NTU Journal of Engineering and Technology (2023) 2 (2) : 16-20
DOI: https://doi.org/10.56286/ntujet.v2i2.689

IRAQI Academic Scientific Journals

P-ISSN: 2788-9971 E-ISSN: 2788-998X
NTU Journal of Engineering and Technology
Available online at: https://journals.ntu.edu.iq/index.php/NTU-JET/index

Intelligent System to Transform Slang Words into Formal Words

1st Ahmed Abdulstar Ibrahim[1], 2nd Ban Shareef Mustafa[2]
1. Northern Technical University, ahmdbdalstaralbdly@gmail.com , 2. University of Mosul, banmustafa66@uomosul.edu.iq

Article Informations

Received: 13-09- 2023,
Revised: 08-10-2023,
Accepted: 10-10-2023,
Published online: 17-10-2023

Corresponding author:
Name:Ahmed Abdulstar
Affiliation :Northern Technical University
Email: ahmdbdalstaralbdly@gmail.com

Key Words:
NLP,
NLP task,
Transformer,
Slang words,
Facebook/bart-base,
Slang to formal words.

A B S T R A C T

Understanding and utilizing informal words not recognized in standard dictionaries poses challenges for users. These words are specific to certain communities, hindering comprehension for those outside. Natural language processing tasks, like translation and summarization, struggle with informal vocabulary and local dialects. Although existing models can translate informal words, comprehensive solutions are elusive due to regional and contextual variations. Developing natural language processing models that consider informal words and local dialects is crucial for future research. This paper presents an updated dataset of informal English words tailored to current usage. Multiple models from the Transformer core library on the Hugging Face platform were trained and evaluated, with the facebook/bart-base model demonstrating high accuracy (training data loss: 0.05299). Continued research and innovation in this field are imperative for effective cross-cultural and intercommunity communication.

THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY LICENSE:
https://creativecommons.org/licenses/by/4.0/

Introduction

In current society, many users face difficulties using formal and appropriate English in their daily lives due to the prevalence and use of colloquial words. It is known that natural language is the language used by humans for communication [9].

Natural language processing is considered an important part of artificial intelligence in the fields of computing and scientific research. To implement these applications, several tools and libraries have emerged to facilitate this process. We have used the available models from Hugging Face through a fine-tuning process, as these models require powerful processors to train artificial intelligence models. The T5-Base model, a text-to-text transformer, was selected. During training on the specified dataset, the model performed accurately and correctly. However, the model produced unreasonable and inaccurate results during testing, displaying incorrect texts. Therefore, we switched to another model called "facebook/bart-base," which also performed accurately and correctly during training on the dataset. When conducting tests, the results were reasonable and accurate, displaying the official texts correctly [13].

It is important to note that the training and execution process for each model takes a long time due to the limited processing power available to us. Among the previous studies:

As previuos studies a present Google's neural machine translation system, has been shown to significantly improve the quality of machine translation. This system could be adapted to formalize slang text by translating informal language into more formal language [10].

In [11], XLNet has been introduced, a generalized autoregressive pre-training method for language understanding. It has been shown to outperform BERT on several NLP tasks and could be a useful tool for formalizing slang text .

[12] present a survey and empirical study of data augmentation and mixed precision training methods for low-resource languages. These methods could be used to improve the performance of NLP models for formalizing slang text in languages with limited training data.

In [4], RoBERTa has been presented, a robustly optimized version of BERT that has been shown to outperform BERT on a range of NLP tasks. RoBERTa's improved performance could make it a useful tool for formalizing slang text.

At last [8], provide a survey of cross-lingual word embedding models, which can be used to improve the performance of NLP models on multilingual tasks. These models could be particularly useful for formalizing slang text in multiple languages.

Background theory

Natural Language Processing

Natural Language Processing (NLP) is an AI branch that enables computers to understand and process human language. NLP tools include NLTK, spaCy, TensorFlow, PyTorch, and Google BERT. NLP focuses on recognizing and understanding the complex elements of natural language, such as vocabularies and sentence structures. NLP tasks include translation and text-to-text conversion. Seq2seq is a deep learning model used for transforming sequences by encoding and decoding them. It has applications in automatic translation and text composition [5,1].

NLP tasks include translation, converting texts from one language to another, and facilitating understanding across cultures. Another task is text-to-text conversion, where texts are transformed without changing the overall meaning, often involving format changes or rephrasing [2].

