A BERT-based Idiom Detection Model

Gihan Gamage
Centre for Data Analytics and Cognition (CDAC)
La Trobe University
Melbourne, Australia
g.gamage@latrobe.edu.au

Daswin De Silva
Centre for Data Analytics and Cognition (CDAC)
La Trobe University
Melbourne, Australia
d.desilva@latrobe.edu.au

Achini Adikari
Centre for Data Analytics and Cognition (CDAC)
La Trobe University
Melbourne, Australia
a.adikari@latrobe.edu.au

Damminda Alahakoon
Centre for Data Analytics and Cognition (CDAC)
La Trobe University
Melbourne, Australia
d.alahakoon@latrobe.edu.au

2022 15th International Conference on Human System Interaction (HSI) | 978-1-6654-6822-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/HSI55341.2022.9869485

***Abstract*—Idioms are figures of speech that contradict the principle of compositionality. This disposition of idioms can misdirect Natural Language Processing (NLP) techniques, which mostly focus on the literal meaning of terms. In this paper, we propose a novel idiom detection model that distinguishes between literal and idiomatic expressions. It utilizes a token classification approach to fine-tune BERT(Bidirectional Encoder Representations from Transformers). It is empirically evaluated on four idiom datasets, yielding an accuracy of more than 0.94. This model adds to the robustness and diversity of NLP techniques available to process and understand increasing magnitudes of free-form text and speech. Furthermore, the social value of this model is in enabling non-native speakers to comprehend the nuances of a foreign language.**

Keywords—Idioms, Natural Language Processing, BERT

I. INTRODUCTION

Idioms are multi-word expressions derived and established by convention and cultural experience. Although a limited subset of language, idioms are frequently used in formal and informal communication for creative, subtle, simple, or stimulating effect. Idioms are distinguished from other figures of speech as the meaning cannot be derived from its individual terms. But the non-literal meaning of the idiom is familiar to the native speakers. For instance, "to have kittens when submitting this paper" means the authors are nervous about the outcome, a highly figurative connotation beyond the literal meaning.

Although idioms enhance the expressiveness of language, an idiomatic expression's contradiction of compositionality means these phrases present a unique challenge in Natural Language Processing/Understanding (NLP/U) settings, from simple techniques such as keyword extraction, topic extraction, sentiment analysis, information retrieval, to complex techniques such as text summarization, machine comprehension, machine translation, question answering, and emotion analysis [1]–[4]. For instance, the example *"The movie's final scene left me with a lump in my throat"* contains strong negativity which literal sentiment classifiers often fail to capture.

Recent literature reports several categories of methods for idiom detection and classification. The most prominent are lexicon based methods that consider idioms to be frozen phrases that can be identified using simple string-matching techniques, such as [5] and [6]. However, idioms can exhibit syntactic changes such as verb and tense alteration that does not impact its figurative meaning. This allows the same idiom to appear in multiple forms [7]. In such cases, simple string matching fails and in response regular expressions [8] and local grammar based methods have been proposed [9], [10]. These two types of methods are primarily affected by the time and effort required of human experts to produce idiom dictionaries and local grammar. Furthermore, detecting diverse forms of idioms and new idiom discovery is infeasible because the classification is limited to the local grammar or regular expressions generated by the human expert.

Machine learning methods have been proposed to address the limits of human involvement and to introduce a degree of automation to idiom detection. Munzy et al.[11] proposed a binary classification model based on Wiktionary to predict whether a given phrase is idiomatic or literal. Wang et al. [12] presented an attention-based mechanism to choose the best matching idiom to be picked for the Chinese idiom cloze test.

More recently, language models based on transformer architectures have been effective in metaphor detection[13]. Skvorc et al.[14] used the contextual embeddings of BERT and ELMo(Embeddings from Language Models)[15] on deep neural networks to outperform existing approaches for detecting Slovenian idioms. To the best of our knowledge this is the first study to utilize a language model for idiom detection.

The rest of the paper is organized as follows. Section 2 describes the idioms datasets, followed by Section 3 which presents the idiom detection model. Section 4 reports on the experiments and results, and Section 5 concludes the paper.

II. DATASETS

A. Idioment Dataset:

Williams et al.[9] have collected, annotated and verified the Idioment Dataset via crowdsourcing. This dataset consisted of 2525 sentences in which 580 different idioms are included. Each idiom is presented across several sentences, and in different forms of expression.

