Emotion Detection from Tweets on Twitter using Boyer-Moore Algorithm

Maharani Devira Pramita - 13516142

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia

[13516142@stei.itb.ac.id](mailto:13516142@stei.itb.ac.id) **|** [devipram99@gmail.com](mailto:devipram99@gmail.com)

*Abstract*—Delivering thoughts in textual form is usually debatable in problem of effectiveness, because of possibility misunderstanding among speakers. This misunderstanding is caused by missing intonation or voice tone in textual form, makes people have their own opinion about emotional state of interlocutors. There is a way to detect what emotion behind text delivered by interlocutors by using pattern matching. Moreover, to fasten the matching process, this paper will discus about pattern matching using Boyer-Moore Algorithm which is faster than comparing each character one by one.

*Keywords*— emotion*,* tweet*,* Boyer-Moore, pattern matching

# I. Introduction

Social media has been booming in Indonesia and most probably among the youth. This moment supports freedom of expression because there is no boundary in social media. People can fake their profile account and write any opinions toward how the government works.

One of social media that is effectively used for expressing thoughts is Twitter. Any thoughts, including positive and negative, could be delivered through tweets on Twitter. Both delivered with positive-minded or full of emotion sometimes cannot be detected easily because of absence of voice tone int text. Moreover, emerged speculation could be different in one to another people’s mind.

Misconception that may be appear is really based on someone’s mindset. A tweet of positive critical thoughts may ignites misunderstanding and pro contra then feuds could happen just because of not paying attention to keywords in detail to really understand the way he talk. Communication will not go effectively without understanding each other emotions.

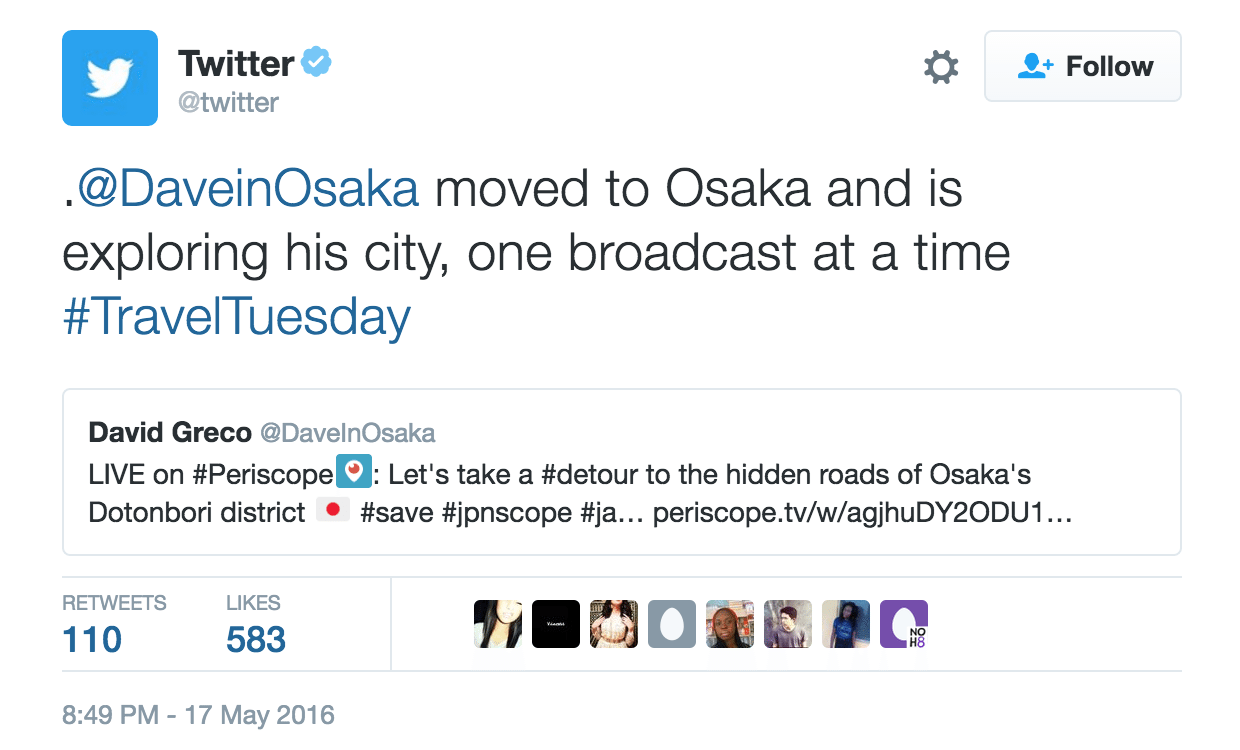
Emotion Detection from text is a recent field of research that is closely related to Sentiment Analysis. Sentiment Analysis aims to detect positive, neutral, or negative feelings from text, whereas Emotion Analysis aims to detect and recognize types of feelings through the expression of texts, such as anger, disgust, fear, happiness, sadness, and surprise.

