EMOTION RECOGNITION MODEL BASED ON INDONESIAN SENTIMENT TEXT USING MACHINE LEARNING AND NEURO-PHYSIOLOGICAL APPROACH

BY

KHODIJAH HULLIYAH

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

FEBRUARY 2022

ABSTRACT

Emotion Recognition in the Brain and Computer Interface (BCI) field is gaining popularity, not only in terms of volume or amount of incoming data but the variety of media used by netizens and the acceleration of increasing information (velocity) as well. Therefore, the development of techniques and algorithm models with various approaches is a significant concern to recognize the netizens' emotions through writing. This study examined the introduction of text-based emotions in the Indonesian language by taking Twitter data as the dataset. The dataset is processed using two approaches; 1) Recognizing emotions automatically based on sentiment text, and; 2) In real-time viewing brain waves using machine learning and Electroencephalogram (EEG) tools by neuro-physiological approach. The output of these tasks is the accuracy of training data and testing data score. Knowing the results of the accuracy of the two approaches is important, as a reference recommendation to see how much emotion affects the writer and the status of the reader. Furthermore, we conducted preliminary research to obtain Indonesian words with raw data from Affective Norm English Words (ANEW) and classify them into four basic emotions: happiness, sadness, anger, and fear. The highest scored calculation for these four emotions are carried out as keywords in crawling Twitter data. After that, it processed using the Long Short-Term Memory (LSTM) model and also using two benchmark models (Random Forest and Support Vector Machine) at the emotion recognition stage based on sentiment analysis. Next, the dataset in the form of brain waves are processed using the same models. In the sentiment analysis approach, the LSTM model has the highest accuracy value than the two benchmarks. Whereas for data using EEG, Random Forest produces the best accuracy value. Consequently, this research contributed to a collection of datasets based on affective Indonesian words. Besides, it provided recommendations for several algorithm models that match the data and the case. This research's novelty value was to recognize emotions using brain waves with stimulation of reading text with a sentiment analysis approach. Future research was still very much needed to get maximum results to provide knowledge that human emotions can be affected by reading emotional texts.

خلاصة البحث

يحظى مجال التعرف على المشاعر في واجهة الدماغ والكمبيوتر (BCI) شعبية كبيرة لدى الكثير من الباحثين، ولا يقتصر اهتمامهم بحجم البيانات الواردة أو مقدارها فقط بل يشمل تنوع الوسائط التي يستخدمها مستخدمو الإنترنت و سرعة ترقية المعلومات. لذلك، فإن تطوير التقنيات والنماذج الخوارزمية بمناهج مختلفة أصبحت من الموضوعات المهمة للتعرف على مشاعر مستخدمي الإنترنت من خلال النص المقروء.ستعالج هذه الدراسة مقدمة النصوص العاطفية باللغة الإندونيسية من خلال المعطيات الواردة في تويتير كمنظومة البياتات. ستتم معالجة البيانات باستخدام نهجين؛ ١) الكشف عن العواطف تلقائيًا من خلال تحليل المشاعر ؛ ٢) عرض موجات الدماغ في الوقت الحقيقي باستخدام أداة مخطط كهربية الدماغ (EEG) مع نهج التعلم الآلي. يترتب من هذه المعالجة الحصول على نتائج التدرييب الدقيقة و درجات الاختبار. وقد تم إجراء الدراسة الأولية للحصول على الكلمات الإندونيسية بالاعتماد على الكلمات الإنجليزية المعيارية العاطفية (ANEW) كالبيانات الخام لتصنيفها إلى أربعة مشاعر أساسية: السعادة والحزن والغضب والخوف. وسيتم استخدام الكلمات الأربعة الأكثر استعمالا ككلمات رئيسية في تحري البيانات عبر تويتر. وتتم معالجة البيانات باستخدام نموذج الذاكرة الطويلة المدى (LSTM) و نموذجي القياس المقارن (الغابة العشوائية و آلية دعم التوجيه) في مرحلة التعرف على العواطف من خلال تحليل المشاعر. ثم تتم معالجة مجموعة البيانات على شكل موجات دماغية باستخدام نهج الشبكة العصبية بنفس النماذج في تحليل المشاعر، يتمتع نموذج LSTM بدقة أكثر مقارنةً بالمعيارين الأخرين. بينما بالنسبة للبيانات المحصول عليها من خلال استخدام EEG فإن طريقة الغابة العشوائية تنتج أفضل بكثير من حيث دقتها. تساهم هذه الدراسة في جمع منظومة البيانات المستندة إلى الكلمات العاطفية الإندونيسية. كما أنها توصي إلى العديد من نماذج الخوارزميات التي تناسب البيانات والحالة. و تتمتع الدراسة بالقيمة الجديدة المتمثلة في التعرف على العواطف باستخدام موجات الدماغ بالإثارة من خلال النصوص المقروءة باستخدام منهج تحليل المشاعر. هذا و لا تزال هناك حاجة لمتابعة الدراسة في المستقبل للحصول على أقصى قدر من النتائج لتوفير المعرفة بأن المشاعر البشرية يمكن أن تتأثر بقراءة النصوص العاطفية.