Seq2seq (Sequence-to-Sequence)

Seq2seq is a deep learning model that transforms sequences from one type to another, such as language translation or text composition. It consists of an encoder and a decoder. The encoder analyzes the input sequence and represents it as a compressed context vector. The decoder generates the output sequence word by word, using the context vector as a starting signal. The model is trained by comparing the generated sequence with the target sequence and updating its parameters based on prediction errors. Seq2seq finds applications in automatic translation and text generation tasks [6].

Hugging Face platform

Hugging Face is an AI platform that simplifies the use of natural language processing. It offers the "Transformers" library, which provides pre-trained models. Key pre-trained models available include [3]:

1. BERT: A versatile model used in tasks like text classification, information extraction, and machine translation.
2. GPT: A model capable of generating coherent and lengthy texts, useful for tasks like dialogue generation and creative writing.

3. GPT-2: An advanced version of GPT known for generating interactive and creative texts
4. GPT-3: A highly capable model with a deep understanding of natural language, used for complex tasks like translation and conversation.
5. T5: A model specialized in transforming texts from one form to another, such as translation and summarization.
6. Seq2Seq: A model used for converting one sequence of texts to another, commonly employed in machine translation and dialogue generation.

These pre-trained models can be fine-tuned on specific tasks or datasets to improve their performance.

Building The dataset

Datasets are essential for training the models available in the Transformer library. After researching and studying the dataset used in previous research and studies, the dataset [14]. has been used. It consists of two columns: the first column contains slang words labeled as "slang," and the second column contains formal words labeled as "formal." However, the dataset is considered outdated and in need of updating to be suitable for training modern models. Modern words were added, and offensive or inappropriate words in our society were removed. Two additional columns were added: the first column contains a sentence for each slang word, and the second column contains a sentence for each formal word. After updating the dataset and making the necessary modifications, the models were trained on the three-column dataset (slang, formal, and context). The models were also trained on the four-column dataset (Slang, Formal, Context, and Text with Formal). It was observed that the models perform better when using the three-column dataset (slang, formal, and context). Table 1 shows part of the data set.

Table 1 A part from the dataset

slang	formal	Context1	Context2
$	dollar	The $ is the official currency of the United States of America	The dollar is the official currency of the United States of America
€	euro	The € is Europe's common currency	The euro is Europe's common currency
a.m	before midday	Let's meet a.m for brunch	Let's meet before midday for brunch
a3	anytime anywhere	a3 works for me	anytime anywhere
		anyplace	anyplace works for me
afaik	as far as i know	afaik, the meeting is still scheduled for tomorrow	as far as i know, the meeting is still scheduled for tomorrow

Model implementation

In the research, a primary dataset representing colloquial English was selected. To enhance its applicability to the current community and achieve improved performance and more accurate results during model training and testing, additional columns were incorporated. The Hugging Face platform offered a range of models, with particular emphasis on leveraging the core Transformer library. Figure 1 shows the detailed steps for implementing the trained model for transforming slang words into formal ones.

Following the completion of training and testing procedures across multiple models, it was observed that the "facebook/bart-base" model exhibited exceptional accuracy in performance. The loss function yielded values of (0.0627) for the training data and (0.04902) for the validation data, indicating its superiority over other models in terms of results.

