DOI: 10.15849/IJASCA.211128.06

Int. J. Advance Soft Compu. Appl, Vol. 13, No. 3, November 2021
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Bilingual Text Classification in English and Indonesian via Transfer Learning using XLM-RoBERTa

Yakobus Keenan Wiciaputra, Julio Christian Young, and Andre Rusli

Department of Informatics Universitas Multimedia Nusantara
keen.wiciaputra@gmail.com
Department of Informatics Universitas Multimedia Nusantara
julio.christian@umn.ac.id
Department of Informatics Universitas Multimedia Nusantara
andre.rusli@lecturer.umn.ac.id

Abstract

With the large amount of text information circulating on the internet, there is a need of a solution that can help processing data in the form of text for various purposes. In Indonesia, text information circulating on the internet generally uses 2 languages, English and Indonesian. This research focuses in building a model that is able to classify text in more than one language, or also commonly known as multilingual text classification. The multilingual text classification will use the XLM-RoBERTa model in its implementation. This study applied the transfer learning concept used by XLM-RoBERTa to build a classification model for texts in Indonesian using only the English News Dataset as a training dataset with Matthew Correlation Coefficient value of 42.2%. The results of this study also have the highest accuracy value when tested on a large English News Dataset (37,886) with Matthew Correlation Coefficient value of 90.8%, accuracy of 93.3%, precision of 93.4%, recall of 93.3%, and F1 of 93.3% and the accuracy value when tested on a large Indonesian News Dataset (70,304) with Matthew Correlation Coefficient value of 86.4%, accuracy, precision, recall, and F1 values of 90.2% using the large size Mixed News Dataset (108,190) in the model training process.

Keywords: *Multilingual Text Classification, Natural Language Processing, News Dataset, Transfer Learning, XLM-RoBERTa*

1 Introduction

While doing our daily activities, we will receive a lot of information. Those informations may present in various forms such as text, pictures, tables, diagrams, audio, video, and so on. Along with the development of technology, we can easily receive and share this kind of information with others through mobile applications on our smartphones every day.

Out of 272.1 million population in Indonesia, 59% or 160 million people are active users of social media, with the average use of social media in Indonesia in 2020 is 3 hours and 26 minutes [1]. Of the total social media users in Indonesia, 99% of them have used messaging applications in the last month [1]. The large use of social media platform, especially in messaging applications, has an impact on the amount of text information we received every day. Thus, all of those text information needs to be processed properly so that there is no disinformation that can harm many parties. There are many ways to process text data, one of them is text classification.

In making a text classification model, sufficient text data is needed so that it can be used as a train data. Most of the current Natural Language Processing research only focuses on 20 languages out of 7000 languages spoken by humans, so there are still many languages that are under-studied or commonly referred as low-resource languages [2]. Based on history, Indonesian itself is a language derivative from Malay, so it is also included in the low-resource languages [3]. Other languages that have not been studied require a machine learning to overcome the problem of insufficient data in those languages. One of the solutions to overcome the lack of data in low-resource languages is to utilize transfer learning which allows a model to perform multilingual text classification.

In solving these problems, there are several previous studies developed by Google and Facebook, including BERT [4], multilingual BERT (mBERT) which is the development of BERT, RoBERTa [5], and XLM [6]. However, there are still some shortcomings and sacrificed aspects in making multilingual text classification model such as lacks of text data, limitations of model in making predictions for low-resource languages, and the accuracy value that still cannot outperform the monolingual model consistently in all languages. The limitations of the model above create an urgency in processing text using machine learning on low-resource languages, including Indonesian text data.

There are 2 language choices that are widely used by Indonesian people, especially the younger generation in using one of the largest social media platforms (Instagram), namely Indonesian and English [7]. As with the case of English, previous works research have tried experimenting and building models to process texts written in Bahasa Indonesia [8] [9] [10], however most only focus on single language classification models. The freedom of using both Indonesian and English use in daily activities on social media creates a new challenge in classifying texts. In addition to requiring larger resources, creating 2 models to classify text in

different languages is also not effective. Based on the level of need and the effectiveness of the work in classifying texts in Indonesia, a model that is able to perform multilingual text classification is needed, especially for Indonesian and English text data.

In 2020, there was a study that developed a new method called XLM-RoBERTa. Cross Lingual Model – RoBERTa or XLM-R is a development of XLM and BERT. The results of this study show, for the first time, that XLM-R makes it possible to create a giant-sized model that can perform multilingual text classification with greater accuracy than most monolingual text classification models and can be used in 100 languages, including low-resource languages, such as Indonesian [11].

The ability of XLM-RoBERTa to perform multilingual Natural Language Processing make XLM-RoBERTa used to improve multilingual language model performance. Previous work has already compared XLM-RoBERTa as multilingual model with monolingual model [12], while other works have also tried XLM-RoBERTa for multilingual sentiment analysis, hate speech detection, and offensive language identification [13], [14], [15], [16], [17], [18] . For Indonesian language, the full utilization of a pre-trained language model like XLM-RoBERTa has successfully improved the result of Indonesian sentiment analysis and hate speech detection [19]. The implementation of XLM-RoBERTa already mentioned also outperforms previous research about multi-label hate speech and abusive language detection in Indonesian Twitter [20].

In this paper, the XLM-R method will be used to perform multilingual text classification for news datasets in different languages (Indonesian and English) to determine news categories based on their titles. The news dataset used is a collection of Indonesian local news in 2017 and English international news from 2012 to 2018. This paper has the main focus and goal to create a model that can perform multilingual text classification on Indonesian language texts using transfer learning with XLM-R which is pre-trained in over 100 different languages.