However, on each tweet, there is (are) some keyword(s) of which emotion tried to be delivered. In this paper will be discussed about some potential keywords toward some emotion such as happiness, sadness, and anger. By using pattern matching, could be classified which emotion is behind tweet.

# II. Theory Based

## A. Twitter

Twitter is a network media where people can upload sort of words limited to 140 characters called a tweet. In this section, people can write anything down then upload it and shared into timeline. Any kind of words could be written down, from words full of happiness or sadness or even anger. Usually, people use Twitter to express attention towards newly updated news or their thoughts of something tickling their minds or even boredom. Besides expressing thoughts into paragraph, people can also make hashtag to make something viral, shoutout to the world.



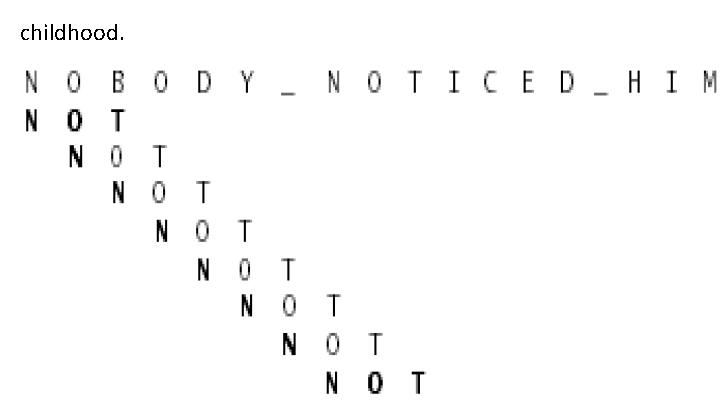
*Picture 1. Example of tweet from Twitter Official Account*

## B. Pattern Matching

Pattern matching in computer science is the checking and locating of specific sequences of data of some pattern among raw data or a sequence of tokens. Unlike pattern recognition, the match has to be exact in the case of pattern matching. Pattern matching is one of the most fundamental and important paradigms in several programming languages. Many applications make use of pattern matching as a major part of their tasks.

Some people are confused between pattern matching and pattern recognition. Whereas pattern recognition looks for a similar or most likely pattern in a given data, pattern matching looks for exactly the same pattern. Pattern matching is not considered part of machine learning, although in some cases it leads to similar results as pattern recognition, while pattern recognition is a supervised machine learning based.

In this paper, pattern matching is a way to detect some emotional words towards text in Twitter. It matches on characters between emotional word (called pattern in pattern matching) and text as the resource. There are three algorithm approaches, they are Brute Force, Knuth-Morris-Pratt (KMP), and Boyer-Moore which is to be explained further more in this paper.



