Using Hashtags to Capture Fine Emotion Categories from Tweets

Dr. V. Sharmila[1]
M.E., Ph.D., [1]Professor,
Department of Computer Science and Engineering.
K.S.R. College of Engineering, Tiruchengode, India.

B. Anusuyadevi[2], G. S. Aishwarya[3],
S. Bhavithrasree[4], S. R. Dharani[5]
[2,3,4,5] UG Students
Department of Computer Science and Engineering.
K.S.R. College of Engineering, Tiruchengode, India.

***Abstract*— Despite recent successes of deep learning in many fields of natural language processing, previous studies of emotion recognition on Twitter mainly focused on the use of lexicons and simple classifiers on bag-of-words models. The central question of our study is whether we can improve their performance using deep learning. To this end, we exploit hashtags to create three large emotion-labelled data sets corresponding to different classifications of emotions. We then compare the performance of several word and character-based recurrent and convolutional neural networks with the performance on bag-of-words and latent semantic indexing models. We also investigate the transferability of the final hidden state representations between different classifications of emotions, and whether it is possible to build a unison model for predicting all of them using a shared representation. We show that recurrent neural networks, especially character-based ones, can improve over bag-of-words and latent semantic indexing models. Although the transfer capabilities of these models are poor, the newly proposed training heuristic produces a unison model with performance comparable to that of the three single models.**

Index Terms—Emotion Recognition, Text Mining, Twitter, Indonesian tweet, Natural language processing.

1. INTRODUCTION

Social media has become a new trend for people to interact and communicate. Hence, the growth rate of social media users is increasing rapidly over the years. A social media which has the highest user growth is Twitter. The content of the Twitter post, which is called as tweet, has been widely used by researchers, government or industry to gain knowledge which helps them to solve everyday problems. Various actual human behaviours can be captured from tweets. One of the most popular tasks is emotion analysis.

Emotion is an ongoing state of mind, characterized by mental, physical, and behavioural symptoms. People emotion can be identified directly through their facial expression and speech. Automatically detecting emotion is crucial because it can be implemented in various fields. In education, emotion analysis can be utilized for intelligent e-learning environment. Moreover, emotion analysis can be used in the business for identifying customer complaint in email. In nowadays world where the technology has grown rapidly, people also tend to express their emotion through text in a social Medias post. In social media data such as Twitter, emotion detection can be beneficial in government to monitor public response regarding policy or political event. Moreover, emotion analysis from social media also can be utilized by companies to monitor public responses about services or product thus help them in deciding the target market. Identifying emotion in Twitter is also challenging because its short text with informal words and unstructured grammar cannot be handled using normal text processing techniques. Because of its importance, several datasets are created as a benchmark to obtain state-of-the-art techniques for emotion analysis. Those standard datasets mostly used for English emotion task. However, the standard dataset for another language is limited.

Indonesian tweet is potential for emotion analysis study. According to Statist, an online statistics portal, Indonesia is marked as the third largest active Twitter users in the Asia Pacific from 2012 to 2018. It can be inferred that conducting emotion analysis for Indonesian tweet would be beneficial for many purposes. However, there is not any public dataset for emotion analysis in Indonesia. Previous works in Indonesian emotion analysis, not publish their dataset for the public. In addition, their datasets are limited in small data dan less variety. Therefore, we construct an Indonesian Twitter dataset for emotion classification task which has various characteristics and available for public.

2. RELATED WORK

In addition, we also propose feature engineering to discover the best features for Indonesian emotion classification. Those features include Bag-of-Words, word embeddings, lexicon-based, Part-Of-Speech (POS) tag, and orthographic features. For classifier, there are three methods used: Logistic Regression, Support Vector Machine, and Random Forest. F1-score is utilized as a metric to evaluate the best performance of feature and classifier. To sum up, our main contributions are:

- We build a dataset for Indonesian emotion classification from Twitter data. This dataset consists of 4.403 tweets which divided into five classes of emotions (love, joy, anger, sadness, fear) and publicly available for research purpose2.
- We propose feature engineering which recommends the best features to identify emotion in Indonesian tweet.