2 Dataset

The design of the dataset is done to obtain maximum results while training the model. The dataset is adjusted to the focus of the research, namely text classification for English and Indonesian. Based on the research focus, 2 main datasets were used, specifically the English News Category Dataset containing English news titles from 2012 to 2018 from HuffPost [21] and the Indonesian News Title Dataset containing Indonesian news titles from detik.com [22]. From the 2 main datasets, several new datasets were made with different content and size adjustment for related research.

Using our main datasets, we obtain 7 new datasets while experimenting with language combination and dataset sizes. English and Indonesian language were combined to create Mixed News Dataset. While for dataset sizes, we created full size datasets (depends on raw dataset's size), medium size datasets, and small size

datasets (2,000 for each language). In result, there are full size English News Dataset (37,886), full size Indonesian News Dataset (70,304), full size Mixed News Dataset (108,190), medium size Mixed News Dataset (75,772), small size English News Dataset (2.000), small size Indonesian News Dataset (2,000), and small size Mixed News Dataset (4,000). New datasets will be used to evaluate model in performing multilingual text classification with 5 different scenarios of trials.

3 Research Methods

3.1 Dataset Preparation

Our raw datasets contain a lot of information that is not required for our experiment. Therefore, data preprocessing is done to maximize our model performance. Data preprocessing in this paper consists of deleting unnecessary categories, columns, rows, and missing values in our dataset.

	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	5/26/2018
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	5/26/2018
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	5/26/2018
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...	5/26/2018
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	5/26/2018
...	...	...	...	...	...	...
200848	TECH	RIM CEO Thorsten Heins' 'Significant' Plans Fo...	Reuters, Reuters	https://www.huffingtonpost.com/entry/rim-ceo-t...	Verizon Wireless and AT&T are already promotin...	1/28/2012
200849	SPORTS	Maria Sharapova Stunned By Victoria Azarenka I...	NaN	https://www.huffingtonpost.com/entry/maria-sha...	Afterward, Azarenka, more effusive with the pr...	1/28/2012
200850	SPORTS	Giants Over Patriots, Jets Over Colts Among M...	NaN	https://www.huffingtonpost.com/entry/super-bow...	Leading up to Super Bowl XLVI, the most talked...	1/28/2012
200851	SPORTS	Aldon Smith Arrested: 49ers Linebacker Busted ...	NaN	https://www.huffingtonpost.com/entry/aldon-smi...	CORRECTION: An earlier version of this story i...	1/28/2012
200852	SPORTS	Dwight Howard Rips Teammates After Magic Loss ...	NaN	https://www.huffingtonpost.com/entry/dwight-ho...	The five-time all-star center tore into his te...	1/28/2012

200853 rows × 6 columns

Fig. 1. Raw English News Dataset

	text	label
0	How Business Leaders Can Help Foster Mental He...	5
1	Pond-Skimming and Other Reasons to Love Spring...	4
2	We Travel With Our Own Germs	3
3	People Who Live Without Screens Don't Sleep An...	3
4	Move The Starting Line: COO Knows Diverse Team...	5
...	...	...
37882	Maybe "Billionnaire" Should Mean Helping 1 Bil...	5
37883	10 Most Scenic Road Trips	4
37884	How Athletes Stay Calm Under Pressure: Breath ...	0
37885	Fat Studies: Bodies, Culture, Health	3
37886	Dudes: If You're Short On Sleep, Your Percepti...	3

37887 rows × 2 columns

Fig. 2. English News Dataset

Our raw English News Dataset contains 200,853 rows, 6 columns, and 41 news categories (Fig. 1). In this paper, we will use news headlines as our input and category as our predicted output, therefore we only use 2 columns, namely category and headline. From 41 news categories which has been provided, we only used 7 categories with main reason to maintain consistency between English and Indonesian dataset.

	date	url	title	category
0	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	Kemnaker Awasi TKA di Meikarta	finance
1	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	BNI Digitalkan BNI Java Jazz 2020	finance
2	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	Terbang ke Australia, Edhy Prabowo Mau Genjot ...	finance
3	02/26/2020	https://finance.detik.com/moneter/d-4916133/oj...	OJK Siapkan Stimulus Ekonomi Antisipasi Dampak...	finance
4	02/26/2020	https://finance.detik.com/berita-ekonomi-bisni...	Saran Buat Anies-RK yang Mangkir Rapat Banjir ...	finance
...	...	...	...	...
91012	02/03/2020	https://travel.detik.com/travel-news/d-4882807...	Ketumpahan Air Panas di Pesawat, Kamu Bisa Tun...	travel
91013	02/03/2020	https://travel.detik.com/fototravel/d-4882796/...	Foto: Bali & 9 Destinasi Paling Instagramable ...	travel
91014	02/03/2020	https://travel.detik.com/travel-news/d-4882794...	Game Bikin Turis Ini Liburan ke Jepang untuk.....	travel
91015	02/03/2020	https://travel.detik.com/travel-news/d-4882792...	Sekeluarga Didepak dari Pesawat, Maskapai Bila...	travel
91016	02/03/2020	https://travel.detik.com/travel-news/d-4882791...	Kapal Raib di Segitiga Bermuda, Nyaris Seabad ...	travel

91017 rows × 4 columns

Fig. 3. Raw Indonesian News Dataset

	text	label
0	Keris Kiai Naga Siluman, Saksi Bisu Ingkar Jan...	2
1	Kata Polisi soal 2 Pria Diduga Maling yang Dii...	2
2	Citilink Buka Rute Surabaya-Jeddah, Alternatif...	4
3	Robot Anjing Patroli Virus Corona di Singapura	6
4	Universal Studio Jepang Kembali Menyapa Wisatawan	4
...	...	...
70299	2020, Mau Mulai Bisnis Forex? Ini Cara Terbaiknya	5
70300	Dolar AS Gebuk Rupiah di Tengah Geger Corona, ...	5
70301	Ahli Bumi-Antariksa UPI Sebut Dentuman Jakarta...	2
70302	Pemkot Banda Aceh Tunggu Restu Ulama untuk Bik...	2
70303	F-PAN DPRD DKI Minta Anies Selesaikan Masalah ...	2