B. EPIE Dataset:

The English Possible Idiomatic Expressions (EPIE) is a comparatively large dataset presented by Saxena et al. [16] This contains 25206 sentences along with lexical instances of 717 idiomatic expressions. Here the extraction was done using several string processing steps using StringNet, other pattern-matching methods and some manual work. Similar to the

Idioment dataset, here each idiom is presented across several sentences. They have split the dataset into two classes. (1) Static idioms Dataset - idioms that do not undergo any lexical modifications, which contains 21891 sentences with 359 Static idioms. (2) Formal idioms Dataset - idioms which occur in sentences with various lexical modifications, which contains 3135 sentences with 358 formal idioms.

C. Theidioms.com Dataset:

This dataset was curated by the authors specifically for this study. We utilized the resources in *theidioms.com* [17] which organizes idioms by topics. We collected 1106 unique idioms and corresponding example sentences. The main difference between this dataset and Idioment and EPIE is the one-to-one mapping between idioms and sentences. Table 1 summarises the four datasets.

Table 1 : Overview of the datasets

Dataset Name	Number of sentences	Number of Idioms	Average sentences per idiom
Idioment	2525	580	4.35
EPIE static	21891	359	60.67
EPIE formal	3135	358	8.75
theidioms.com	1106	1106	1

III. Methodology

Our proposed approach is based on BERT (Bidirectional Encoder Representations from Transformers) [18], [19]. The novelty of BERT compared to other language models is the use of mask language modelling for training [20], in contrast to right-to-left or left-to-right language modelling. This allows BERT to effectively encode information from both ends of each transformer layer which enhances the performance.[21] DistilBERT[22] is a computational improvement of the baseline BERT model, and we have used HuggingFace Transformers library for accessibility [23].

We have modeled the idiom detection as a token classification problem and fine-tuned the pre-trained DistilBERT model. We preprocessed all four datasets for token classification by splitting sentences into words and generating word-wise labels. When creating the labels, we marked idiomatic tokens as 1 and non-idiomatic tokens as 0 (See Fig. 1).

As the tokenizer is working as a word-piece tokenizer, single word is further split into sub-tokens. Thus, token-label alignment was done by aligning the original token with the first-word piece and populating the rest with the non-idiomatic label 0. Sentence size was used as 40 where the padding and truncation were applied appropriately. Weight_decay was set to 0.01 and warmup_steps were set to 500. We selected the batch size as 16 for training and 5 for evaluation where used training epochs {3, 10} depending on the dataset.

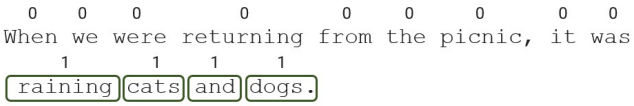

Fig. 2 : Finetuning BERT for token classification

Figure 2 depicts an instance of token classification. Here non idiomatic tokens are annotated with 0's where idiomatic tokens are annotated with 1's. Generally, in a given text there is a majority of non-idiomatic tokens. In idiom detection system, as well as correctly identifying the idiomatic tokens, it is also important that model can recognize non idiomatic tokens as not a part of idiom. Because there can be same wording as idioms without idiomatic sense. Moreover, by modelling this as token classification we can locate the idiom within the text other than just classify sentences for idiom existence.

IV. Experiments and Results

We conducted a series of experiments on all four datasets to evaluate the effectiveness of the proposed model. This includes four experiments focusing on the individual datasets, followed by three experiments that combine the datasets to investigate model robustness.