The earliest study in emotion mining in text was conducted by Alm et al. They identified emotion expresses in children fairy tales using Valence and Arousal model. The dataset built in their research has been widely used in emotion analysis study. On the other hand, the initial study

https://www.statista.com/statistics/303861/twitter-users-asia-pacificcountry/

https://github.com/meisaputri21/

Indonesian-Twitter-Emotion-Dataset of emotion analysis on Twitter data was introduced by Mohammad. They used n-gram and emotion lexicon based features for detecting the emotion in English tweet based on Ekman's emotion model. Since then, the study of emotion analysis using tweet is increased, both using supervised and unsupervised methods.

Most emotion analysis studies utilize emotion lexicon for classification features. There are several emotion lexicons for English which have been widely used for emotion classification, such as NRC emotion lexicon and Word Net Affect (WNA) lexicon which construct based on Ekman's emotion class. However, there is only one emotion lexicon for Indonesian which was developed by Shaver based on Shaver's emotion definition. Therefore, the study of emotion analysis in Indonesia mostly uses n-gram based feature instead of lexicon-based. Early research on Indonesian emotion analysis on tweet data was conducted by Arifin ET. Al. They use Non-negative Matrix Factorization, an extension of TF-IDF model, to classify emotion in tweets. TF-IDF based features also used by to classify emotion in Indonesian tweet. On the other hand, The et al. used more various features for detecting emotion in Indonesian tweet, including n-gram, linguistic, sentiment lexicon, and orthographic features. They used Shaver's emotion word list as query filters in data collection thus their dataset consists of explicit emotion only. However, all experiments in Indonesian emotion analysis are conducted using their own dataset because there is no standard dataset for Indonesian emotion classification which publicly available.

In recent years, word embed dings dominantly used as a feature for emotion classification. Word embedding features for English emotion detection has been implemented by Heirs et al. They compared the use of basic Bag-of-words (BOW) features and word embedding's (Word2Vec and Glove). The results of their experiment show that combining basic BOW features and word embed dings can improve the performance. Word embed dings for tweet emotion classification also used by Vora et al. Using Random Forest, their model can achieve 91% precision for four classes of emotion in English tweet. However, word embed dings have not been yet utilized for Indonesia emotion classification task.

3. METHODOLOGY

There are two main processes conducted in this study: Emotion classification and Dataset building.

3.1 EMOTION CLASSIFICATION

Paul Ekman studied facial expressions to define a set of six universally recognizable basic emotions: anger, disgust, fear, joy, sadness and surprise

Robert Plutchik defined a wheel-like diagram with a set of eight basic, pairwise contrasting emotions; joy – sadness, trust – disgust, fear – anger and surprise – anticipation.

We consider each of these emotions as a separate category, and we disregard different levels of intensities that Plutchik defines in his wheel of emotions. Profile of Mood States [6] is a psychological instrument for assessing the individual's mood state. It defines 65 adjectives that are rated by the subject on the five-point scale. Each adjective contributes to one of the six categories. For example, feeling annoyed will positively contribute to the anger category. The higher the score for the adjective, the more it contributes to the overall score for its category, except for relaxed and efficient whose contributions to their respective categories are negative. POMS combines these ratings into a six-dimensional mood state representation consisting of categories: anger, depression, fatigue, vigour, tension and confusion. Since POMS is not publicly available, we used the structure from Norcross et al, which is known to closely match POMS's categories. We supplemented it with additional information from the BrianMac Sports Coach website1 Comparing to the original structure, we discarded the adjective blue, since it only rarely corresponds to an emotion and not a colour, and word-sense disambiguation 1. https://www.brianmac.co.uk/pomscoring.htm tools were unsuccessful at distinguishing between the two meanings. We also removed adjectives relaxed and efficient, which have negative contributions, since the tweets containing them would represent counter-examples for their corresponding category.
For each category we used the following adjectives:

- **Anger**: angry, peeved, grouchy, spiteful, annoyed, resentful, bitter, ready to fight, deceived, furious, bad tempered, rebellious,
- **Depression**: sorry for things done, unworthy, guilty, worthless, desperate, hopeless, helpless, lonely, terrified, discouraged, miserable, gloomy, sad, unhappy,
- **Fatigue**: fatigued, exhausted, bushed, sluggish, worn out, weary, listless,
- **Vigour**: active, energetic, full of pep, lively, vigorous, cheerful, carefree, alert,
- **Tension**: tense, panicky, anxious, shaky, on edge, uneasy, restless, nervous,
- **Confusion**: forgetful, unable to concentrate, muddled, confused, bewildered, uncertain about things.

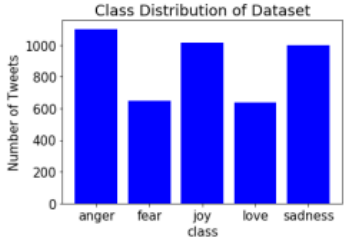
From now on, we will refer to these classifications as Ekman, Plutchik and POMS.

3.2 DATASET CHARACTERISTICS

Our dataset is built based on manual annotation by two annotators. There are 7.500 tweets that should be annotated by annotators. After annotation, the proportion of five basic emotion class, no-emotion, and multi-label emotion are 64%, 32%, and 4% respectively.

In this study, we consider to focus on five basic emotion classes. To measure the quality of annotation, we calculate the Kappa score of five basic emotion classes. The Kappa score of our annotation is 0.917 which considered being very good. The final dataset is taken from the dataset with the agreed label, which consists of 4.403 tweets.
The distribution of our dataset is summarized in Figure.

1. Figure 1Class Distribution of Dataset



To shows that there is a balanced number of joy, anger, and sad class. On the hand, the number of love and fear tweet are limited. To show the variety of our data, we put on the example of tweets in anger class.

First example:

hari ini libur, rencananya mau nonton Jurassic World, tapi kayanya gajadi deh mengingat kon- disi yg gak fit bgt ini sebel. Rusak rencana sebelanga.. sebel akutu (Today is holiday, I am going to watch Jurassic World, but maybe it should be canceled because I am extremely not fit. annoying. What a broke plan. I am annoyed.)

Second example:

Ini aja membuktikan anda sudah TIDAK BENAR....!!! MASA NAPI KORUPTOR
BISA PUNYA HP DI PENJARA ITU SDH MELANGGAR ATURAN.... DAN ANDA DG ENAKNYA MELANGGAR ATURAN...!! INI MENANDAKAN BAHWA ITULAH
KARAKTER ANDA. (It proves that you are NOT TRUE!!! HOW CAN THE CORUPTOR CONVICT HAVE A HAND PHONE IN THE PRISON THAT HAVE BEEN BREAKING THE RULES ... AND YOU ENJOY BREAK THE RULES..!! THIS INDICATES THAT'S YOUR CHARACTER)

The first example contains emotion word, i.e. annoying, hence anger emotion can be indicated explicitly. On the other hand, the second example does not contain any emotion words, but we can identify this tweet as anger because of capitalized characters and exclamation mark. This kind of implicit emotion can be captured in our dataset because we do not use emotion words list on the data collection process. This characteristic is different from another Indonesian tweet dataset which commonly contains explicit emotion only.