70304 rows × 2 columns

Fig. 4. Indonesian News Dataset

Our raw Indonesian dataset contains 91,017 rows, 4 columns, and 9 news categories (Fig. 3). The same data preprocessing method was applied to raw Indonesian News Dataset, resulting in Indonesian News Dataset with 70,304 rows, 2 columns, and 7 news categories (Fig. 4). Both English and Indonesian News Dataset contains 7 categories and already mapped into integer numbers such as follows: sports (0),

food (1), world news (2), healthy & living (3), travel (4), business & finance (5), tech & internet (6). Both English and Indonesian News Dataset will be used to create 7 new datasets with different content and sizes as mentioned before.

3.2 Training Scenarios

To find the best model for multilingual text classification, we made 5 different training scenarios for our model. Every training scenario differs by contents (language combination) and sizes (small, medium, and large). Also, every trained model will be tested on all dataset (except dataset used for its training). Here are our training scenarios:

1. Model trained with full size English News Dataset (37,886). The first scenario is done to find out how good is our model to perform text classification on another dataset just by using English in its training phase.
2. Model trained with medium size Mixed News Dataset (75,772). For our second scenario, we used the combination of English and Indonesian language with data ratio 1:1, specifically 37,886 lines for English news dataset and 37,886 lines for Indonesian news dataset, with total lines of 75,772.
3. Model trained with full size Mixed News Dataset (108,190). For our third scenario, we also used the same language combination (English and Indonesian) yet with different size and data ratio. In this scenario, we combined full size English News Dataset (37,886) with full size Indonesian News Dataset (70,304) so that our Mixed News Dataset has the total lines of 108,190 with data ratio of 1:1.85 between English and Indonesian data.
4. Model trained with small size English News Dataset (2,000). For our fourth scenario, we used the smaller size English News Dataset with 2,000 lines as a training dataset for our model. The purpose of this fourth scenario was to find out the impact of training dataset size with model performance on multilingual text classification.
5. Model trained with small size Mixed News Dataset (4,000). For our fifth scenario, we also used language combination between English and Indonesian, but with smaller dataset size. For each language, we will use 2,000 lines of data which return a total of 4,000 lines for both languages, with data ratio of 1:1. This scenario has a purpose to find out the impact of language combination even though the size of the dataset was significantly reduced.

3.3 Evaluation Metrics

There are 2 evaluation metrics to evaluate our model performance, namely Confusion Matrix and Matthew Correlation Coefficient (MCC). Confusion Matrix is one of performance measurement metrics for machine learning classification that

has the output of two or more classes. The calculation of confusion matrix considering the value between the predicted results and the actual value. In 2 classes problem, there are 4 variables that will be used in confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In result, there will be 4 output value, specifically accuracy, precision, recall, and F1-score.

To train our model, we used datasets mentioned before. However, our datasets have one main problem, that is data imbalance. Data imbalance occurs when the sample size in the data classes is unevenly distributed [23]. One way to overcome the data imbalance is to use a calculation algorithm that considers the sample size used in a dataset. The Matthew Correlation Coefficient or MCC is an algorithm that can be used to calculate the value of data that has an imbalance sample size [24]. The calculation of Matthew Correlation Coefficient could be done by using categories from confusion matrix such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The calculation process will return a value with a range from -1 to +1, with value of -1 represents perfect negative correlation, value of 0 represents that the classifier is no better than a random flip of a fair coin, and value of +1 represents perfect positive correlation.

4 Results and Discussion

In this experiment, we used 5 training scenarios as mentioned before. For every trial scenario, there are 2 results that can be used to evaluate our model. The first result was obtained from confusion matrix calculation algorithm with values such as accuracy, precision, recall, and F1-score while the second result was obtained from Matthew Correlation Coefficient (MCC) calculation algorithm. In its implementation, we use hyperparameters as follows: batch size = 32, learning rate = 0.00002, and epochs = 4.

4.1 Model Trained with Full Size English News Dataset (37,886)

The result shows that our model could perform best on small size English News Dataset (2,000) with MCC value of 88.30%, accuracy, precision, recall, and F1-score value of 90.3%. While this model was tested on Indonesian News Dataset, it could only obtain the highest MCC value of 43.30% for small size dataset (2,000) and 42.20% for full size dataset (70,304). Even though the first scenario model did not perform well when tested on Indonesian News Dataset, it shows big potential for our model to be developed furthermore.