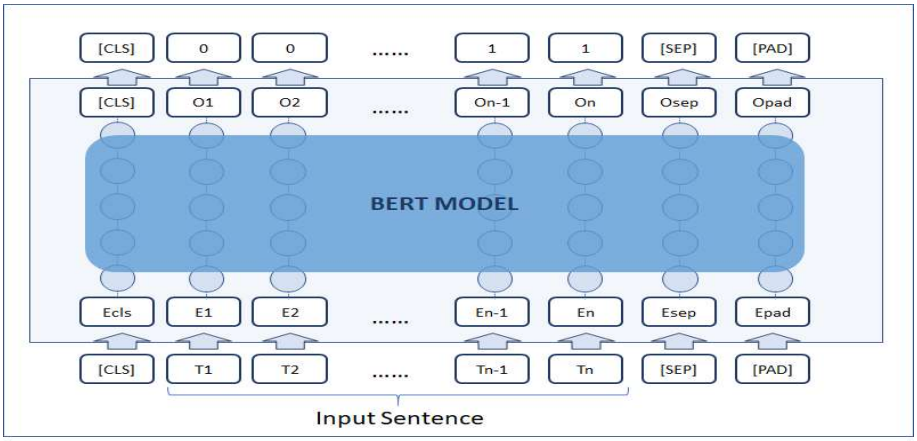

Fig. 1 : Finetuning BERT for token classification

Table 2 : Evaluation metrics for experiments 1-7. The Overall Accuracy, F1 Score, Precision, Recall is represented as pair of values for non-idiomatic term prediction and idiomatic term prediction respectively

Experiment	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Accuracy	0.980	0.972	0.997	0.994	0.960	0.948	0.944
F1 Score	[0.989 0.904]	[0.985 0.815]	[0.998, 0.984]	[0.996 0.959]	[0.978 0.756]	[0.972 0.657]	[0.970 0.633]
Precision	[0.992 0.882]	[0.978 0.888]	[0.999, 0.979]	[0.996 0.962]	[0.974 0.799]	[0.960 0.773]	[0.958 0.752]
Recall	[0.986 0.926]	[0.991 0.754]	[0.998, 0.989]	[0.997 0.957]	[0.983 0.717]	[0.984 0.571]	[0.983 0.548]

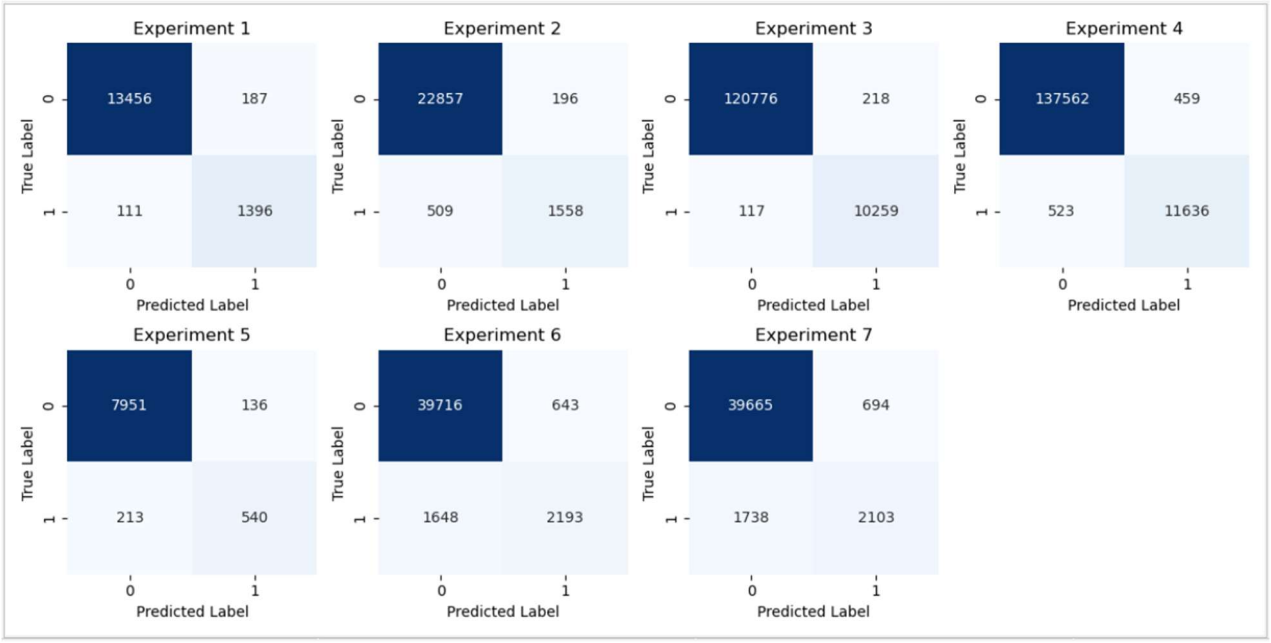

Fig. 3 : Confusion matrices for the experiments

The experiment configurations are as follows,

1. Use the Idioment dataset. 80%, 20% train, and test split.
2. Use the EPIE Formal dataset. 80%, 20% train, and test split.
3. Use the EPIE static dataset. 80%, 20% train, and test split.
4. Merged EPIE Formal and Static datasets. 80%, 20% train, and test split.
5. Use the theidioms.com dataset. 80%, 20% train, and test split.
6. Using Idioment dataset as the training set and theidioms.com dataset as the test set.
7. Using EPIE Formal dataset as the training set and theidioms.com dataset as the test set.

