Multilingual Sentiment Analysis on Social Media Disaster Data

Muhammad Jauharul Fuady[1,2], Roliana Ibrahim[1]
[1]School of Computing, Universiti Teknologi Malaysia, Malaysia
[2]Faculty of Engineering, Universitas Negeri Malang, Indonesia
jauharul@graduate.utm.my, jauharul@um.ac.id, roliana@utm.my

***Abstract*—The use of social media in disaster situations is inevitable, but it is also the case that information presented through this medium can include both public opinion and general information. In a multicultural nation like Malaysia, people like to use codeswitch sentences, through which they mix several languages to express their opinions. Sentiment analysis can be used to classify the subjectivity of social media data, by considering the multilingual aspect of Malaysian users who may experience disaster. In this paper, the authors propose a multilingual sentiment classifier used to understand how Malaysians react during a disaster. The proposed model collects disaster data from social media, which is then classified through a deep learning algorithm, so as to analyze the sentiments of people affected by disasters. The experiment results show that a multilingual sentiment classifier can achieve 0.862 accuracy and 0.864 F1-score which is considered suitable for analyzing social media data. The classification result shows that most Malaysians use social media to disseminate information during disaster periods.**

Keywords—sentiment analysis, codeswitch, multilingual

I. INTRODUCTION

Since the growth of social media, including Facebook and Twitter, people can more easily express their opinion through the Web. The use of social media during disasters has increasingly helped to disseminate information and generate public awareness [1-3]. The abundance of data gathered from social media can be used for understanding public opinions during disaster periods. Guy et al. [1] studied public tweets during earthquakes. The results of their research revealed that the integration of citizen reports and seismically-derived earthquakes have improved earthquake detection in a specified area. David, Ong, and Legara [2] have shown that when the Typhoon Haiyan hit the Philippines, most tweets evolved over time. The latest research by Ragini, Anand, and Bhaskar [3] has revealed that sentiment analysis can be used to determine the basic needs of people affected by disaster.

The contents of the information represent a mix of public opinion and the institutions involved in disaster relief and response [4]. Malaysia is a multiethnic and multicultural country, with a population made up of Malays, Chinese, Indians and other migrants [5]. The official language in Malaysia is Malay, with English also being spoken by most of the population [6]. This affects how Malaysians express their opinions through social media. While most of the time Malaysians use Malay or English, sometimes they mix these languages to express their opinions.

Sentiment analysis is the field of studying and analyzing people's opinions, sentiments, evaluations, appraisals, attitudes and emotions [7]. However, methods for sentiment analysis have been largely limited to the English language. Consequently, many developed approaches cannot be readily applied to other languages, which usually do not have the wealth of labeled data that is exclusive to English [8]. Multilingual sentiment analysis is a topic within sentiment analysis, which aims to solve a sentiment classification task from a cross-language point of view. The importance of this area is that it exploits the existing labeled resources in a source language, building a model in another target language [9]. It saves us from manually labeling data for all languages, which is expensive and time-consuming. Researchers attempted to use machine translation to leverage labeled English data, to com- pensate for the relative lack of training materials in other languages [10].

Sentiment analysis is sensitive to language and domain, from which training data is drawn. Medhat, Hassan, and Korashy [11] have suggested building a bridge between the source and the target, to attain convincingly-labeled documents in the target language and domain. With the many freely-available English sentiment corpora on the Web, Wan [9] proposed a co-training approach to building a cross-lingual sentiment analysis model. Bollegala, Mu, and Goulermas [12] introduced an approach to transferring cross-domain knowledge, by adapting an existing sentiment classifier to previously unseen target domains.