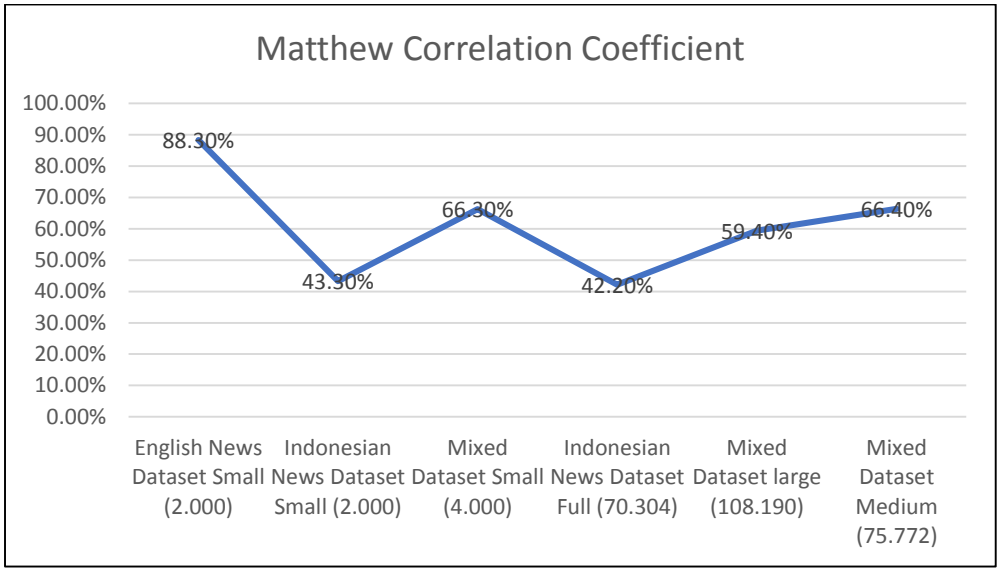

Fig. 5. MCC Value on First Trial Scenario

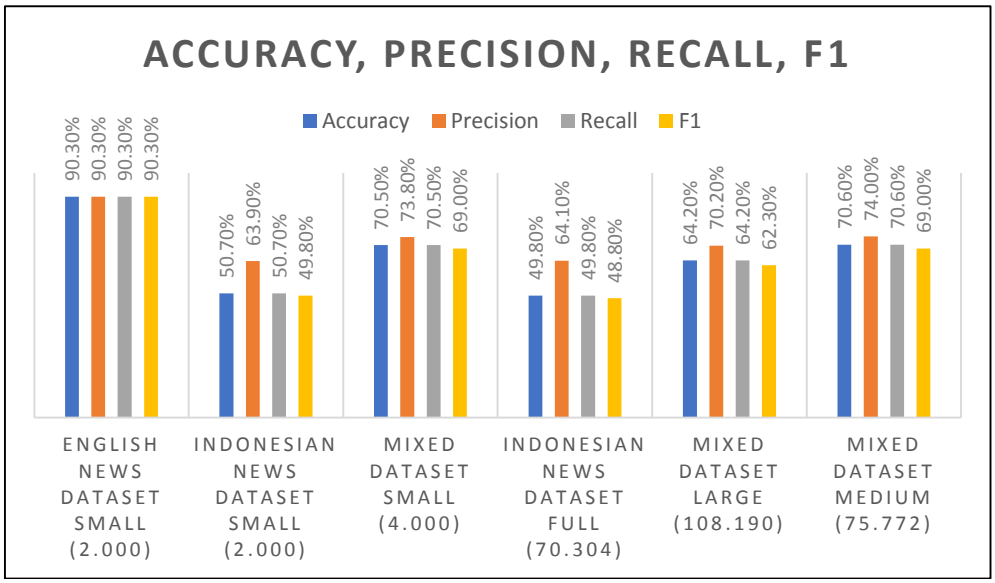

Fig. 6. Confusion Matrix Value on First Trial Scenario

4.2 Model Trained with Medium Size Mixed News Dataset (75,772)

The result shows that our model could perform best on full size English News Dataset (37,886) with MCC value of 89.30%, accuracy, precision, recall, and F1-score value of 91.10%. This model also shows much better performance when tested on Indonesian News Dataset with its best MCC value of 86.40% for full size Indonesian News Dataset (70,304).