*Picture 2. Example of pattern matching using Brute Force Algorithm (match character one by one)*

## C. Boyer-Moore Algorithm

Boyer-Moore algorithm is one of the most efficient string matching algorithm. It was founded in 1977 by Robert S. Boyer and J Strother Moore. This algorithm is very suitable for a text or strings which have large alphabet and when the pattern is too long. Boyer Moore algorithm is specified as follows

//Preprocessing

Compute R(x) for each x ϵ Σ;

Compute L’(i) and l(i) for each i = 2, … , n+1;

//Search

k ← n

while k ≤ m do

i ← b; h ← k

while i > 0 and P[i] = T[h] do

i ← i – q; h ←h-1;

endwhile;

if i = 0 then

Report an occurrence at T[h+1…k]

k ← k + n – l(2);

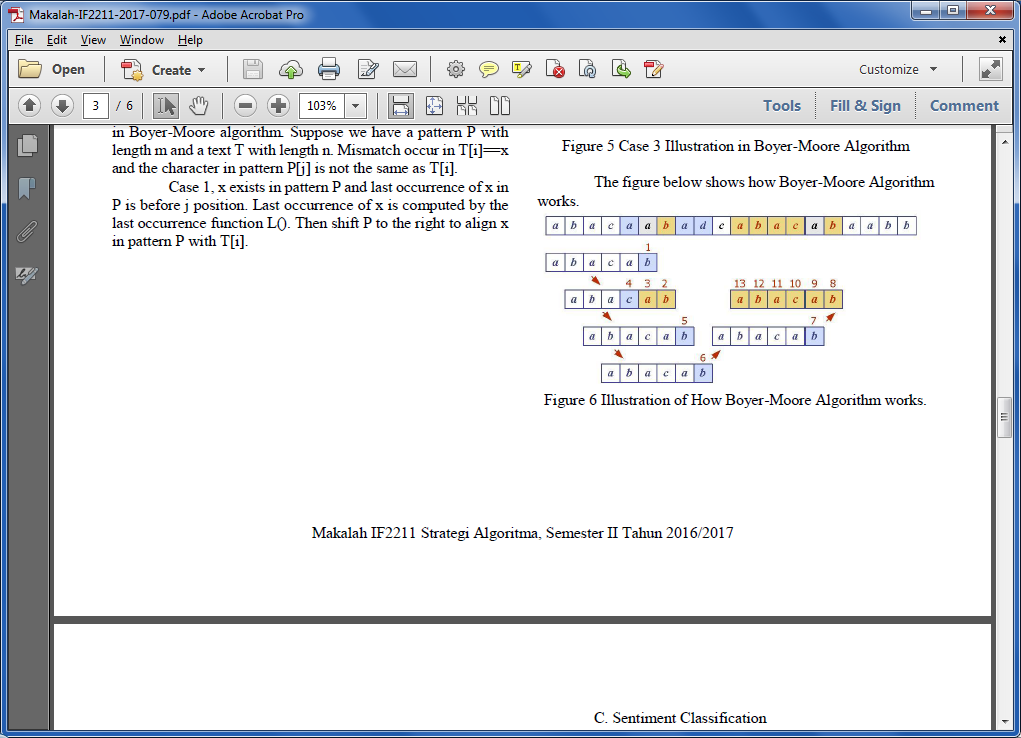
else //mismatch at P[i]

Increase k by the maximum shift given by shifting rule.

endif;