Words for expressing sentiments can differ significantly between languages. Mistranslations between languages can change positive sentiments into negative sentiments. Zou, Wan, and Xiao [10] introduced an approach to filling vocabulary gaps between source and target languages, by using a machine translation system. Mikolov et al. [13] constructed vector spaces for words in English and Spanish, finding that the relative positions of semantically-related words are preserved across languages. Abdalla and Hirst [14] showed an approach to effectively leveraging sentiment knowledge without the need for accurate translation. Multilingual corpora separate the corpus by labeling documents according to their language. However, multilingual speakers often switch between languages when expressing opinions on social media. Lignos and Marcus [15] identified that approximately 11% of tweets contain code-switched content. Solorio et al. [16] introduced code-switched corpus for training. Barman et al. [17] used the dictionary-based approach to detecting language.

There are two main approaches used for sentiment analysis, including lexicon based and machine learning approaches [11]. Zhang, Wang and Liu [18] have shown that deep learning produces state-of-the-art prediction results for sentiment analysis. Chaturvedi et al. [19] has introduced meta-level feature representation, which generalizes well through new domains and languages by using a deep learning method. Vilares [20] has suggested an approach for building a multilingual model trained on a multilingual dataset, which does not require a language recognition step.

This study analyzed the problem of multilingual sentiment analysis from the Malaysian social media context, with data collected during disaster periods. The authors consider multilingual aspects within a sentence type known as codeswitch. They explore the applications of sentiment

978-1-7281-4160-2/19/$31.00 ©2019 IEEE

analysis, and show how sentiment classification in social media can be used to understand how people react during disasters.

II. Model Design and Implementation

This paper's authors propose a sentiment model for multilingual social media during disaster periods, consisting of three phases including the data preprocessing phase, the learning phase, and the classification phase. This research's main contribution is to the learning and classification phase. The learning phase includes data preprocessing and machine translation, providing a comparable multilingual corpus. The multilingual corpora then treated as input for learning multilingual sentiment model. The classification phase analyzes the social media data, using the model with the cross-domain approach. The overall architecture of the proposed model is shown in Fig.1.

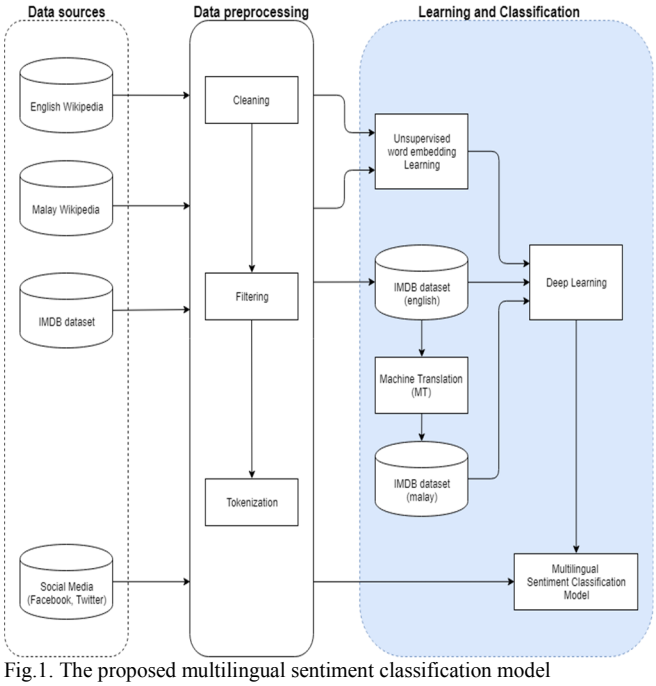

Fig.1. The proposed multilingual sentiment classification model

A. Data Sources

In this study, several lots of data have been collected for different purposes, including training data for multilingual word embeddings, training data for multilingual sentiment classification, and social media data for disasters. The August 2018 article dump of the English and Malay Wikipedia sites was used as the training data for multilingual word embeddings. The data was cleaned so to retain only the normally-visible article text on Wikipedia web pages, by removing HTML codes, URLs, and hashtags symbols. The distribution of content between English and Malay languages was different. The volume of English and Malay data was then balanced, through filtering and randomizing datasets into 100,000 documents. The content was then tokenized into words, as the training input for the next phase. The Large Movie Review Dataset [21] was used as the training data for sentiment classification learning, containing 50,000 movie reviews for training and testing, with balanced amounts of positive and negative sentiments. The first stage here was to perform a data preprocessing phase, synchronizing the words statistics from another dataset, so to minimize the out-of-vocabulary (OOV) words.