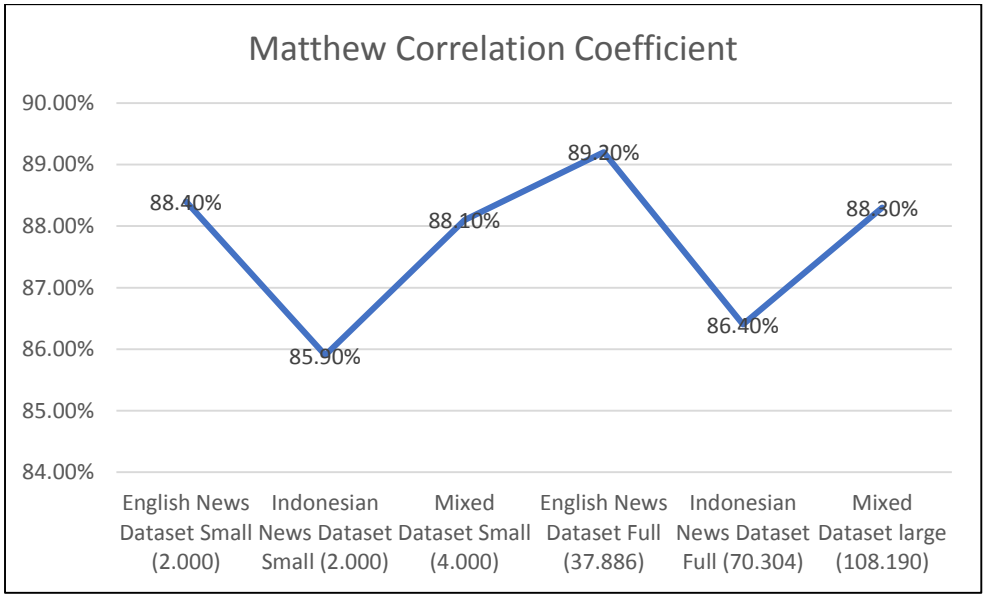

Fig. 7. MCC Value on Second Trial Scenario

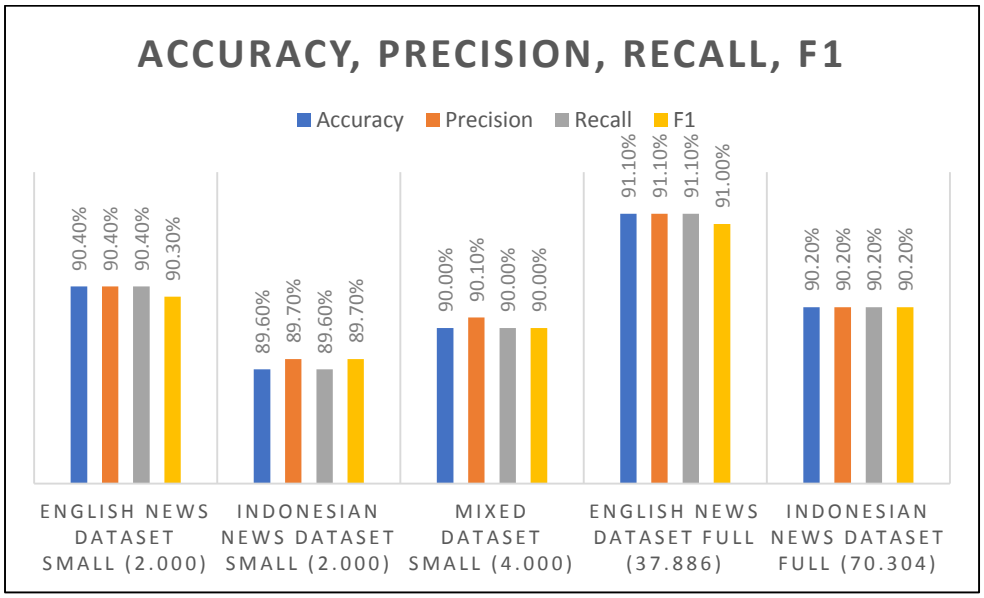

Fig. 8. Confusion Matrix Value on Second Trial Scenario

4.3 Model Trained with Full Size Mixed News Dataset (108,190)

The result shows that our model could perform best on large size Indonesian News Dataset with MCC value of 90.80%, accuracy, recall, F1-score of 93.3% and precision value of 93.4%. From our previous scenario (with medium size Mixed News Dataset), the model managed to get better performance on all test datasets, specifically on Indonesian News Dataset. From this scenario, we knew that the

amount of data combination used in a training dataset has an impact for our model performance.

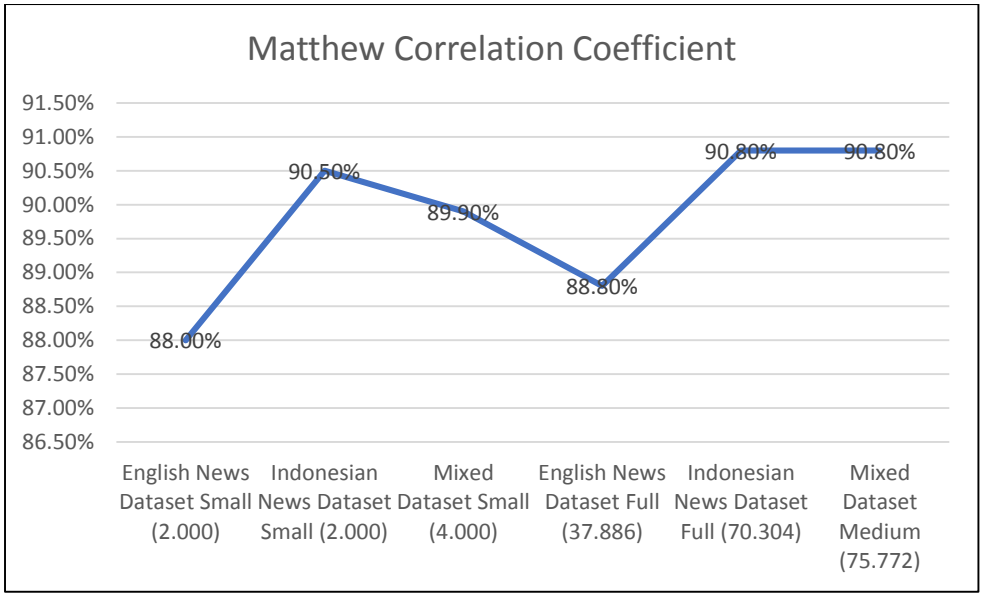

Fig. 9. MCC Value on Third Trial Scenario

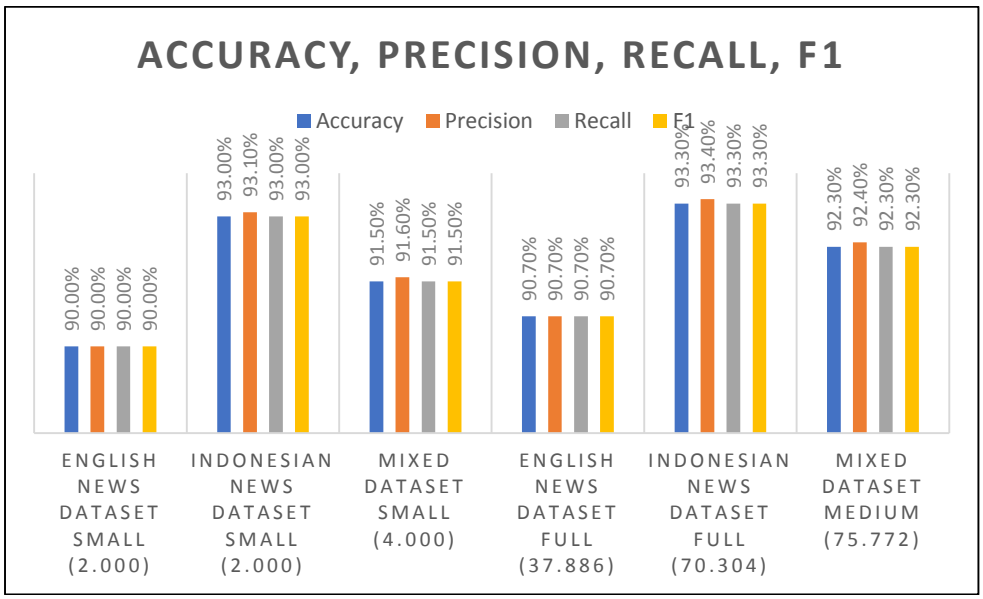

Fig. 10. Confusion Matrix Value on Third Trial Scenario

4.4 Model Trained with Small Size English News Dataset (2,000)

The result shows that our model could perform best on large size English News Dataset (37,886) with MCC value of 21.30%, accuracy and recall value of 35.80%, recall value of 37.20%, and F1-score of 25.80%. From this scenario, we found out

that dataset size has a major impact for our model performance, specifically for multilingual text classification.

