ENHANCING TEXT CLASSIFICATION WITH LIDA AND ENSEMBLE METHODS

[1] Mohammad Adnaan Ahmed, [2] Dr. K. Santhi Sree

[1] Student, Department of Information Technology, University College of Engineering, Science and Technology, JNTUH Hyderabad

[2] Professor & Head of Department, Department of Information Technology, University College of Engineering, Science and Technology, JNTUH Hyderabad

ABSTRACT: Developing a high-performance text classification model in a low-resource language is challenging due to the lack of labeled data. Meanwhile, collecting large amounts of labeled data is costinefficient. One approach to increase the amount of labeled data is to create synthetic data using data augmentation techniques. However, most of the available data augmentation techniques work on English data and are highly language-dependent as they perform at the word and sentence level, such as replacing some words or paraphrasing a sentence. We present Language-independent Data Augmentation (LiDA), a technique that utilizes a multilingual language model to create synthetic data from the available training dataset. Unlike other methods, our approach worked on the sentence embedding level independent of any particular language. We evaluated LiDA in three languages on various fractions of the dataset, and the result showed improved performance in both the LSTM and BERT models. Furthermore, we conducted an ablation study to determine the impact of the components in our method on overall performance. The source code of LiDA is available at https://github.com/yest/LiDA.

Keywords *– Data augmentation, low-resource language, text classification.*

1. INTRODUCTION

Text classification is the most widely known task in Natural Language Processing (NLP) due to its various applications across many domains. Spam detection, sentiment analysis, emotion detection, topic detection are a few examples of text classification applications. Nowadays, the performance of text classification applications is tremendous because of the deep learning algorithm. However, to achieve such high performance, a deep learning algorithm requires enormous amounts of labeled data. In a low-resource language such as Indonesian, creating a high-performance text classification model is challenging due to the insufficient labeled data. Moreover, collecting enormous labeled data is difficult and costly. One approach to overcome this problem is to create synthetic data by using data augmentation techniques. Data augmentation is a method to create synthetic data from original data. On textual data, this technique aims to transform original sentences into synthetic sentences, and is generally done at the

word or sentence level. For the text classification task, several data augmentation techniques can be used to create synthetic data. At the word level, the simplest strategy is through random word replacements. This approach replaces random words in the sentence with synonyms [1], [2], [3], or the closest words in the word embedding space [4], or a predicted word from the language model [5], [6]. Another technique is changing the sentence structure by deleting some words, inserting a word in a random place, or by swapping some words within the sentence [3]. At the sentence level, the most popular technique is back-translation. This technique produces a sentence that has the same meaning as the original sentence using machine translation models [7], [8], [9], [10].