APPROVAL PAGE

The thesis of Khodijah Hulliyah has been approved by the following:

Normi Sham Awang Abu Bakar
Supervisor

Amelia Ritahani
Co-Supervisor

Murni Mahmud
Internal Examiner

Mazidah Puteh
External Examiner

Media Anugerah Ayu
External Examiner

Noor Mohammad Osmani
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigation, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Khodijah Hulliyah

Signature Date, February, 15th 2022

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

THE IMPACT OF MOBILE INTERFACE DESIGN ON INFORMATION QUALITY OF M-GOVERNMENT SITES

I declare that the copyright holders of this dissertation are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Khodijah Hulliyah

Signature Date, February, 15th 2022

I dedicate this thesis to my late parents who have laid the foundation for what I experience in this life, my husband who always supports and accompanies me, also to my children who are great and love me

ACKNOWLEDGEMENTS

First of all, I would like to say that I am so grateful to Allah SWT who has allowed me to complete all the steps in my thesis project. Without His mercy and blessing, it would have been impossible for me to work hard and reach all of the objectives of this project.

Firstly, it is my utmost pleasure to dedicate this work to my lovely husband Aceng Abdul Azis, my sons (Izzat, Amar, Eki, and Aras), my brothers and my sisters (and their families) who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your endless support, encouragement, love, and patience.

I gratefully acknowledge my supervisor, Assoc. Prof. Dr. Normi Sham Bt. Awang Abu Bakar for her full support, assistance, supervision, and for giving me spirit in undertaking this project, so that I can complete my work. I would like also to express my acknowledgment to Prof. Abdul Wahab Abdul Rahman for his initial supervision.
I also would like to express my gratitude to Br. Nashrul Hakim, Ph.D., Br. Aang Subiyakto Ph.D., Kiai Ahmad Rusydi, Br. Octaviano, Br. Rusnandi, Br. M. Qomarul Huda Ph.D., Br. Yusuf Durrahman, Br. Herlino Nanang, Sr. Viva Arifin, my all colleagues at Informatics Engineering (TI) Study Program, UIN Jakarta, and my friends in Mahallah Sumayyah for their help and support during my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANEW	Affective Norm English Words
ANIW	Affective Norm Indonesian Words
ANN	Artificial Neural Network
BCI	Brain Computer Interface
CBE	Computer Based of Emotion
CL	Computational Linguistics
CT	Computed Technology
DL	Deep Learning
ED	Emotion Detection
EEG	Electroencephalogram
fMRI	functional Magnetic Resonance Imaging
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RF	Random Forest
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SAM	Self Assessment Manikin
SC	Sentiment Classification
SVM	Support Vector Machine
V/A	Valence and Arousal

CHAPTER ONE

INTRODUCTION

1.1 INTRODUCTION

In this digital era, information plays a pivotal role in human daily life. Language is one of the ways to transfer information as well as it is also a way of communicating views or messages orally or in written text. Moreover, language is not only used for communication but also imparts the emotion associated with it. Writing text is a way of stating the language you want to express, as it was stated by Kaur, written text is one good source for expressing ideas, emotions, and feelings (Kaur & Saini, 2014).

Man created by God has a brain. The behavior of the human brain is very complex and this makes it difficult to interpret. Human emotions may come from brain activity. However, the relationship between the two is still very rarely studied (Liu and Meng 2016), where human emotion is a complex phenomenon that comes from the human brain while there is no clear knowledge of its generation mechanism, even-though human feelings can be easily expressed in the form of writing. The two ways that become interesting topics in analyzing a person's emotions in a text; 1) is how to conduct emotional recognition in a sentiment text with the method of machine learning and 2) by using brain waves.