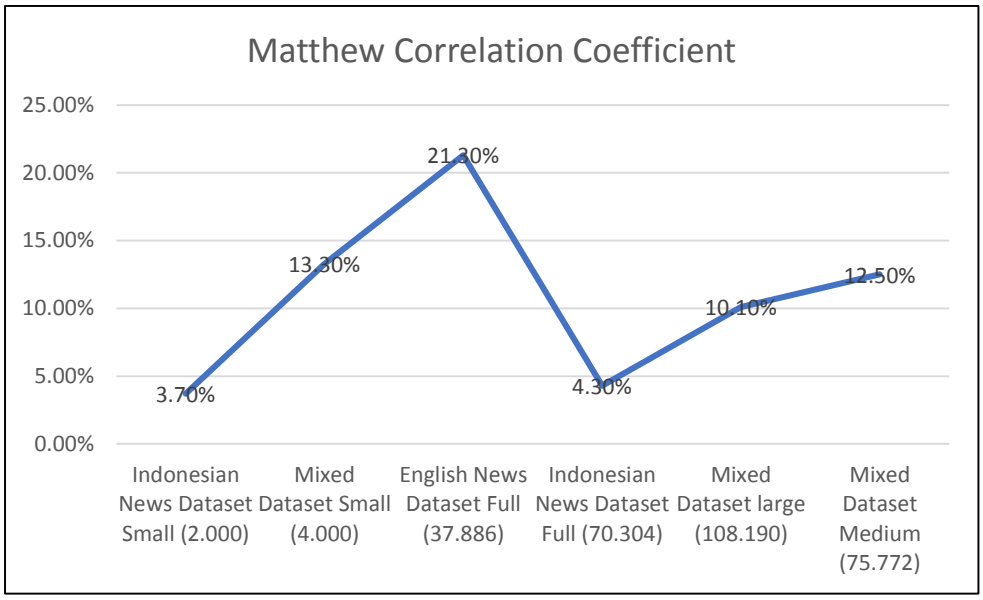

Fig. 11. MCC Value on Fourth Trial Scenario

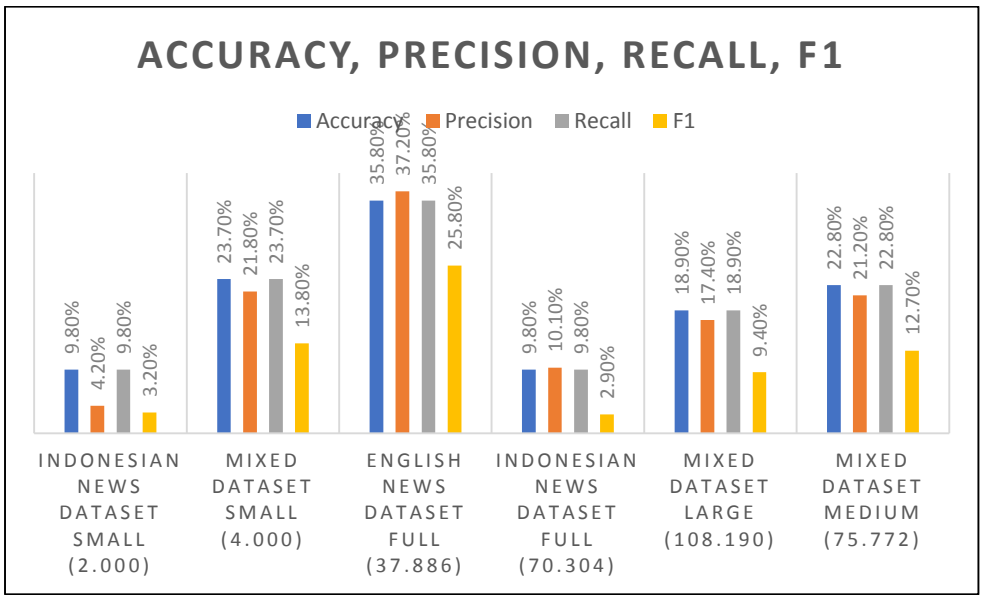

Fig. 12. Confusion Matrix Value on Fourth Trial Scenario

4.5 Model Trained with Small Size Mixed News Dataset (4,000)

The result shows that our model could perform best on small size English News Dataset (2,000) with MCC value of 63.70%. Different from MCC value, this scenario has the best value of confusion matrix when tested on small size Indonesian News Dataset (2,000) with accuracy and recall value of 72.3%, precision value of 71.20%, and F1-score of 71.00%. The result of this scenario also shows that language combination could improve model performance significantly even when using small dataset size (4,000), specifically for multilingual text classification.

Based on all of our experiments, we had the best performance for multilingual text classification by using large size Mixed News Dataset (108,190) as our training dataset. Our model shows that it has better performance for multilingual text classification when using language combination in our training dataset. However, our model couldn't get good performance for multilingual text classification just by using English News Dataset (37,886) as our training dataset. It shows that zero shot transfer learning in our experiment has not reached its maximum capability yet. Therefore, further research and experiment needs to be done in order to obtain good performance for multilingual text classification using zero-shot transfer learning.