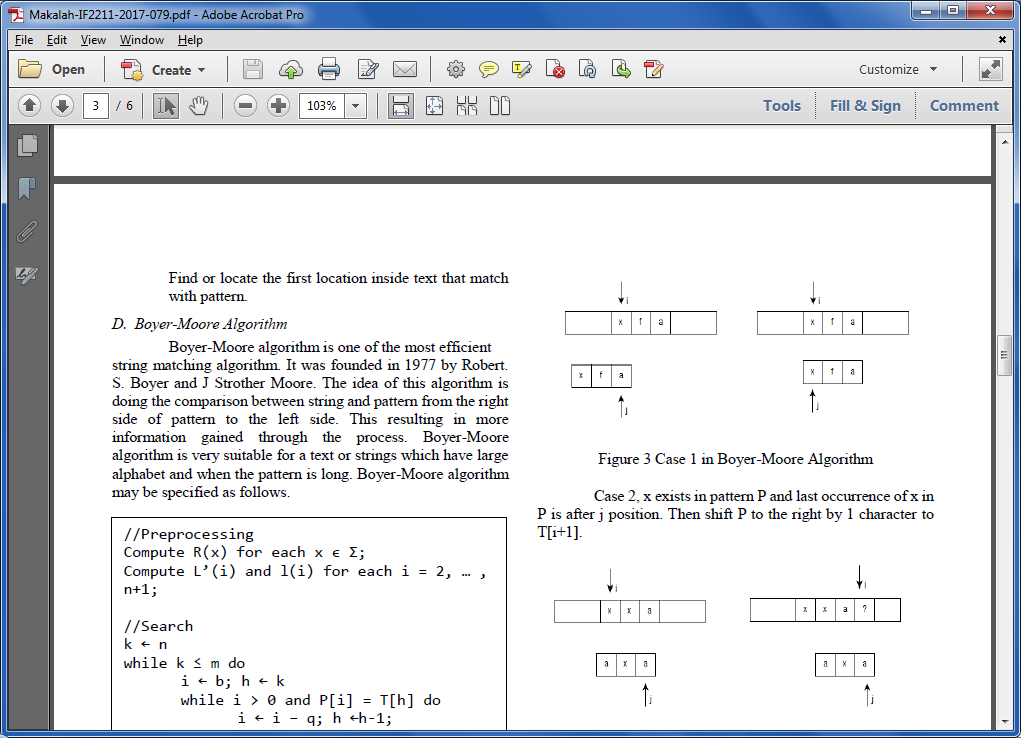
endwhile;

In Boyer-Moore Algorithm, matching process is done by character jumping starting from last character of pattern and there are three approaches based on cases. First case is when character ‘p’ in pattern doesn’t match with character ‘x’ in text while ‘x’ is exist before ‘p’ in pattern. Second case is when character ‘p’ in pattern doesn’t match with character ‘x’ in text while ‘x’ is exist after ‘p’ in pattern. Third case is when character ‘p’ in pattern doesn’t match with character ‘x’ in text while ‘x’ isn’t exist anywhere in pattern.



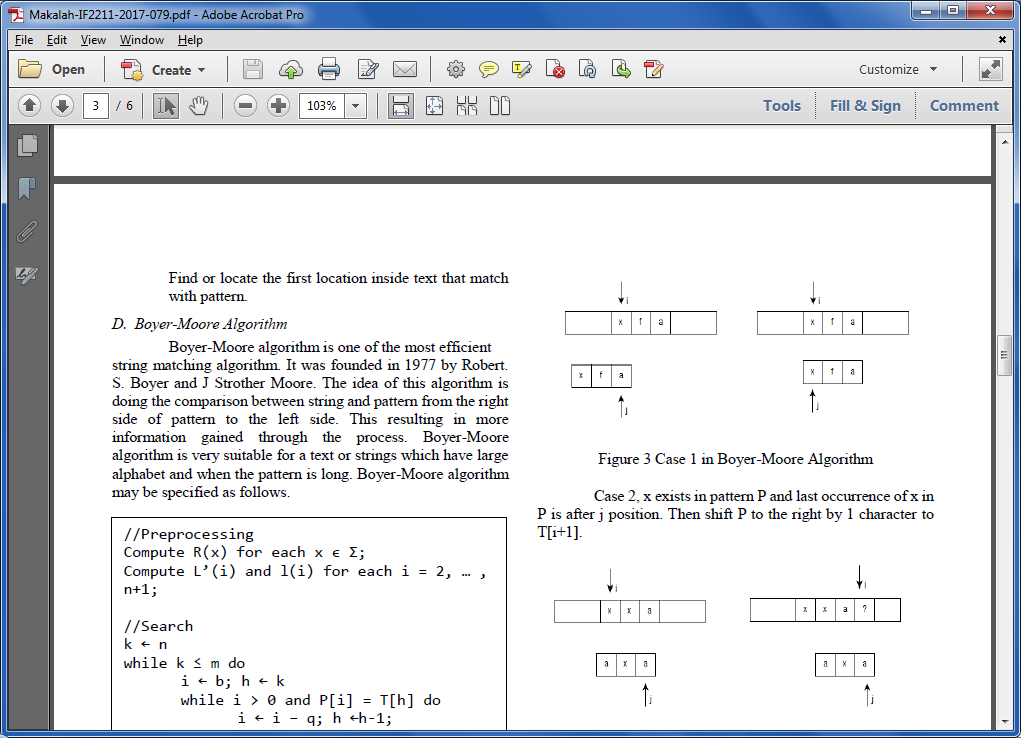
*Picture 3. Explain how Boyer-Moore Algorithm works*