3.3 BAG-OF-WORDS & LATENT SEMANTIC INDEXING MODELS

To set the baseline performance, we first experimented with common approaches to emotion detection. Within the realm of pure machine learning (as opposed to using, say emotion lexicons), one of the most frequently used approaches is to use simple classifiers on the bag-of-words (BoW) models. We studied two approaches for transforming raw text into BoW model. Vanilla BoW is a model without any normalization of tokens. Normalized BoW reduces the dimensionality of feature space by these transformations: all @mentions are truncated to a single token, all URLs are truncated to a single token , all numbers are truncated to a single token , three or more same consecutive characters are truncated to a single character (e.g. looooooove → love), TABLE 5 The number of features of BoW and LSI models for combined train and validation sets using different token normalizations. The name bigrams stands for a model consisting of combination of unigrams and bigrams. Ekman Plutchik POMS BoW LSI BoW LSI BoW Unigrams Vanilla 45,484 523 58,146 500 183,727 Unigrams Norm. 35,555 316 44,009 299 129,841 Bigrams Vanilla 204,453 5,433 284,467 6,183 1,248,037 Bigrams Norm. 187,533 3,955 256,889 4,390 1,081,598 all tokens are lower-cased. The aim of these normalization techniques is to remove the features that are too specific. For each of these two models, we run experiments on counts of unigrams as well as unigrams and bigrams. Hereafter, we will refer to the combination of unigrams and bigrams simply as bigrams. Tokenization was done using Tweet POS tagger. For each model, we filtered out tokens and bigrams occurring in less than five tweets. These four BoW models served as a basis for experiments with latent semantic indexing (LSI). We determined the number of dimensions to keep so that 70% of the variance was retained. While the threshold comes from the number of retained dimensions is in the range that empirical studies show as appropriate. LSI experiments were only performed for Ekman and Plutchik, since calculating the decomposition for POMS was not possible with the computation resources we had at our disposal. The dimensionality of BoW and LSI models is shown in Table 5. We experimented with the following classifiers: Support Vector Machines with linear kernel (SVM), Na¨ıve Bayes (NB), Logistic Regression (LogReg) and Random Forests (RF). Regularization parameters for SVM, LogReg, and the number of trees for RF were selected using linear search.

4. EXPERIMENT AND RESULT

We implement our proposed features which have been described in Section III to our built dataset. In addition, we also applied our proposed features into Indonesian tweet dataset for comparison. Their dataset consists of 942 tweet which has similar emotion classes but has different characteristics from ours. Their dataset has explicit emotion because it was build based on emotion words list. On the other hand, our dataset has more variety of data as mentioned in Section IV. We compare the contribution of different features in different Machine Learning classifier for both datasets. We implement several individual features as mentioned in Section III as well as the combination of those individual features. The results of our experiment are summarized in TA-BLE I. We examine the use of different individual features and the combined features. The results show that the use of emotion words list as our baseline feature performs better on The's dataset which contains emotion words explicitly.

This feature achieves 57.85% on F1-score when Logistic Regression applied. On the other hand, the highest F1-score for this baseline feature on our new built dataset is 43.09%. The use of emotion word list is not enough to capture the emotion expressed in our dataset due to the variety of our data.

Other individual features are Bag-of-Words and word embeddings. The use of Bag-of-Words can boost perfor- mance on both datasets. For word embeddings features, we compare the use of Word2Vec and FastText features. In general, FastText obtain better score both in two datasets although Word2Vec perform better on The's dataset when Logistic Regression applied. The great result obtained when we combine the emotion word list, Bag-of-Words, and FastText features. For the lexicon-based feature, InSet sentiment lexicon, get the best F1-Score compared to Vania's lexicon and emoticon list. Vania's lexicon contains formal words while InSet lexicon is developed using Twitter data thus it more suitable for our task. However, there is a slight difference of F1-score obtained from emoticon lexicon feature in both datasets. Combining Vania's lexicon, InSet Lexicon and emotion list obtain slightly higher F1-score than the result of individual InSet. In addition, we examine the effect of combine

emotion words list, Vania's Lexicon, InSet lexicon, and emoticon list for feature combination. The result shows that the combination of these features achieve.

Better performance compare to emotion word list only. To boost the performance of emotion classification model, we also examine the use of POS tag and or- thographic features. The results show that both features not perform well as individual feature, but shows better performances when combined.