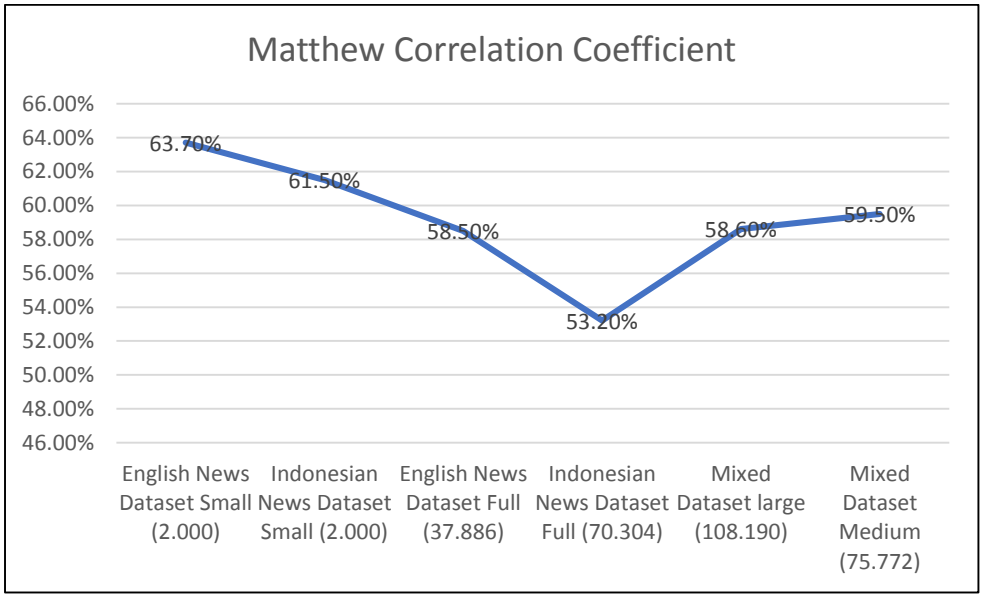

Fig. 13 MCC Value on Fifth Trial Scenario

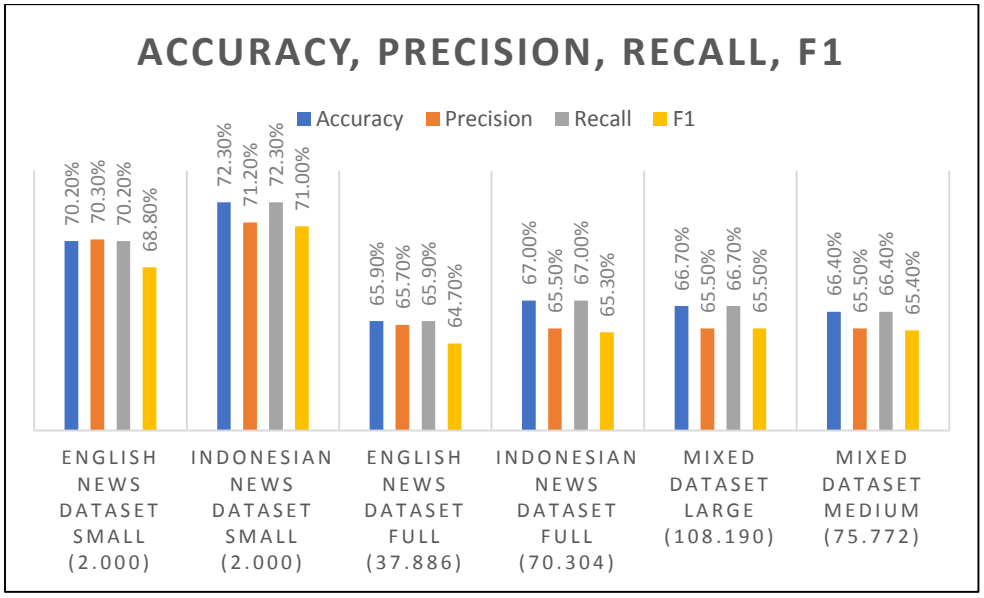

Fig. 14 Confusion Matrix Value on Fifth Trial Scenario

5 Conclusion and Future Works

This paper presented the result of our model that was pre-trained using XLM-R to perform multilingual text classification (English and Indonesian). In general, model was evaluated with 2 performance metrics, namely Matthew Correlation Coefficient (MCC) and Confusion Matrix. Evaluation is done with 5 different scenario using training datasets as a differentiator.

The result of this paper shows that this study is able to perform multilingual text classification with very good accuracy using a model that was trained with large size Mixed News Dataset (108,190). This model managed to get the highest score when tested on a large English News Dataset (37,886) with MCC value of 88.8%, accuracy, precision, recall, and F1 value of 90.7%. While tested on large size Indonesian News Dataset (70,304), this model could obtain MCC value of 90.8%, accuracy, recall, F1-score values of 93.3%, and 93.4% precision value.

Based on our experiments and trial scenarios, it can be concluded that XLM-R model is capable for multilingual text classification using zero-shot transfer learning method. Furthermore, there are 2 factors that have a significant impact on model performance, specifically language combination in a training dataset and the size of dataset used for training. The best dataset size for this research is the largest dataset of the 5 available datasets with a total of 108,190 rows. By combining languages and improving our dataset size, we could obtain the optimum model to perform multilingual text classification for English and Indonesian language.

In this paper, hyperparameter configuration is not experimented yet. Also, the goal to implement multilingual text classification is still limited to 2 languages, namely English and Indonesian language. Therefore, future works could implement hyperparameter configuration on the training phase using XLM-RoBERTa to achieve better model performance for multilingual text classification. In addition, future works could implement XLM-RoBERTa to perform multilingual text classification using zero shot transfer learning on more low-resource languages, especially local languages in Indonesia.

