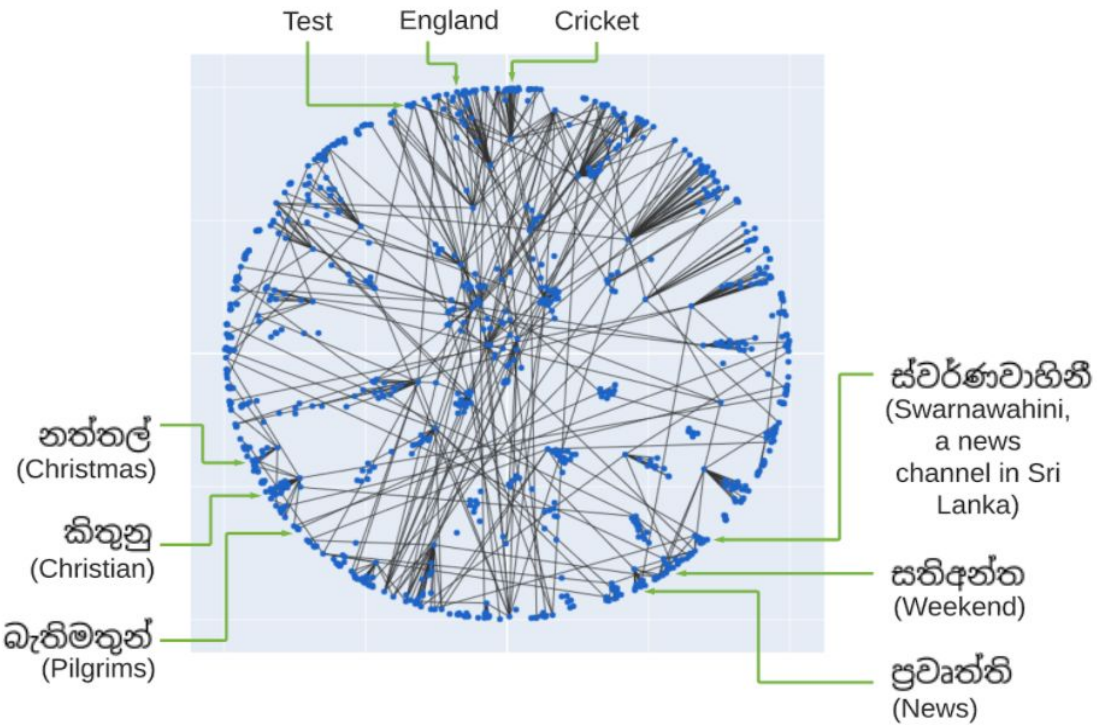


Figure: Poincar´e embeddings illustrations

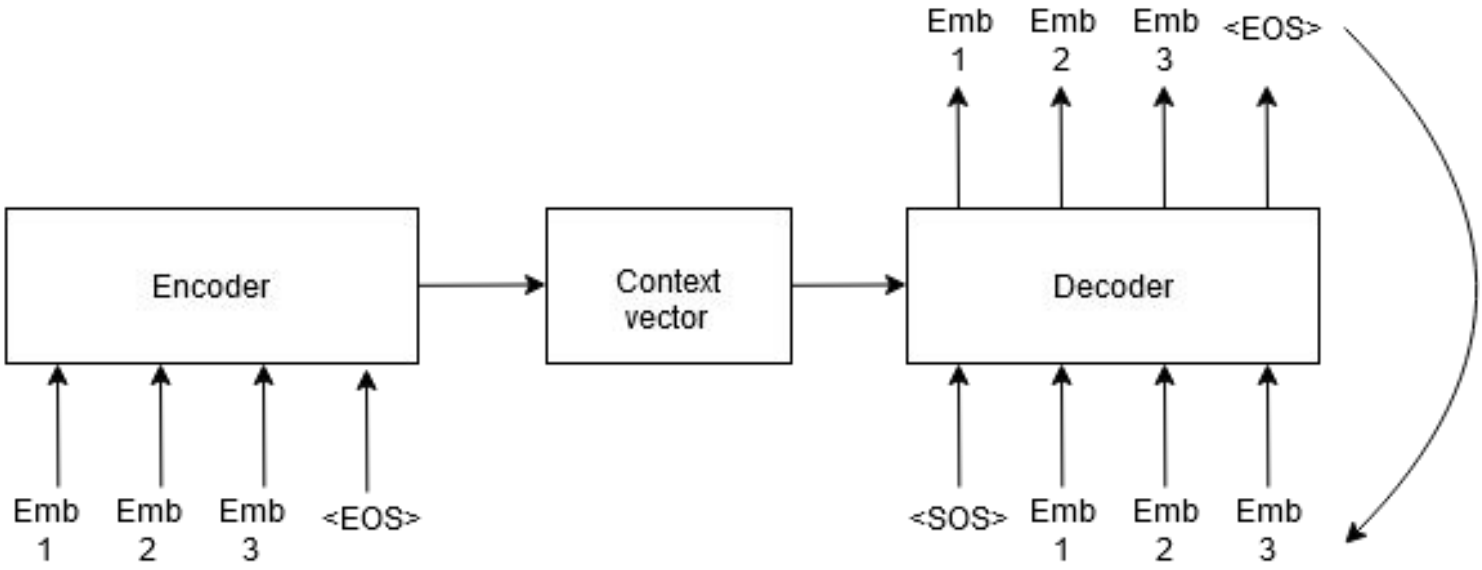
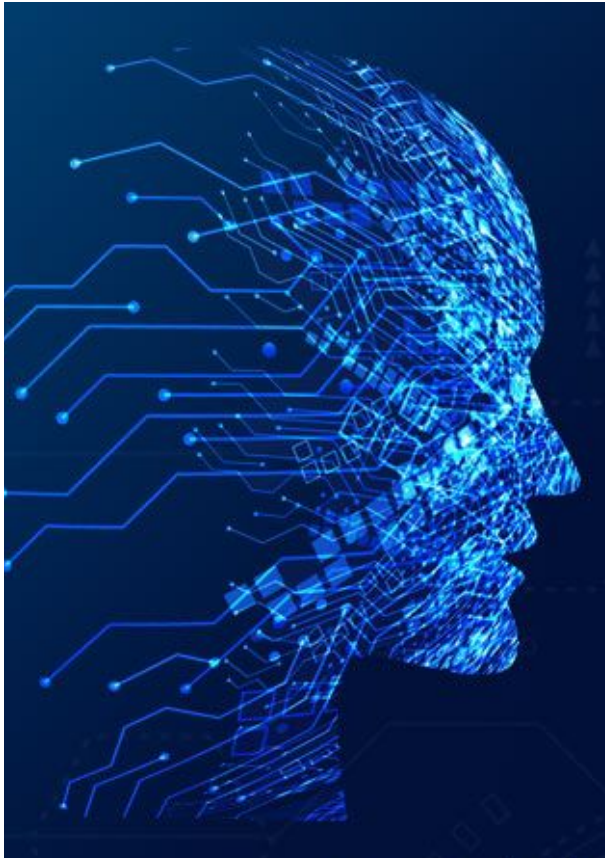


Figure: Two tier embeddings with Seq2seq model

5. Results



Word Embedding	F1 Score
fastText (Sinhala News comments) [13]	81.37
fastText [16-17]	**83.76**
Glove [18]	82.28
Word2Vec [19]	83.56
Hyperbolic [20]	82.84

[13] L. Senevirathne, P. Demotte, B. Karunanayake, U. Munasinghe, andS. Ranathunga, "Sentiment analysis for sinhala language using deep learning techniques," 2020.
[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information,"Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.
[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification,"arXiv preprint arXiv:1607.01759, 2016.
[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space,"arXiv preprint arXiv:1301.3781, 2013.
[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), pp. 1532–1543, 2014.
[20] M. Nickel and D. Kiela, "Poincar´e embeddings for learning hierarchical representations,"Advances in neural information processing systems, vol. 30,pp. 6338–6347, 2017.