The paragraph from each documents will be tokenized into sentences and further split into words. The word appearance in the dataset will then be counted and indexed ordered by its number of appearance. The limitation of the word dictionary will be applied to limit the memory being used such as no further compromise with unrecognizable word that will affect the analysis. The common word and continuation of characters such as punctuation will be removed to produce a clear word dictionary that represent the language. Tokens such as `start`, `end`, and `pad` also added to easily convert the sentence into array of integer so it can be recognize by the model.

B. Learning phases

This paper has sought to make use of the annotated English corpus for the sentiment classification of Malaysian social media data, without using any Malay resources. Labeled English movie reviews and unlabeled Malaysian social media data were than prepared, with the need to consider codeswitch sentences. First a multilingual word embedding model was created, covering a broad domain of terms. Then a multilingual sentiment classification model was built to classify Malaysian social media data.

The authors applied the word2vec [13] and Ballahur's approaches [22], so to learn a multilingual word embedded using English and Malay Wikipedia data. The multilingual word embedding model was used as the dictionary for the embedding layer in the next phase. The dictionary was constructed through a blended approach which could not differentiate between the original English or Malay words, but still preserved the semantic relationships between words.

The authors used Google Machine Translation to produce a comparable Malay dataset from the English dataset, given that it is one of the state-of-the-art machine translation systems used today [23]. The Vilares' approach [20] was then applied, by combining English and Malaysian datasets into a single dataset. The authors used a multilayer deep learning approach to train the model, and used the multilingual word embedding model in the embedding layer.

The deep learning model is built using embeddings layer as its input layer with the 50 dimensional vectors and 100,000 words. The word index limitation is considered to maximize the coverage of word that can be captured from the data while minimizing the out-of-vocabulary value. The embeddings layer can achieve 4.8% oov value from the English dataset, 3.9% oov value from the Malay dataset, and 7% oov value from the mixture of English and Malay dataset. The sentence in the dataset is converted into a fixed 256 length of integer array with padding approach applied when the sentence is shorter than the array length, and cut at the array length when the sentence is longer.

The second layer of the deep learning model is an `AveragePool` layer which returns a fixed-length vector for each data by averaging over the input. This layer will handle the variable length input in a simplest way possible. The next layer of the deep learning model is a `Dense` layer which accept the fixed-length vector through a fully-connected 50 hidden units with the `RELU` activation function. The `Dropout` layer then applied with value of 0.2 as the next layer to add a generalized version of the model before the final layer. The last layer is a `Dense` layer connected with a single node which use the `sigmoid` activation function representing the probability value of the classification.

The learning process itself is conducted by using the keras platform with tensorflow backend. During training, the validation data constructed from 40% of the training data split randomly. The learning process run by using Adam optimizer, binary cross-entropy loss function, and accuracy as validation metric. The model then run to fit in 20 epochs with 1024 batch size.

C. Evaluations

Performance of sentiment analysis model is measured experimentally and assessed using recall, precision, accuracy, error-rate, and F1-score. Calculation of these standard measures is performed based on the following values:

- True Positives (TP): the number of documents that are correctly classified belong to that class
- True Negatives (TN): the number of documents that are correctly classified not belong to that class
- False Positives (FP): the number of documents that are incorrectly classified belong to that class
- False Negatives (FN): the number of documents that are incorrectly classified not belong to that class

Recall is the portion of correctly classified instances against all actual instances for each class. Precision is the portion of correctly classified instances against all classified instances for each class. Accuracy is the portion of correctly classified instances against all classified instances. Error rate is the portion of documents that are classified to the wrong class. F1-score is the harmonic average of precision and recall.

The authors applied a multilayer deep learning approach as the deep learning algorithm when training the multilingual sentiment classifier with multilingual comparable corpora. They also applied the same approach with the English and Malay datasets when building the model. They wanted to compare the performance of the deep learning approach with the lexicon-based approach, and then compare the same deep learning algorithm across different datasets, singular language and a multilingual model. Table I shows information about the multilingual model and its comparison with singular language.