Acknowledgements

We would like to thank our colleagues from the Faculty of Engineering and Informatics in Universitas Multimedia Nusantara who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

References

[1] S. Kemp, "Digital 2020: Indonesia," 18 February 2020. [Online]. Available: https://datareportal.com/reports/digital-2020-indonesia.

[2] Eberhard, D. M., G. F. Simon and C. D. Fennig, Ethnologue: Languages of the World, Twenty-fourth edition ed., Dallas, Texas: SIL International, 2021.

[3] Y. Murakami, "Indonesia Language Sphere: an ecosystem for dictionary development for low-resource languages," *Journal of Physics: Conference Series*, pp. 1-8, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformer for Language Understanding," *Google AI Language*, pp. 1-16, 2019.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *Facebook AI*, pp. 1-13, 2019.

[6] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," *Facebook AI*, pp. 1-10, 2019.

[7] R. Abraham, "Pemilihan Bahasa dan Presentasi Diri dalam Media Sosial Instagram di Kalangan Mahasiswa Program Studi Inggris Universitas Indonesia," *Universitas Indonesia*, pp. 1-18, 2017.

[8] N. Elvina, A. Rusli and S. Hansun, "Implementasi Algoritma Naive Bayes pada LINE Bot untuk Penyaringan Pesan Berdasarkan Kepentingan Organisasi," Universitas Multimedia Nusantara, Tangerang, 2018.

[9] A. Rusli, A. Suryadibrata, S. B. Nusantara and J. C. Young, "A Comparison of Traditional Machine Learning Approaches for Supervised Feedback Classification in Bahasa Indonesia," *International Journal of New Media Technologies,* vol. 7, no. 1, pp. 28-32, 2020.

[10] G. P. Wiratama and A. Rusli, "Sentiment Analysis of Application User Feedback in Bahasa Indonesia Using Multinomial Naive Bayes," in *5th International Conference on New Media Studies (CONMEDIA)*, Bali, 2019.

[11] A. Conneau and K. Khandelwal, "Unsupervised Cross-lingual Representation Learning at Scale," *Facebook AI,* pp. 1-12, 2020.

[12] C. Wang and M. Banko, "Practical Transformer-based Multilingual Text Classification," *NAACL 2021,* pp. 121-129, 2021.

[13] X. Ou and H. Li, "XLM-RoBERTa for Multi-language Sentiment Analysis," *FIRE 2020: Forum for Information Retrieval Evaluation,* pp. 1-6, 2020.

[14] L. Xu, J. Zeng and S. Chen, "Fine-tune XLM-RoBERTa for Hate Speech Identification," *FIRE 2020: Forum for Information Retrieval Evaluation,* pp. 1-8, 2020.

[15] Y. Zhao and X. Tao, "Offensive Language Identification based on XLM-RoBERTa with DPCNN," *Association for Computational Linguistics,* pp. 216-221, 2021.

[16] T. Kanan, G. G. Kanaan, R. Al-Shalabi and A. Aldaaja, "Offensive Language Detection in Social Networks for Arabic Language Using Clustering Techniques," *International Journal of Advances in Soft Computing & Its Applications,* vol. 13, no. 2, 2021.

[17] T. Kanan and E. A. Fox, "Automated arabic text classification with P-Stemmer, machine learning, and a tailored news article taxonomy," *Journal of the Association for Information Science and Technology,* vol. 67, no. 11, pp. 2667-2683, 2016.

[18] T. Kanan, A. Aldaaja and B. Hawashin, "Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents," *Journal of Internet Technology,* vol. 21, no. 5, pp. 1409-1421, 2020.

[19] I. F. Putra and A. Purwarianti, "Improving Indonesian Text Classification Using Multilingual Language Model," *2020 International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA),* pp. 1-5, 2020.

[20] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *Association for Computational Linguistics,* vol. Proceedings of the Third Workshop on Abusive Language Online, pp. 46--57, 2019.

[21] R. Misra, "News Category Dataset," Kaggle, June 2018. [Online]. Available: https://www.kaggle.com/rmisra/news-category-dataset.

[22] Ibrahim, "Indonesian News Title," Kaggle, 14 June 2020. [Online]. Available: https://www.kaggle.com/ibamibrahim/indonesian-news-title.

[23] S. Daskalaki, I. Kopanas and N. Avouris, "Evaluation of Classifier for an Uneven Class Distribution Problem," *Applied Artificial Intelligence,* vol. 20, no. 5, pp. 381-417, 2006.

[24] S. Boughorbel, F. Jarray and M. El-Anbari, "Optimal classifier for imbalanced data using Matthew Correlation Coefficient metric," *PLoS ONE,* pp. 1-17, 2017.

Notes on contributors

Yakobus Keenan Wiciaputra holds a bachelor's degree in Computer Science which he obtained from Universitas Multimedia Nusantara in 2021. His field of interests include the use of applied natural language processing and application development to contribute in making the world a better place to live, especially in this digital era.

Julio Christian Young received his bachelor's degree in Computer Science in 2017 from Universitas Multimedia Nusantara and his master's degree in 2019 from Universitas Indonesia, both in Indonesia. He is working as a lecturer in the Department of Informatics at Universitas Multimedia Nusantara in Indonesia. His field of specialization is in the application of machine learning techniques in natural language processing.

Andre Rusli received his bachelor's degree in Computer Science in 2014 from Universitas Multimedia Nusantara in Indonesia and his master's degree in 2017 from Tokyo Denki University in Japan. He is working a lecturer in the Department of Informatics at Universitas Multimedia Nusantara. His research interests include cross-lingual natural language processing and the application of NLP to support language education.