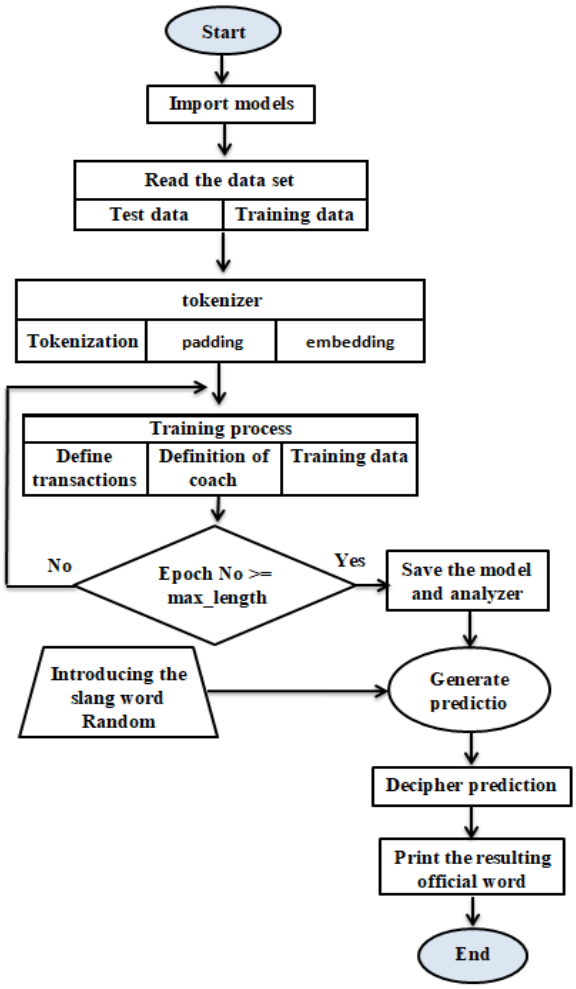

Figure 1 Smart model implementation

Model T5-base (Text-To-Text Transfer Transformer)

The modified dataset has been utilized for training the T5-base model (Text-to-Text Transfer Transformer). The model achieved a loss function value of (0.3176) and an accuracy function value of (0.04902) on the validation data. Table 2 shows a portion of the training data set. Figure 2 illustrates the graphical representation of the loss and accuracy values for the T5-base model.

Table 2 A portion of the training data set

slang	formal	Context
utc	coordinated universal time	The webinar will start at 3 PM utc
w/o	without	I can't go to the party w/o my phone
wassup	what is up	Was sup with your phone? It keeps ringing
wtg	way to go	You finished the project ahead of schedule - wtg!
wtpa	where the party at	I heard there's a party tonight - wtpa?

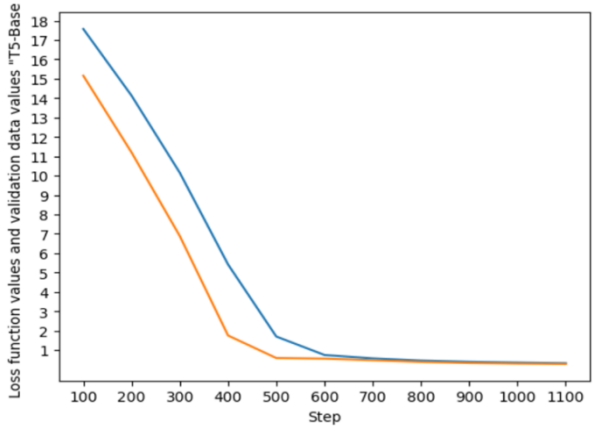

Figure 2 Training Loss and Validation Loss results chart (T5-base) model

After completing the training process and saving the best parameters, the model has been tested on test dataset. It gives un promosied results. Table 3 shows part of the results of the testing process.

Table 3 A part from of the results of the testing process

Slang word (input)	Word prediction using T5-Base	The formal word specified in the test dataset
Tfw	fw-fw-fw-	the feeling when
Np	npp tonp powerppnp brainnpp to	no problem

Model BART (facebook bart-base)

The modified dataset has been utilized for training the bart-base model (Facebook Bart-Base). The model achieved a loss function value of (0.0627) and an accuracy function value of (0.049024) on the validation data. Figure 3 shows the graphical representation of the loss and accuracy values for the BART model.

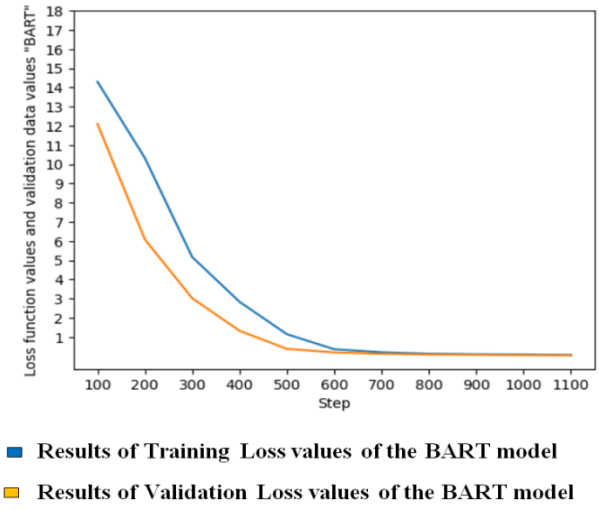

Figure 3 Graphical representation of loss and ccuracy values for the BART model.