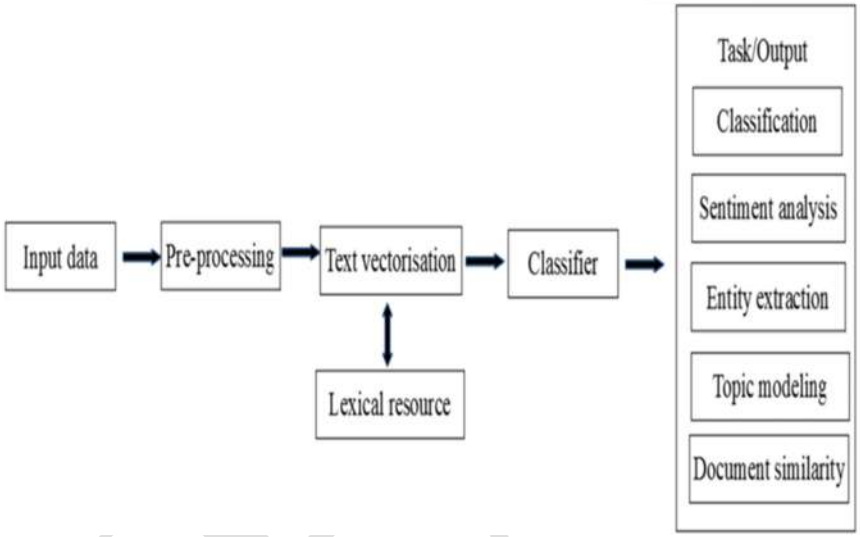

Fig.1: Example figure

Another technique at the sentence level is the generative method. This technique creates synthetic sentences using a text generation model where the model generates tokens sequentially based on the probability of word occurrences formulated from previous word sequences [11], [12], [13], [14]. 1.2 SCOPE Nevertheless, since they operate at the word and phrase level, such as substituting a few words or paraphrasing a sentence, the majority of the data augmentation techniques that are now available only function with English data and are extremely language-dependent. We introduce Language-independent Data Augmentation (LiDA), a method that draws on the available training dataset and a multilingual language model to generate synthetic data.

2. LITERATURE REVIEW

Character-level convolutional networks for text classification:

This article offers an empirical exploration on the use of character-level convolutional networks (ConvNets) for text classification. We constructed several large-scale datasets to show that character-level convolutional networks could achieve state-of-the-art or competitive results. Comparisons are offered against traditional models such as bag of

words, n-grams and their TFIDF variants, and deep learning models such as word-based ConvNets and recurrent neural networks.

Siamese recurrent architectures for learning sentence similarity:

We present a siamese adaptation of the Long Short-Term Memory (LSTM) network for labeled data comprised of pairs of variable-length sequences. Our model is applied to as- sess semantic similarity between sentences, where we ex- ceed state of the art, outperforming carefully handcrafted features and recently proposed neural network systems of greater complexity. For these applications, we provide word- embedding vectors supplemented with synonymic informa- tion to the LSTMs, which use a fixed size vector to encode the underlying meaning expressed in a sentence (irrespective of the particular wording/syntax). By restricting subsequent operations to rely on a simple Manhattan metric, we compel the sentence representations learned by our model to form a highly structured space whose geometry reflects complex semantic relationships. Our results are the latest in a line of findings that showcase LSTMs as powerful language models capable of tasks requiring intricate understanding.

EDA: Easy data augmentation techniques for boosting performance on text classification tasks:

We present EDA: easy data augmentation techniques for boosting performance on text classification tasks. EDA consists of four simple but powerful operations: synonym replacement, random insertion, random swap, and random deletion. On five text classification tasks, we show that EDA improves performance for both convolutional and recurrent neural networks. EDA demonstrates particularly strong results for smaller datasets; on average, across five datasets, training with EDA while using only 50% of the available training set achieved the same accuracy as normal training with all available data. We also performed extensive ablation studies and suggest parameters for practical use.

BAE: BERT-based adversarial examples for text classification:

Modern text classification models are susceptible to adversarial examples, perturbed versions of the original text indiscernible by humans which get misclassified by the model. Recent works in NLP use rule-based synonym replacement strategies to generate adversarial examples. These strategies can lead to out-of-context and unnaturally complex token replacements, which are easily identifiable by humans. We present BAE, a black box attack for generating adversarial examples using contextual perturbations from a BERT masked language model. BAE replaces and inserts tokens in the original text by masking a portion of the text and leveraging the BERT-MLM to generate alternatives for the masked tokens. Through automatic and human evaluations, we show that BAE performs a stronger attack, in addition to generating adversarial examples with improved grammaticality and semantic coherence as compared to prior work.

Contextual augmentation: Data augmentation by words with paradigmatic relations:

We propose a novel data augmentation for labeled sentences called contextual augmentation. We assume an invariance that sentences are natural even if the words in the sentences are replaced with other words with paradigmatic relations. We stochastically replace words with other words that are predicted by a bi-directional language model at the word positions. Words predicted according to a context are numerous but appropriate for the augmentation of the original words. Furthermore, we retrofit a language model with a label-conditional architecture, which allows the model to augment sentences without breaking the label-compatibility. Through the experiments for six various different text classification tasks, we demonstrate that the proposed method improves classifiers based on the convolutional or recurrent neural networks.