For character jumping itself will be explained in this section. In the first case, pattern will be shifted right until character ‘x’ in text is matched with character ‘x’ in pattern.



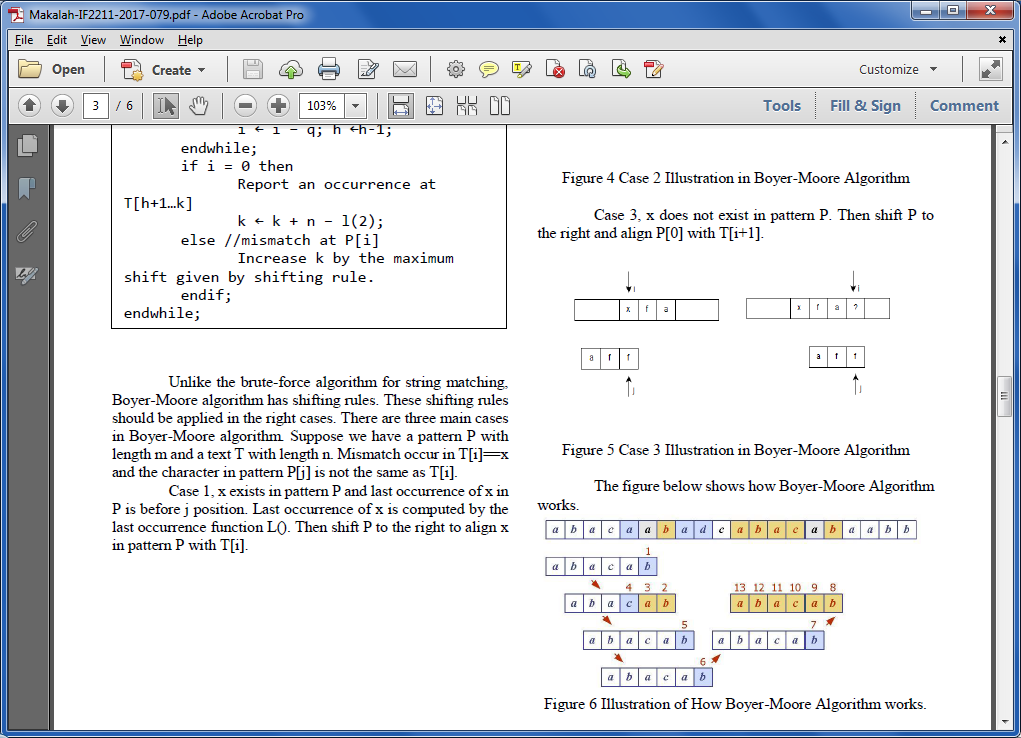
*Picture 4. Case 1 of Boyer-Moore Algorithm*

In the second case, pattern will be shifted left until character ‘x’ in text is matched with character ‘x’ in patter.



*Picture 5. Case 2 of Boyer-Moore Algorithm*

In the third case, because there is no character ‘x’ exist in pattern, so that pattern will be shifted right all.