Word Embedding	Sentence Embedding	F1 Score
Word2Vec	Avg Pooling	87.01
	Seq2seq [21] GRU [22]	85.75
	Seq2seq [21] GRU [22] + Attention [24]	**87.29**
	Seq2seq [21] LSTM [23]	86.01
	Seq2seq [21] LSTM [23] + Attention [24]	86.53
Glove	Max Pooling	**86.22**
	Seq2seq [21] GRU [22]	85.16
	Seq2seq [21] GRU [22] + Attention [24]	85.12
	Seq2seq [21] LSTM [23]	85.16
	Seq2seq [21] LSTM [23] + Attention [24]	85.12

[21] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems,pp. 3104–3112, 2014.
[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling,"arXiv preprint arXiv:1412.3555, 2014.
[23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks,"IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, pp. 5998–6008, 2017

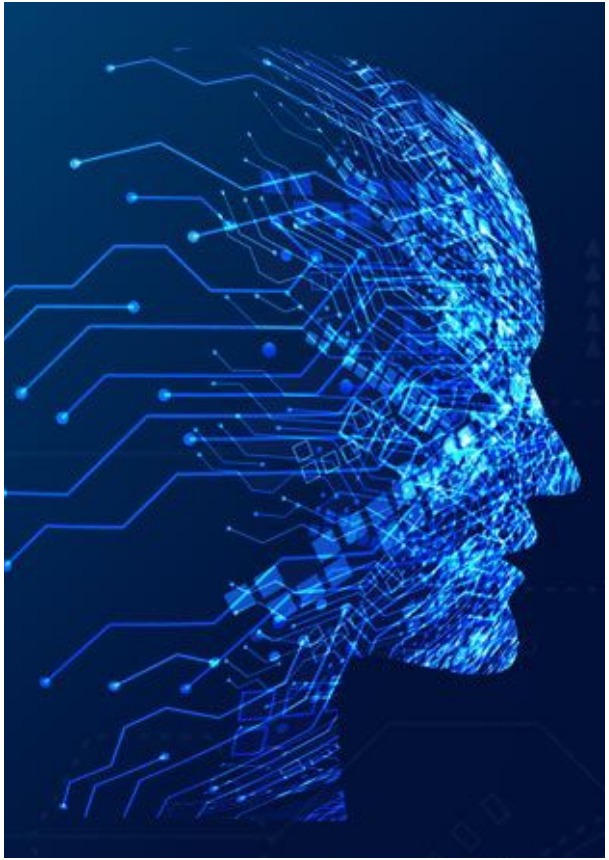
Word Embedding	Sentence Embedding	F1 Score
fastText	Avg Pooling	87.93
	Seq2seq [21] GRU [22]	86.23
	Seq2seq [21] GRU [22] + Attention [24]	**88.04**
	Seq2seq [21] LSTM [23]	86.60
	Seq2seq [21] LSTM [23] + Attention [24]	87.72
Hyperbolic	Max Pooling	85.77
	Seq2seq [21] GRU [22]	86.13
	Seq2seq [21] GRU [22] + Attention [24]	**86.54**
	Seq2seq [21] LSTM [23]	85.81
	Seq2seq [21] LSTM [23] + Attention [24]	86.30

[21] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems,pp. 3104–3112, 2014.
[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling,"arXiv preprint arXiv:1412.3555, 2014.
[23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks,"IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, pp. 5998–6008, 2017

Significance of the Outcome

- **Embeddings for Sinhala language**
  - Uses a significantly large dataset
  - Tests a more granular structure with word and sentence embeddings
  - Considers multiple combinations of existing tools
  - Introduces hyperbolic embeddings for sentiment data in Sinhala for the first time

6. Conclusions



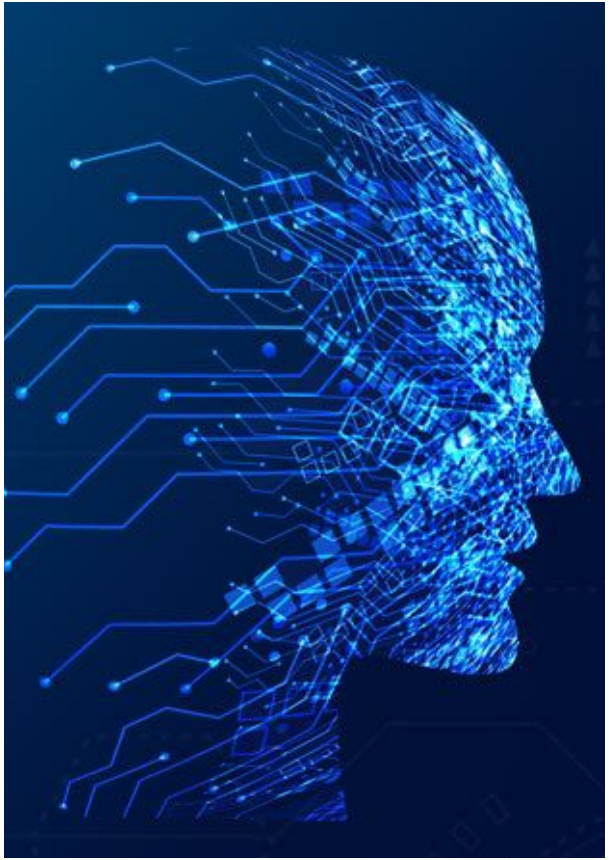
- **Introducing a two tier architecture is more efficient than relying on simple word embedding architecture**