3. METHODOLOGY

Many data augmentation approaches can be utilised to produce synthetic data for the text classification problem. The easiest method at the word level is to substitute words at random. In this method, the sentence's random words are changed to their synonyms [1], [2], [3], their nearest word neighbours [4], or a predicted word from the language model [5], [6]. Another method involves altering the sentence structure by deleting words, adding words at random, or switching around terms already present in the text [3]. The most widely used method at the sentence level is back translation. This method uses machine translation models [7], [8], [9], and [10] to create a sentence that has the same meaning as the original text.

Disadvantages:

1. changing the sentence structure by deleting some words, inserting a word in a random place, or by swapping some words within the sentence
2. In a low-resource language such as Indonesian, creating a high-performance text classification model is challenging due to the insufficient labeled data

In this paper, we introduce LiDA: A Languageindependent Data Augmentation technique for text classification. Our approach works at the sentence embedding level, unlike previous methods that perform at the word and sentence level. Our approach was inspired by data augmentation techniques in computer vision where synthetic images would be created by transforming the original image vector with some functions such as flipping, shifting, rotating, zooming, etc. Similarly, our approach transformed sentence embedding with some functions to create a new synthetic sentence embedding.

Advantages:

1. The result shows that LiDA increased the model's performance in both the LSTM and BERT models.

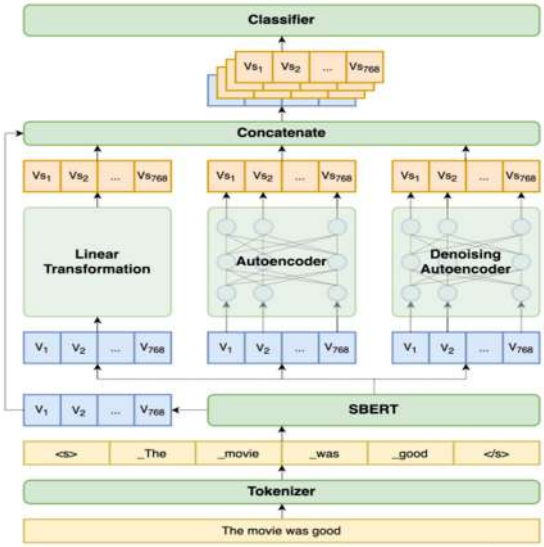

Fig.2: System architecture

MODULES:

To implement aforementioned project we have designed following modules

- Data exploration: using this module we will load data into system
- Processing: Using the module we will read data for processing
- Splitting data into train & test: using this module data will be divided into train & test
- Model generation: Building the model - Sentiment Annotation - BERT - LSTM with BERT Transformer - LSTM with Keras Tokenizer - CNN + LSTM with BERT Transformer - LSTM + GRU with Keras Tokenizer
- User signup & login: Using this module will get registration and login
- User input: Using this module will give input for prediction
- Prediction: final predicted displayed

4. IMPLEMENTATION

Here in this project we are used the following algorithms

Sentiment Annotation – Sentiments are providing helpful insights that often drive business decisions. Sentences are annotated as positive, negative, or neutral when training a machine learning model to analyze sentiments. Our text annotation tool will speedup your sentiment annotation.

BERT – BERT is an open source machine learning framework for natural language processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context.

LSTM with BERT Transformer – The BiLSTM network and BERT are used to acquire a more profound semantics knowledge of each word and contextual sentimental correlation.

LSTM with Keras Tokenizer – This method creates the vocabulary index based on word frequency and then it basically takes each word in the text and replaces it with its corresponding integer value from the word_index dictionary.

CNN + LSTM with BERT Transformer – A CNN can learn features from both spatial and time dimensions. An LSTM network processes sequence data by looping over time steps and learning long-term dependencies between time steps. A CNN-LSTM network use convolutional and LSTM layers to learn from the training data.

LSTM + GRU with Keras Tokenizer- the LSTM (Long -short-term memory) and GRU (Gated Recurrent Unit) have gates as an internal mechanism, which control what information to keep and what information to throw out. By doing this LSTM, GRU networks solve the exploding and vanishing gradient problem.

5. CONCLUSION

We introduced LiDA, a data augmentation technique for text classification. LiDA is not dependent on a particular language and does not require the features of a language, therefore it is suitable for low-resource languages. The experimental results showed that LiDA could improve the performance of the multilingual text classification model without the need for language adjustments. We hope that LiDA can promote the development of universal data augmentation techniques.