In addition, one's emotions are so difficult to predict, and has sentiment polarity is insufficient to convey the precise effectiveness of the writers (Almashraee, 2016), because this is related to one's experience and knowledge. Especially if these needs will be carried out by computers, there will be many methods and techniques that must be done to obtain accurate results. Therefore, machine learning becomes an important method used by researchers in studying the analysis of emotional texts (Choudhury, Wang, Carlson, & Khanna, 2019). This is because the behavior of the above methods is very close to human behavior and there are two distinctive features of machine learning which are training data and test data. As stated earlier, emotional recognition research requires a corpus or large dataset to see the accuracy of the classification of emotions in a text.

On the other hand, the introduction of text-based emotions in the digital era has become an important part of the branch of NLP, namely computational linguistics (CL). Emotions can be expressed by a person's speech, facial expression, and written text, which is respectively known as speech, facial, and text-based emotion (Choudhury et al., 2019). A sufficient amount of work has been done regarding speech and facial emotion recognition but text-based emotion recognition systems are getting to gain the attention of researchers. Furthermore, human beings have the power to feel different kinds of emotions because the life of every human being is filled with many emotions, such as joy, fear, anger, and sadness. In Addition, for using a computer, the categorization of text in these emotional states in CL is known as sentimental analysis/emotion detection (Kaur & Saini, 2014). Besides that, emotional words contained in a sentence will affect the sentiment content of the sentence, and emotional words will have different meanings based on the culture or language of a country as well.

Ekman et al. proposed the notion of basic emotions that were universal and found across cultures (Ekman 1987). They proposed the two-dimensional model in which emotions were given coordinates denoting the degree of valence (the positive or negative quality of emotion) and arousal (how responsive or energetic the subject is). That theory could be the based material of corpus for decision making in several fields such as the medical and psychology fields (Sianipar, van Groenestijn, & Dijkstra, 2016).

At present, so many researchers in the field of computational linguistics are studying to classify and recognize emotions and calculate their accuracy. The usage of machine learning techniques to classify sentiments (Pang 2019), how to classify emotional models with some technical approaches (Rout et al. 2018; Bruna, Avetisyan, and Holub 2016), and some benchmarks to detect and calculate the accuracy of emotions or sentiment (Parupalli, Rao, and Mamidi 2018; Ren, Cheng, and Han 2017) as well as how to recognize emotions by approaching reading texts directly or by reading the writings (Daşdemir, Yıldırım, and Yıldırım 2017; Li, Chao, and Zhang 2019; Geethanjali et al. 2017; Saif M. Mohammad 2015; Yu et al. 2016). Even because it is so complex in analyzing text, there is now a new method of machine learning that is deep learning with a neural network approach (Antariksa, Purnomo WP, & Ernawati, 2019; Habimana, Li, Li, Gu, & Yu, 2020; Schmidhuber, 2014).

Furthermore, researchers who examine the recognition of emotions with a variety of methods and techniques in machine learning aims to find out what emotional sentences emerge quite often in social media. In addition, some articles also do a lot of real-time emotional recognition studies with various stimulation media (Handayani, 2017; Iurp et al., 2016; Kamaruddin & Abdul Rahman, 2013; T. M. Li, Chao, & Zhang, 2019; Wang, Nie, & Lu, 2014). So, what needs to be done is to do an experiment that detects emotional words that exist on social media with machine learning techniques, then it will be examined whether the sentiment of the text affects the reader by taking data from the reader's brain waves using a brain wave recorder. Also, one of the centers of activity in the human body when reading aloud is the brain. However, until now, EEG research that specifically examines reading aloud, especially in Indonesian, has never been carried out (Nurhadi & Rahma, 2017). Therefore, the contribution of this research is twofold:

1) Collecting corpus through social media (in this case, using Twitter data) and
2) The corpus will be used as a source of stimulation data in detecting emotions in real-time captured using EEG.

In Carley et al's article, entitled Twitter Usage in Indonesia, Indonesia is indeed in the top five user countries and invests in social media in general, and Twitter in particular. In early 2012 Indonesia's Twitter user population was 29.4 million, the fifth-largest in the world. In 2013, CNN nicknamed Indonesia "the country of Twitter". In 2014, Indonesia ranked the fifth tweeting country with 29 million Indonesian users, and Jakarta was responsible for 2.4% of the 10.6 billion Twitter posts made between January and March 2014 (Carley, Malik, Kowalchuck, Pfeffer, & Landwehr, 2018). Meanwhile, the corpus of emotions using Indonesian is not yet already available, so in this research, data from Affective Norm English Words (ANEW) was referred available and the advancement of that study (Bradley & Lang, 1999; Delatorre, Salguero, León, & Tapscott, 2019).