- **fastText word embedding combined with sentencer embedding structure of Seq2seq model with GRU and attention layers can be identified as the best performing model**

- **Hyperbolic embedding does not surpass fastText and Word2Vec embeddings due to not having a proper parser**

- **Glove embeddings lacks the performance due to not having a proper pre-trained Glove model for Sinhala language**

7. Future Work



- **A parser that is customized for Sinhala language which can help for hyperbolic embedding**
- **Developing a well trained Glove model for Sinhala language**
- **Pure Sinhala embeddings instead of Sinhala-English codemixed embedding**
- **Testing the embeddings with different deep learning models an the use of transformer models**
- **Testing the embeddings with external databases**

References

[1] Y. Wijeratne and N. de Silva, “Sinhala language corpora and stopwords from a decade of sri lankan facebook,” arXiv preprint arXiv:2007.07884, 2020

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,”Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.

[3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,”arXiv preprint arXiv:1607.01759, 2016.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,”arXiv preprint arXiv:1301.3781, 2013.

[5] 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), pp. 1532–1543, 2014.

[6] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Advances in neural information processing systems,pp. 3104–3112, 2014.

[7] K. Cho, B. Van Merri ̈enboer, C. Gulcehre, D. Bahdanau, F. Bougares,H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,”arXiv preprint arXiv:1406.1078, 2014.

[8] Q. Lu, N. de Silva, S. Kafle, J. Cao, D. Dou, T. H. Nguyen, P. Sen,B. Hailpern, B. Reinwald, and Y. Li, “Learning electronic health records through hyperbolic embedding of medical ontologies,” inProceedings of the10th ACM International Conference on Bioinformatics, Computational Bi-ology and Health Informatics, pp. 338–346, 2019

[9] M. Nickel and D. Kiela, “Poincar ́e embeddings for learning hierarchical representations,”Advances in neural information processing systems, vol. 30,pp. 6338–6347, 2017.

[10] M. Leimeister and B. J. Wilson, “Skip-gram word embeddings in hyperbolic space,”arXiv preprint arXiv:1809.01498, 2018.

[11] B. Dhingra, C. J. Shallue, M. Norouzi, A. M. Dai, and G. E. Dahl, “Embedding text in hyperbolic spaces,”arXiv preprint arXiv:1806.04313, 2018.