As tabulated below (Table 2), the proposed idiom detection model consistently generates an accuracy of more than 94%. The highest accuracy is recorded for the EPIE - Static dataset, 99.4%.

As we have modeled the idiom detection problem as a token classification, we used confusion matrices to visualize the prediction of idiomatic terms and non-idiomatic terms (See Fig.). Across all instances, we report high accuracies for predicting non-idiomatic terms than detecting idiomatic terms.

Experiments 1 to 4 are shuffled and split into 80% training and 20% testing, respectively. Thus, the test set contains unseen idioms that were not included in the training set. Experiment 5 only uses *theidioms.com* dataset. When comparing with other datasets, this dataset contains a one-to-one mapping with the sentences and idioms. When we split the dataset into training and testing partitions, the test set contained entirely new set of idioms that was not included in the training set. In this experiment the accuracy was 96% where all non-idiomatic matrices exceeded 97% and all idiomatic matrices exceeded 71%. This further validates that DistilBERT has learned the underlined figurative nature of the idioms and the ability to identify entirely new idioms unseen during training. In experiments 6 and 7, we have used a combination of datasets, with nuances like different sentence lengths, and this too led to an accuracy of more than 94%.

Table 3 : Evaluation metrics comparing with other embeddings and NB classifier. The Overall Accuracy, F1 Score, Precision, Recall and Confusion metrices are represented as pair of values for non-idiomatic term prediction and idiomatic term prediction respectively

Experiment	BOW embeddings with Naïve Bayes Classifier	Word2vec embeddings with Naïve Bayes Classifier	Our approach
Accuracy	0.439	0.549	**0.960**
F1 Score	[0.512 0.339]	[0.662 0.323]	**[0.978 0.756]**
Precision	[0.876 0.216]	[0.849 0.224]	**[0.974 0.799]**
Recall	[0.362 0.775]	[0.542 0.579]	**[0.983 0.717]**
Confusion Matrix	[[1608 2829] [227 784]]	[[2414 2036] [427 588]]	**[[7951 136] [213 540]]**

Furthermore, this idiom detection model highlights the significance of contextual embeddings compared to first-generation neural embeddings such as Bag of Words (BOW)[24], Word2vec[25] in detecting figurative components in a text corpus. Contextual embeddings help to retain the phenomenon of polysemy which captures different senses of a given word based on the context (e.g., paper as a material, as a newspaper, as a scientific work, and as an exam). Here, the scope of the context is typically a sentence, and a word will get a unique vector for each figurative or literal meaning it takes [14]. This capability intensifies the uniqueness of the idiomatic segments in a given sentence even if it is unseen by the model.

As shown in table 03, we compared our language model-based token classification method against existing popular embedding generation methods for the experiment 5. As embedding generation methods, we used BOW and Word2vec. Naïve Bayes was used as the classifier. This further elaborate the significance of contextual embedding in figurative language detection problem.

V. Conclusion

Idiomatic expressions are an outcome of cultural convention and practice. They do not convey the literal meaning of the constituent words, thereby mislead most NLP/U techniques. It is also a challenge in language translation when a non-native speaker is unfamiliar with such figurative expressions. In this paper, we presented a novel idiom detection model that distinguishes between non-idiomatic and idiomatic expressions. It utilizes a token classification approach to fine-tune DistilBERT. This model was empirically evaluated on four idiom datasets, yielding an accuracy of more than 94%, that confirms its effectiveness at idiom detection from free-form text data. As future work, we intend to extend this work to enhance the accuracy in sentiment classification by incorporating the detection of idiomatic phrases. Furthermore, we plan to expand the algorithm to detect other types of figures of speech as well idioms in other languages.

References

[1] G. Salton, “Representations of Idioms for Natural Language Processing: Idiom type and token identification, Language Modelling and Neural Machine Translation,” *Doctoral*, Jan. 2017, doi: https://doi.org/10.21427/D77H8K.