Previously, we did preliminary research on emotion/sentiment Indonesian corpus based on ANEW (Hulliyah, Awang Abu Bakar, and Ismail 2018). The study began by building a set of emotion words as a database and subsequently expanding this set to validate the correlation with the proposed data set (corpus) above for identification of sentiment classification (Hirschberg & Manning, 2016). The output obtained is to get a collection of emotional words as a reference for labeling, then we will classify the data using several techniques in analyzing sentiments using machine learning and deep learning method to obtain emotional classifications. The Twitter application is used by users to express personal opinions or clarification of something (public statement). Therefore, the text used in writing the Twitter status contains many emotional words as expressions of their feelings. Also, the status of Twitter is often used as a medium to influence the people who read it.

To summarize, three main components must be explored in answering the above questions:

1) The availability of dataset/corpus of emotional words,
2) Finding and analyzing the good modeling of emotion recognition based on Indonesian sentiment text with Twitter data as the dataset, and
3) Looking for the representative model based on the real-time dataset using EEG tools.

1.1.1 Affective Norm English Words (ANEW)

The emotion domain has a direct relationship with the identification of relevant pieces of social cognition and the relevant textual data for understanding sentiment. The affective norm in English word (ANEW) was introduced by Bradley, et.al, which provided the standard of emotional data set ratings for a big figure of terms in the English language, it collects the affective terms in three areas; valence, arousal, and dominance (Bradley and Lang 1999).

ANEW is a corpus that was created by Bradley in 1999 where he did research using Self-Assessment Manikin (SAM) sheets. In the questionnaire sheets distributed to the participants, about 100-150 words were made to experiment to get a rating for each word in the affective word group and the respondents were divided into several groups of male and female. The format dimensions are pleasure, arousal, and dominance. The ANEW corpus aims to provide standardization of the base material for the researchers who conducted the study of emotion and attention.

The ANEW corpus is in great demand by researchers in psychology and computer science for understanding the basic emotions classification, as this has become a reference to find a deeper meaning into human emotions discretely or dimensionally. The affective ratings of ANEW detected 1024 effective words in classifying emotions based on the dimensions of valence and arousal with 4 quadrants of classification (Delatorre et al., 2019).

In this research, we take all the affective words in ANEW, which will then be translated into the Indonesian language by certified linguists. Then we involved 30 students as participants for the primary data to obtain a group of emotion words based on valence and arousal (VA) where valence determines the range from negative to positive of emotions, and arousal denotes the range from calmness to exciting of emotions. We take the words that have the highest value in every 4 groups of basic emotions namely: happy, sad, fear, and angry. The results of this preliminary research, are used as raw data when building corpus sentiment consists of using 2 methods based on physiology and psychology approaches (Hulliyah, Awang Abu Bakar, and Ismail 2018).

1.1.2 Sentiment Analysis

Emotional recognition research in the last ten years has attracted many researchers, in line with a large number of microblogging and social media that have emerged along with the development of internet technology. In the business world and government, the status or comments of netizens or citizens of a country becomes an important part of the decision-making process, because from social media we can identify one's sentiments/emotions. The sentiment analysis technique in the Natural Language

Processing (NLP) field is widely used to find out the emotions or sentiments that occur whether positive, negative, or neutral.

Sentiment analysis (SA) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to identify, extract, measure, and systematically study affective states and subjective information (Pasayat, 2018). Almost all aspects of life that involve social media is working on and developing their work. Furthermore, millions of transaction processes are currently carried out online, which means that there are millions of data in the form of text. Therefore, understanding text or words is the center of attention for many institutions to understand the desires or feelings of consumers. In other words, sentiment analysis determines the feelings or emotions contained in the text.

However, there are many algorithms offered in the sentiment analysis process. Choosing the right algorithm, very much depends on the available data and the desired output needs. Therefore, conducting benchmarks becomes important, to ensure the results of this research are optimal or not.

In this study, we conducted an analysis of which algorithms are effective in classifying emotions. We chose Twitter data as a data source based on the emotional words that we obtained from previous preliminary research.

1.1.3 Machine learning

An important part of the machine learning process is analyzing the algorithm that will be used. To produce optimal output. The process of training computers to understand human language words must be done, by taking training data and test data. Furthermore, the available models need to be compared to get the best results. The three chosen algorithm models are SVM (with the concept of classification), random forest (using