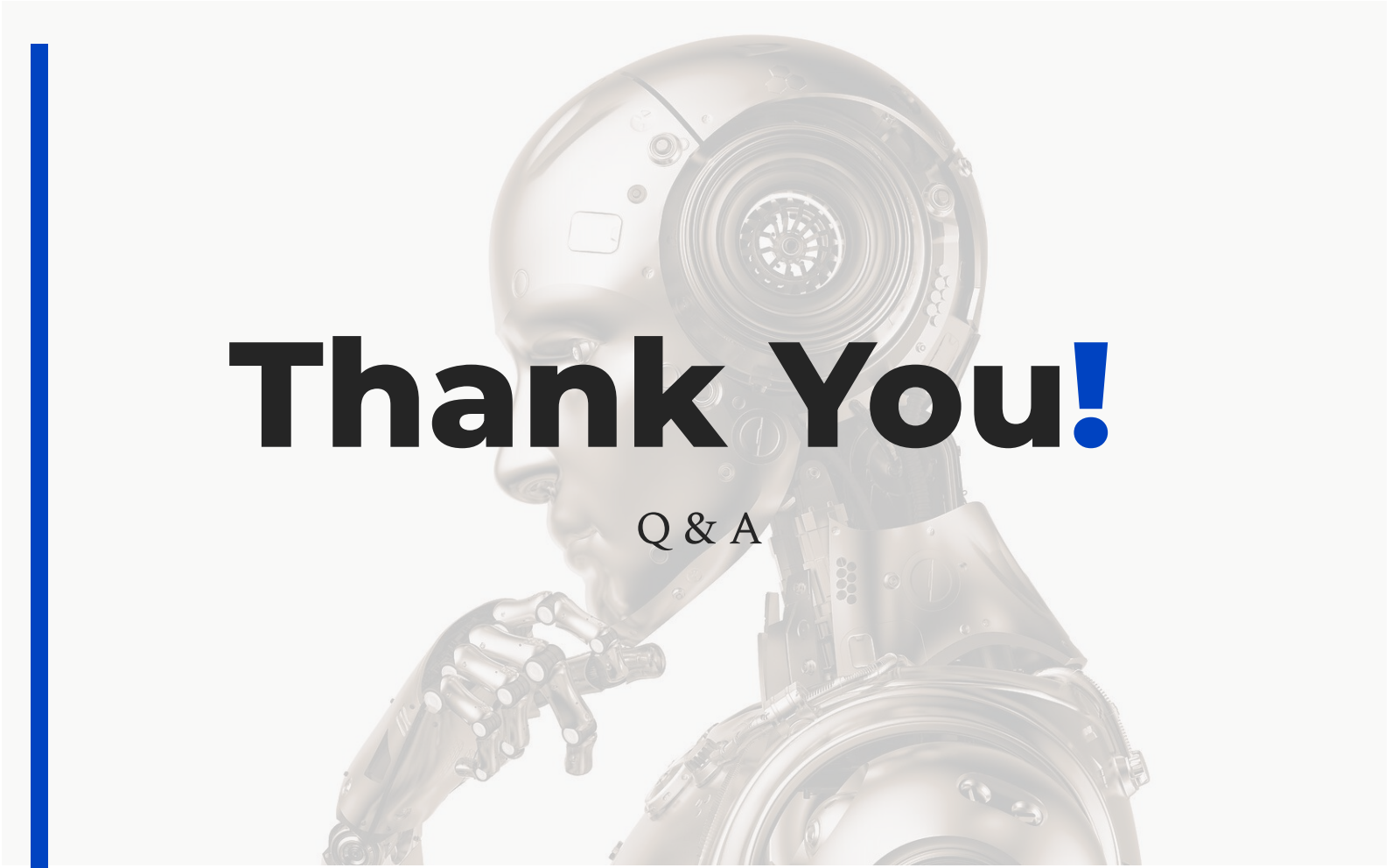
[12] N. de Silva, “Survey on publicly available sinhala natural language processing tools and research,”arXiv preprint arXiv:1906.02358, 2019.

[13] L. Senevirathne, P. Demotte, B. Karunanayake, U. Munasinghe, andS. Ranathunga, “Sentiment analysis for sinhala language using deep learning techniques,” 2020.

[14] Vihanga Jayawickrama, Gihan Weeraprameshwara, Nisansa de Silva, and Yudhanjaya Wijeratne. 2021. Seeking sinhala sentiment: Predicting facebook reactions of sinhala posts. In 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter), pages 177–182. IEEE.

[15] Gihan Weeraprameshwara, Vihanga Jayawickrama, Nisansa de Silva, and Yudhanjaya Wijeratne. 2022. Sentiment analysis with deep learning models: a comparative study on a decade of sinhala language facebook data. In 2022 The 3rd International Conference on Artificial Intelligence in Electronics Engineering, pages 16–22.
[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information,"Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.
[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification,"arXiv preprint arXiv:1607.01759, 2016.
[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space,"arXiv preprint arXiv:1301.3781, 2013.
[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), pp. 1532–1543, 2014.
[20] M. Nickel and D. Kiela, "Poincar ´e embeddings for learning hierarchical representations,"Advances in neural information processing systems, vol. 30,pp. 6338–6347, 2017.
[21] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems,pp. 3104–3112, 2014.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,”arXiv preprint arXiv:1412.3555, 2014.
[23] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,”IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, pp. 5998–6008, 2017



Thank You!

Q & A