For increasing the F1-score, we consider implementing several feature combination scenarios. We take the best feature for each feature group and combine those features. Based on the results in TABLE I, it can be inferred that the most significant features are formed based on the combination of Emotion Words List, Bag-of-Words and FastText. This combination achieve 73.72% of F1-Score in The's dataset and 68.39% in our new dataset. Adding lexicon and additional features (orthographic and POS tag) to the combination of basic features can increase the F1-Score. Both The's dataset and our new dataset achieve the highest F1-score when the combination of basic (emotion word list, Bag-of-Words, FastText), Lex (Vania's lexicon, InSet lexicon, emoticon list), orthography and POS tag features used in the Logistic Regression model. This combination achieves 75.98% of F1-Score on The's dataset and 69.73% of F1-Score on our new dataset. Regarding the classifier model, Logistic Regression performs the best in almost scenarios, followed by Support Vector Machine and Random Forest. In general, our pro-posed feature combination can boost performance in both datasets. The implementation of our proposed features to The's dataset can achieve 75.98% F1-score which is better

TABLE II. EVALUATION OF EACH EMOTION CLASS ON OUR NEW

5. DATASET

Class	Precision	Recall	F1-Score
love	64%	75%	69%
joy	81%	60%	69%
anger	61%	81%	70%
sadness	89%	72%	80%
fear	65%	53%	59%
avg/total	70%	68%	68%

compared to the result of The et. al. implementation with the same dataset with 71.96% accuracy. On the other hand, the implementation of our combined features on our new dataset achieve 69.73%, which outperforms the baseline by 26,64%. Due to the variety and complexity of our new dataset, which consists of explicit and implicit emotion, the learning model cannot perform better than the implementation in The's dataset, which contains explicit emotion only. We present the detail evaluation of each emotion class of our new built dataset in TABLE II. The best-combined features and Logistic Regression classifier are used in this evaluation. TABLE II shows that a balanced score of precision and recall is achieved by sadness class. Sadness class obtains the best evaluation in precision, i.e. 89%. It means that there is only 11% false positive for sadness label. Recall score for sadness class is also quite high, i.e 72%. On the other hand, joy class achieves high precision but low recall. There is 40% of joy class is predicted as false negative. In contrast, anger class obtains low precision but high recall. The lowest score of precision and recall is obtained from fear class. The limited number of samples in fear class impacts to its classification performance.

6. CONCLUSION

The central aim of the paper was to explore the use of deep learning for emotion detection. We created three large collections of tweets labeled with Ekman's, Plutchik's and POMS's classifications of emotions. Recurrent neural networksindeed outperform the baseline set by the common bag-of-words models. Our experiments suggest that it isbetter to train RNNs on sequences of characters than on sequences of words. Beside more accurate results, such approach also requires no preprocessing or to kenization. We discovered that transfer capabilities of our models werepoor, which led us to the development of single unison model able to predict all three emotion classifications at once. We showed that when training such model, instead of simply alternating over the data sets it is better to sample training instances weighted by the progress of training. We proposed an alternative training strategy that sample straining instances based on the difference between trainand validation accuracy and showed that it improves over alternating strategy. We confirmed that it is possible to train a single model for predicting all three emotion classifications whose performance is comparable to the three separate models. As a first study working on predicting POMS's categories, we believe they are as predictable as Ekman'sand Plutchik's. We also showed that searching for tweets containing POMS adjectives and later grouping them according to POMS factor structure yields a coherent data set whose labels can be predicted with the same accuracy as other classifications.

7. REFERENCES

1. W. G. Parrott, Ed., Emotions in social psychology: Essential readings. New York, NY, US: Psychology Press, 2001.
2. N. Gupta, M. Gilbert, and G. Di Fabbrizio, "Emotion detection in email customer care," in Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Association for Computational Linguistics, 2010
3. J. E. The, A. F. Wicaksono, and M. Adriani, "A twostage emotion detection on indonesian tweets," in 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Oct 2015,
4. T. Daouas and H. Lejmi, "Emotions recognition in an intelligent elearning environment," Interactive Learning Environments, vol. 0, no. 0, pp. 1–19, 2018.
5. Zainal Arifin, Y. Arum Sari, E. Kamilah Ratnasari, and S. Mutrofin, "Emotion Detection of Tweets in Indonesian Language using Non-Negative Matrix Factorization,"
6. https://www.google.co.in/?gfe_rd=cr&ei=ZB9GVKuJL63V8gfDzIDABw&gws_rd=ssl