The trained model showed a promose results over the test dataset . Table 4 shows part from the the results of the testing process.

Table 4 A part from of the results of the testing process

Slang word (input)	Word prediction using BART	The formal word specified in the test dataset
Omg	oh my god	oh my god
Np	no problem	no problem
u4e	yours for ever	you never know

Results and Discussion

The results of this study show that the "facebook/bart-base" model is highly accurate in translating slang words to formal language, outperforming other models. The model exhibited high accuracy during training, with a training data loss of 0.0627 and a validation data loss of 0.049024. These results indicate that the model successfully learned the patterns and relationships between slang words and their corresponding formal counterparts The Table 5 shows the training results of the two models. The fine-tuning process and updated dataset contribute to the model's effectiveness. These findings have implications for improving cross-cultural communication and highlight the importance of continued research in natural language processing.

The Table 5 the training results of the two models

Model Training Results "T5-Base"			Model Training Results "facebook/bart-base"	
Step	Training Loss	Validation Loss	Training Loss	Validation Loss
100	17.5723	15.159283	14.2934	12.093348
200	14.1312	11.199493	10.3269	6.085192
300	10.1506	6.884148	5.1654	3.017894
400	5.4191	1.751378	2.8324	1.323754
500	1.6972	0.587541	1.1445	0.380795
600	0.7438	0.557606	0.3623	0.196137
700	0.5689	0.473745	0.2043	0.125959
800	0.4559	0.389569	0.1277	0.089293
900	0.3926	0.333231	0.0974	0.069948
1000	0.3535	0.304218	0.0855	0.052999
1100	0.3176	0.287276	0.0627	0.049024
{'eval loss': 0.30421847105026245, 'eval_runtime': 0.4237, 'eval samples_per_second': 18.879, 'eval_steps_per_second': 9.44, 'epoch: 10.0)			{'eval loss': 0.05299945548176765 4, 'eval_runtime': 7.3034, 'eval samples_per_second': 1.095, 'eval_steps_per_second' : 0.548, 'epoch: 10.0)	

Conclusion

In conclusion, the utilization of fine-tuning has proven to be instrumental in enhancing the performance and accuracy of models in predicting new data, particularly in the context of translating slang words to formal language. By adapting to the specific characteristics of both informal and formal languages, linguistic models and artificial intelligence applications have achieved significant improvements in their predictive capabilities. The process of fine-tuning with a hugging face Pretrained models play a pivotal role in advancing the development of translator systems, enabling them to effectively bridge the gap between informal and formal language usage.

References

[1] Elena, C., Danilo, C., Passaro, L. C., & Rachele, S. (2021). Preface to the fifth workshop on natural language for artificial intelligence (nl4ai). In CEUR WORKSHOP PROCEEDINGS (Vol. 3015).

[2] Gour, A. (2020). AI-based Natural Language Processing (NLP) Systems. JOURNAL OF ALGEBRAIC STATISTICS, 11(1), 48-58.

[3] Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 470, 443-456.

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2022). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Nozza, D., Passaro, L., & Polignano, M. (2022). Preface to the sixth workshop on natural language for artificial intelligence (nl4ai). In CEUR Workshop Proceedings. (seleziona...).

[6] Ogundepo, O. J., Oladipo, A., Adeyemi, M., Ogueji, K., & Lin, J. (2022, July). AfriTeVA: Extending? small data? pretraining approaches to sequence-to-sequence models. In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing (pp. 126-135).

[7] Rothman, D. (2021). Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more. Packt Publishing Ltd.

[8] Ruder, S., Vulic, I., & Søgaard, A. (2022). A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, 65, 569-631.

[9] Satapathy, R., Cambria, E., Nanetti, A., & Hussain, A. (2020). A review of shorthand systems: From brachygraphy to microtext and beyond. Cognitive Computation, 12, 778-792

[10] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … & Hughes, T. (2023). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[11] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2023). XLNet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5754-5764).

[12] Zhang, Y., Liu, K., & Sun, M. (2023). Token-level and sequence-level loss smoothing for RNN language models with data augmentation and mixed precision training methods for low-resource languages: a survey and an empirical study.

[13] Rothman, D., & Gulli, A. (2022). Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3. Packt Publishing Ltd.

[14] Slang Translator: A database for translating slang words to formal language. Created in 2017 by Rishabh Verma. Source link: https://github.com/rishabhverma17/sms_slang_translator/blob/master/slang.txt