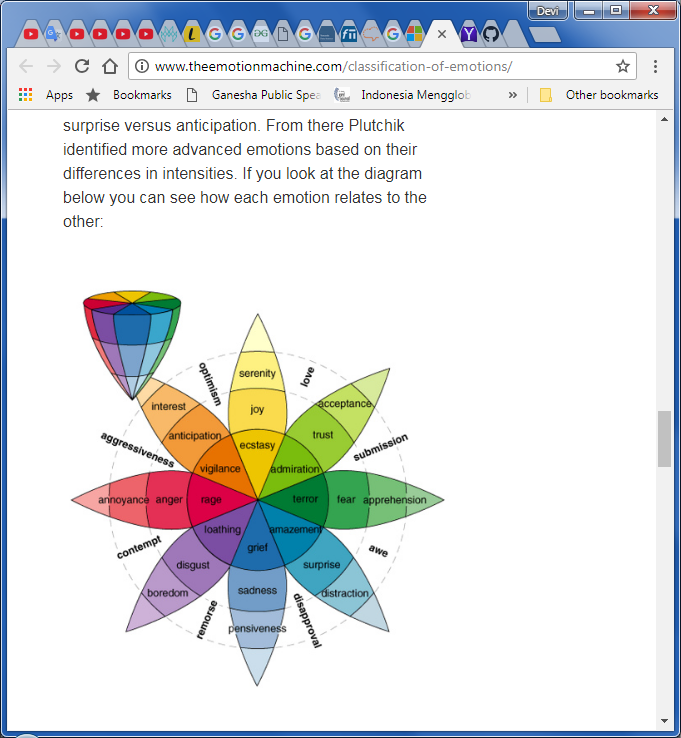
*Picture 6. Case 3 of Boyer-Moore Algorithm*

These procedures is really fasten the matching process compared with brute force which shift pattern one by one character.

## D. Emotion

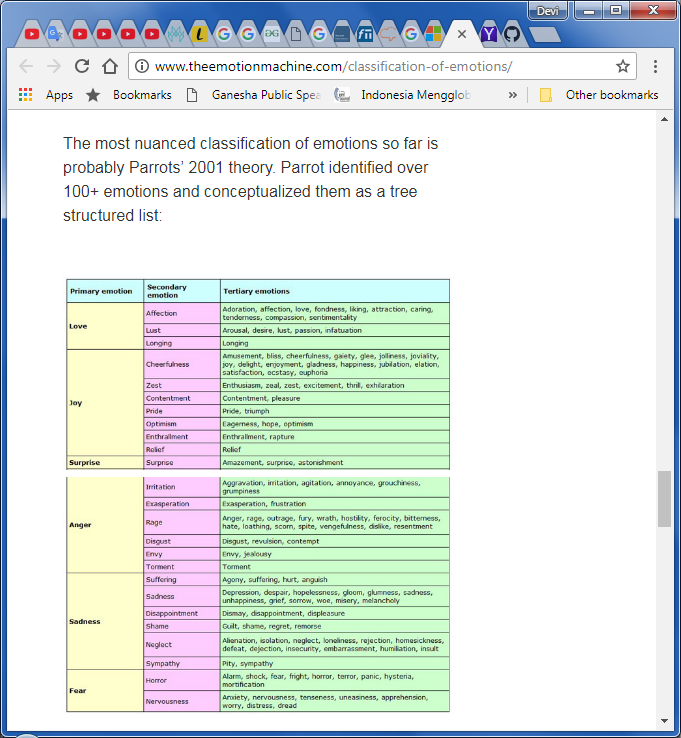
Psychologists did tackle to differentiate one emotion to another, because it involves many different factors for each emotion. When analyzed in everyday lives, “happiness”or “anger” could be come in many different degrees, qualities, and intesities. Despite the difficulties, plenty of psychologists have attempted to classify our emotions into different categories, based on “pleasure” or “pain”.