REFERENCES

[1] X. Zhang, J. Zhao, and Y. LeCun, ‘‘Character-level convolutional networks for text classification,’’ in Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 649–657.

[2] J. Mueller and A. Thyagarajan, ‘‘Siamese recurrent architectures for learning sentence similarity,’’ in Proc. 13th AAAI Conf. Artif. Intell., 2016, pp. 2786–2792.

[3] J. Wei and K. Zou, ‘‘EDA: Easy data augmentation techniques for boosting performance on text classification tasks,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), Hong Kong, 2019, pp. 6382–6388.

[4] W. Y. Wang and D. Yang, ‘‘That’s so annoying!!!: A lexical and framesemantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., Lisbon, Portugal, Sep. 2015, pp. 2557–2563.

[5] S. Garg and G. Ramakrishnan, ‘‘BAE: BERT-based adversarial examples for text classification,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 6174–6181.

[6] S. Kobayashi, ‘‘Contextual augmentation: Data augmentation by words with paradigmatic relations,’’ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., New Orleans, Louisiana, 2018, pp. 452–457.

[7] A. W. Yu, D. Dohan, Q. Le, T. Luong, R. Zhao, and K. Chen, ‘‘Fast and accurate reading comprehension by combining self-attention and convolution,’’ in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–15.

[8] R. Sennrich, B. Haddow, and A. Birch, ‘‘Improving neural machine translation models with monolingual data,’’ in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, Berlin, Germany, 2016, pp. 86–96.

[9] S. Edunov, M. Ott, M. Auli, and D. Grangier, ‘‘Understanding backtranslation at scale,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., Brussels, Belgium, 2018, pp. 489–500.

[10] A. Sugiyama and N. Yoshinaga, ‘‘Data augmentation using backtranslation for context-aware neural machine translation,’’ in Proc. 4th Workshop Discourse Mach. Transl. (DiscoMT), Hong Kong, 2019, pp. 35–44.

[11] S. Qiu, B. Xu, J. Zhang, Y. Wang, X. Shen, G. De Melo, C. Long, and X. Li, ‘‘EasyAug: An automatic textual data augmentation platform for classification tasks,’’ in Proc. Companion Proc. Web Conf., New York, NY, USA, Apr. 2020, pp. 249–252.

[12] G. Rizos, K. Hemker, and B. Schuller, ‘‘Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification,’’ in Proc. 28th ACM Int. Conf. Inf. Knowl. Manag., New York, NY, USA, Nov. 2019, pp. 991–1000.

[13] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, ‘‘Do not have enough data? Deep learning to the rescue!’’ in Proc. 34th AAAI Conf. Artif. Intell., 32nd Innov. Appl. Artif. Intell. Conf., (IAAI), 10th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI), New York, NY, USA, 2020, pp. 7383–7390.

[14] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, ‘‘GPT3Mix: Leveraging large-scale language models for text augmentation,’’ in Proc. Findings Assoc. Comput. Linguistics, (EMNLP), Punta Cana, Dominican Republic, 2021, pp. 2225–2239.

[15] F. Bond and K. Paik, ‘‘A survey of wordnets and their licenses,’’ in Proc. 6th Global WordNet Conf. (GWC), 2012. 64–71.

[16] N. Reimers and I. Gurevych, ‘‘Sentence-BERT: Sentence embeddings using Siamese BERT-networks,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLPIJCNLP), Hong Kong, Nov. 2019, pp. 3982–3992.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ 2018, arXiv:1810.04805.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, ‘‘Unsupervised cross-lingual representation learning at scale,’’ in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 8440–8451.

[19] X. Dong and G. De Melo, ‘‘Cross-lingual propagation for deep sentiment analysis,’’ in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 5771–5778.

[20] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, ‘‘ERNIE 2.0: A continual pre-training framework for language understanding,’’ in Proc. AAAI Conf. Artif. Intell., vol. 34, Apr. 2020, pp. 8968–8975.

IJESR