As seen from the table, the accuracy of the lexicon-based approach is 0.666 for Sentiwordnet which represent the english dataset, and 0.579 for Sentiwordnet-Bahasa which represent the malay dataset. The accuracy values of the deep learning model of English and Malay classifiers are 0.874 and 0.854, respectively, and 0.862 on the multilingual classifier. The overall accuracy gained from the deep learning model is higher than the lexicon-based means that the deep learning models perform better than the lexicon-based. It can be seen that the multilingual classifier achieved a reasonable accuracy, one greater than 0.800, to be used as a multilingual classifier for unlabeled Malaysian social media data.

TABLE I. THE MULTILINGUAL SENTIMENT CLASSIFICATION MODEL.

	Senti wordnet	Senti wordnet-Bahasa	English-classifier	Malay-classifier	Multi lingual
Recall	0.666	0.768	0.881	0.864	0.871
Precission	0.704	0.558	0.862	0.848	0.856
Accuracy	0.666	0.579	0.874	0.854	0.862
F1-score	0.651	0.646	0.871	0.855	0.864

The F1-score of the lexicon-based approach is 0.651 for Sentiwordnet, and 0.646 for Sentiwordnet-Bahasa. The F1-score of the deep learning model of English and Malay classifiers are 0.871 and 0.855, respectively, and 0.864 on the multilingual classifier. The overall F1-score of the deep learning models are higher than the lexicon-based approach means that the models will perform better when predict the positive or negative sentiment of the document.

III. RESULTS

The authors then used the multilingual sentiment classifier to analyze the collected social media data. The multilingual sentiment classifier is chosen since the word dictionary coverage is much larger than both English and Malay classifier and can capture the semantic relationship of words from English, Malay, and the mixing between English and Malay. The social media data that will be analyzed consists a mixed between English sentences, Malay sentences, and codeswitch sentences. Table II presents the results of the classification of the social media data, using the multilingual sentiment classifier.

TABLE II. THE SENTIMENT CLASSIFICATION RESULT

	Classification		
	Positive	***Neutral***	***Negative***
Polarity Range	0.7 – 1.0	0.4 – 0.7	0.0 – 0.4
Dcouments	89,496	224,158	16,538
Distribution	0.271	0.679	0.050

It can be seen from the table that neutral sentiments dominate the social media data, representing 67.9% of the total documents. This means that during the disaster period, social media tends to be used as an information dissemination tool rather than as a means for spreading positive or negative sentiment. The 27.1% of positive sentiments is quite large in comparison, representing 5% of the negative sentiments. This means that Malaysians have a more positive attitude when disasters occur. This classifier can also correctly predict the sentiment of codeswitch sentences included in the social media data. Examples of this include "*oh no! heading towards west coast of sabah soon. tempias memang sampai kk ni...*" and "*pantai barat sabah gerimis mengundang but on off.. hopefully nothing bad happen until new year...*"

IV. CONCLUSION AND FUTURE WORKS

In this paper, it has been proposed that a combined dataset and word embeddings during learning will help address the problem of multilingual sentiment classification. The experimental results indicate the effectiveness of the proposed approach. The classification results have also indicated that most Malaysians use social media to disseminate information, as shown by the large number of neutral sentiments.

In future work, this paper's authors will improve sentiment classification accuracy through the morphological similarity between words in different languages will be exploited, so to minimize the unknown words encountered during the learning phase on word embeddings. The Long Short-Term Memory (LSTM) approach will be considered, to increase the performance of multilingual classification. Affection-level sentiment analysis should be employed to further understand the emotions of people affected by disaster.

ACKNOWLEDGMENT

This research was funded and sponsored by IDB on behalf of the Directorate General of Resources for Science, Technology, and Higher Education with a Letter of Assignment 056/D1.1/PR/4IN1/II/2018.