There are many theories come up about group of emotion. Ekman’s List of Basic Emotions (1972) stated six basic emotions, they are anger, disgust, fear, happiness, sadness, and surprise. Plutchik’s Wheel of Emotions (1980) demonstrated how different emotions can blend into one another and create new emotions in between. There are eight primary bipolar emotions, joy vs sadness, anger vs fear, trust vs disgust, surprise vs anticipation. After combined, it will come up with new emotion in between two bipolar emotion, just like below diagram explains furthermore.



*Picture 7. Diagram of Emotion by Plutchik*

Parrot’s Classification of Emotions (2001) has been the most nuanced. It identifies over 100+ emotions and form it as a tree structured list, where the first level is love, joy, surprise, anger, sadness, fear, and the other levels are explained more in table.



*Picture 8. Table of Emotions as representative of Parrot’s Classification tree based structure*

Emotion detection from text has many applications. For example an stuxent sending harsh email to his colleague or lecturer. A tool that can analyze the email for emotions and alert the employee about its harshness before sending it comes in very handy to protect the student’s state. Another example is emotion-based search engine that ranks documents according to the emotion requested by the user. Such an engine could prove to be very beneficial to users in a certain emotional state and can improve the effectiveness of the information retrieval process. Other useful tools that can benefit from emotion detection include recommender systems that aim to personalize recommendations based on the user’s emotions.

Emotion classification can be divided into two different categories : coarse-grained and fine-grained level (positive or negative) can be accurately perceived from text. There is Linguistic Inquiry and Word Count (LIWC) to classify emotions as positive or negatice. They found that positive emotions are expressed using more exclamation marks and words, while negative emotions are expressed using more affective words. This method is limited to positive or negative emotions (happiness vs sadness).

Besides that, classifying emotions on a fine grained level (Ekeman emotions) requires semantic and syntatic analysis of the sentence and can be done using three methods: (1) Keyword-based detection, (2) Learning-based detection, (3) Hybrid detection.

Keyword-based detection. This classifying method is done by searching for emotional keywords in the input sentence. A research done by Osgood used a multidimensional scaling for visualizing the affective words in order to compute similarity ratings between them. The dimensions are “evaluation”, “potency”, and “activity”. “Evaluation” quantifies how much a word referes to a pleasant or unpleasant event, “potency“ quantifies the emotional intensity of a word (strong or weak), and “activity” refers to word whether it is passive or active form. The problem is this method suffer from ambiguity in the keyword which may have different meanings according to usage in text and incapability of recognizing emotions within sentences that do not contain emotional keywords, and lack of linguistic information.

Learning-based detection. This method use machine learning approaches, emotion is detected by using classification approaches based on training dataset. Strapparava developed a system that used several variations of Latent Semantic Analysis to identify emotions in text when no affective words exist. However, their approach achieved low accuracy because it is not context sensitive and lacks the semantic analysis of the sentence. On this method, input data is preprocessed first and labeling it using a classifier.

Hybrid detection. In this method, emotions are detected by using a combination of emotional keywords and learning patterns collected from training datasets, in addition to information from different sciences, like human psychology. Wu’s novel stated that level emotion mining based on detecting predefined semantic labels and attributes of the sentence, then classifying emotions based on psychological patterns of human emotions called emotion generation rule (EGR). F. Chaumartin developed a linguistic rule-based system UPAR7, which use dependency graph where the root is considered the main subject. Each word in the sentence is rated individually for each emotion, then the rating of main subject is boosted, as it is more important that the rest of the words in the sentence. Yang proposed a hybrid model that includes lexicon-keyword spotting. The results generated form the aforementioned techniques are integrated using a vote-based system. Ghazi built hierarchical classification which defines in multiple levels while classifying emotions by first classifying whether a sentence holds an emotions or not, then classifying emotion as either positive or negative, and finally classifying emotion on a fine grained level.

# III. Problem Solving Analysis

From people’s opinion on Twitter, it can be known probability emotion that they tried to deliver by scanning from each word on tweets. The method would be like this :

*Picture 9. Research methods step by step*

This paper refers to Ekman’s List of Emotions which are happiness, sadness, anger, disgust, fear, surprise, but only use happiness, sadness, and anger as research. Also applies method fine-grained classification using keyword-base which is implemented on making a dictionary of emotional keyword.