[2] G. Salton, R. Ross, J. D. Kelleher, G. D. Salton, and R. J. Ross, “An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese,” *Conference papers*, Jan. 2014, doi: https://doi.org/10.21427/D7GG6N.

[3] E. Avramidis, V. Macketanz, U. Strohriegel, A. Burchardt, and S. Möller, “Fine-grained linguistic evaluation for state-of-the-art Machine Translation,” in *Proceedings of the Fifth Conference on Machine Translation*, Nov. 2020, pp. 346–356.

[4] S. Pelosi, “Semantically Oriented Idioms for Sentiment Analysis. A Linguistic Resource for the Italian Language,” *Advances in Intelligent Systems and Computing*, vol. 1151 AISC, pp. 1069–1077, Apr. 2020, doi: 10.1007/978-3-030-44041-1_92.

[5] B. B. Klebanov, J. Burstein, and N. Madnani, “Sentiment Profiles of Multiword Expressions in Test-Taker Essays: The Case of Noun-Noun Compounds,” *ACM Trans. Speech Lang. Process.*, vol. 10, no. 3, Jul. 2013, doi: 10.1145/2483969.2483974.

[6] C. Cacciari, “Processing multiword idiomatic strings: Many words in one?,” *The Mental Lexicon*, vol. 9, no. 2, pp. 267–293, 2014, doi: https://doi.org/10.1075/ml.9.2.05cac.

[7] G. I. Yusifova, “Syntactic Features of English Idioms,” *International Journal of English Linguistics*, vol. 3, no. 3, p. p133, May 2013, doi: 10.5539/IJEL.V3N3P133.

[8] A. Feldman and J. Peng, “Automatic Detection of Idiomatic Clauses,” in *Computational Linguistics and Intelligent Text Processing*, 2013, pp. 435–446.

[9] L. Williams, C. Bannister, M. Arribas-Ayllon, A. Preece, and I. Spasić, “The role of idioms in sentiment analysis,” *Expert Systems with Applications*, vol. 42, no. 21, pp. 7375–7385, Nov. 2015, doi: 10.1016/J.ESWA.2015.05.039.

[10] Priyanka and R. M. K. Sinha, “A system for identification of idioms in Hindi,” *2014 7th International Conference on Contemporary Computing, IC3 2014*, pp. 467–472, 2014, doi: 10.1109/IC3.2014.6897218.

[11] “Automatic Idiom Identification in Wiktionary - ACL Anthology.” https://aclanthology.org/D13-1145/ (accessed Mar. 24, 2022).

[12] X. Wang, H. Zhao, T. Yang, and H. Wang, “Correcting the Misuse: A Method for the Chinese Idiom Cloze Test,” pp. 1–10, Nov. 2020, doi: 10.18653/V1/2020.DEELIO-1.1.

[13] X. Chen, C. Wee, B. Leong, M. Flor, and B. B. Klebanov, “Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task,” pp. 235–243, Jul. 2020, doi: 10.18653/V1/2020.FIGLANG-1.32.

[14] T. Škvorc, P. Gantar, and M. Robnik-Šikonja, “MICE: Mining Idioms with Contextual Embeddings,” *Knowledge-Based Systems*, vol. 235, Aug. 2020, doi: 10.1016/j.knosys.2021.107606.

[15] M. E. Peters *et al.*, "Deep Contextualized Word Representations," *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 2227–2237, 2018, doi: 10.18653/V1/N18-1202.

[16] P. Saxena and S. Paul, "EPIE Dataset: A Corpus for Possible Idiomatic Expressions," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12284 LNAI, pp. 87–94, Sep. 2020, doi: 10.1007/978-3-030-58323-1_9.

[17] "The Idioms - Largest Idiom Dictionary." https://www.theidioms.com/ (accessed Mar. 24, 2022).

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.

[19] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017.

[20] W. L. Taylor, "'Cloze Procedure': A New Tool for Measuring Readability:," *https://doi.org/10.1177/107769905303000401*, vol. 30, no. 4, pp. 415–433, Oct. 2016, doi: 10.1177/107769905303000401.

[21] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 833–844, 2019, doi: 10.18653/V1/D19-1077.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019.

[23] T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," Oct. 2019.

[24] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1–4, pp. 43–52, Dec. 2010, doi: 10.1007/s13042-010-0001-0.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, Jan. 2013, doi: 10.48550/arxiv.1301.3781.