REFERENCES

[1] Guy, M., Earle, P., Ostrum, C., Gruchalla, K., and Horvath, S., "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies", in *International Symposium on Intelligent Data Analysis*, 2010, pp. 42-53: Springer.

[2] David, C.C., Ong, J.C., and Legara, E.F.T., "Tweeting supertyphoon haiyan: evolving functions of Twitter during and after a disaster event", *PloS one,* vol. 11, no. 3, p. e0150190, 2016.

[3] Ragini, J.R., Anand, P.R., and Bhaskar, V., "Big data analytics for disaster response and recovery through sentiment analysis", *International Journal of Information Management,* vol. 42, pp. 13-24, 2018.

[4] Nagy, A. and Stamberger, J., "Crowd sentiment detection during disasters and crises", in *Proceedings of the 9th International ISCRAM Conference*, 2012, pp. 1-9.

[5] Omar, N. and Dan, W.C., "Multiculturalism and Malaysian children's literature in English", *The English Teacher,* p. 19, 2017.

[6] Omar, A.H., "The linguistic scenery in Malaysia", Dewan Bahasa dan Pustaka, Ministry of Education, Malaysia, 1992.

[7] Liu, B., "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies,* vol. 5, no. 1, pp. 1-167, 2012.

[8] El-Beltagy, S.R. and Ali, A., "Open issues in the sentiment analysis of Arabic social media: A case study", in *Innovations in information technology (iit), 2013 9th international conference on*, 2013, pp. 215-220: IEEE.

[9] Wan, X., "Co-training for cross-lingual sentiment classification", in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, 2009, pp. 235-243: Association for Computational Linguistics.

[10] Zhou, X., Wan, X., and Xiao, J., "Cross-lingual sentiment classification with bilingual document representation learning", in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, vol. 1, pp. 1403-1412.

[11] Medhat, W., Hassan, A., and Korashy, H., "Sentiment analysis algorithms and applications: A survey", in *Shams Engineering Journal,* vol. 5, no. 4, pp. 1093-1113, 2014.

[12] Bollegala, D., Mu, T., and Goulermas, J.Y., "Cross-domain sentiment classification using sentiment sensitive embeddings", *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 2, pp. 398-410, 2016.

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J., "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111-3119: Curran Associates, Inc.

[14] Abdalla, M. and Hirst, G., "Cross-lingual sentiment analysis without (good) translation", 2017, pp. 506-515: Asian Federation of Natural Language Processing.

[15] Lignos, C. and Marcus, M., "Toward web-scale analysis of codeswitching", in *87th Annual Meeting of the Linguistic Society of America*, 2013.

[16] Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P., "Overview for the first shared task on language identification in code-switched data", in *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 2014, pp. 62-72.

[17] Barman, U., Das, A., Wagner, J., and Foster, J., "Code Mixing: A challenge for language identification in the language of social media", in *Proceedings of the first workshop on computational approaches to code switching*, 2014, pp. 13-23.

[18] Zhang, L., Wang, S., and Liu, B., "Deep learning for sentiment analysis: A survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* p. e1253, 2018.

[19] Chaturvedi, I., Cambria, E., Welsch, R.E., and Herrera, F., "Distinguishing between facts and opinions for sentiment analysis: survey and challenges", *Information Fusion,* vol. 44, no. June 2017, 2018.

[20] Vilares, D., Alonso, M.A., and Gomez-Rodriguez, C., "Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora", in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Lisboa, Portugal, 2015, pp. 2--8: Association for Computational Linguistics.

[21] Mass, A.L, Daly, R.E., Pham, P.T, Ng, A.Y., and Potts, C., "Learning word vectors for sentiment analysis", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011, pp. 142-150: Association for Computational Linguistics.

[22] Balahur, A. and Turchi, M., "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis", *Computer Speech & Language,* vol. 28, no. 1, pp. 56-75, 2014.

[23] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J., "Google's neural machine translation system: bridging the gap between human and machine translation", *arXiv preprint arXiv:1609.08144,* 2016.