## A. Dictionary

Dictionary is a collection of keywords that is unique for each emotion, refers to happiness, sadness, and anger. Keywords from this dictionary will be used as pattern to be matched with tweets on Twitter timeline.

This dictionary is made based on number of occurrences on Twitter, which is most probably used to express certain emotion. Suppose below are example of keywords each represents people’s emotion through their tweet written in Bahasa Indonesia

1. Happiness : ‘wow’, ‘asyik’, ‘yeay’
2. Sadness : ‘yaaah’,
3. Anger : ‘apa sih’, ‘?!’, ‘kok gitu sih’, ‘egois’

## B. Text

Collecting tweets as group of data from Twitter using crawling process. This data will be then used as source text. As other research usually tokenized their input text to find match word more easily, in this paper it isn’t needed because of usage Boyer-Moore Algorithm that helps pattern matching process much more faster.

Many tweets could be crawled from Twitter, but in this paper only takes the ones contain emotional words. The text crawled could be about public opinion towards news update in political, product survey, or movie reviews. These kind of tweets most probably contains adjective words that represent opinions.

To identify contextual tweet, one of key determination is by seeing the hashtag such as

* Political issue : #DUAPERIODE refers to Joko Widodo’s second reign proposal, #2019GANTIPRESIDEN refers to Prabowo as he wants to be candidate on 2019 election
* E-commerce promotion : #MulaiAjaDulu refers to Tokopedia, #TravelokaDulu refers to Traveloka, #TiketKemanapun refers to Tiket.
* Movie review : #INFINITYWARS, #FF8

Choosing these tweet could be done after crawling tweets from specific account then filterize it using those hashtags or any other popular tags on Twitter trending topics.

## C. Pattern Matching

In group of tweets, keep it still separated by each tweet uploaded, in a form just like what is remained on Twitter, so that tweet classification could be done easier.

Pattern matching works between words from dictionary being match with source text (tweets). Each tweet is matched with every word from dictionary, then make score once it is matched. If a tweet contains more than one emotional words from different categories, then score of each category will be compared.

Using Boyer-Moore algorithm, it fasten matching process, because once unmatch and the unmatched character in text doesn’t exist anywhere on pattern, then pattern will be shifted right as much as length of character in pattern. If the unmatched character in text is exist on pattern, there are two possibilities, if it exists before unmatched character in pattern then it will be shifted right until matched between two characters, and if it exists after unmatched character in pattern then it will be shifted left until matched between two characters.

## D. Emotion

After finish matching all pattern from dictionary towards a tweet, then compare the score of each emotion categories. Takes the biggest one. Therefore, the probabily emotion behind written text is revealed, could be happiness, sadness, or anger. Same procedures are implemented on next tweets.

This step is kind of text classification which classified tweets into different group of emotions labeled with happiness, sadness, and anger.

# IV. Implementation and Analysis

## A. Sample Data

Example tweets crawled from Twitter

|  |  |
| --- | --- |
| 1 | wow! Bagus banget atraksi sirkus tadi |
| 2 | yaaah kenapa bisa mati lampu waktu aku belajar |
| 3 | Akhirnya aku bisa pergi ke dufan, yey! |
| 4 | Ih apa sih kok sikapnya jutek padaku |
| 5 | Wow egois sekali dia pikir pendapatnya saja yang selalu benar ?! |

*Picture 10. Table of Sample Tweets in Bahasa Indonesia*

## B. Results

|  |  |
| --- | --- |
| 1 | Happiness |
| 2 | Sadness |
| 3 | - |
| 4 | Anger |
| 5 | Anger |

*Picture 11. Table results of emotion detection from tweets*

## C. Analysis

On first tweet, there is found pattern of ‘wow’ which comes from group Happiness so that this tweet is classified into represent happiness.

On second tweet, there is found pattern of ‘yaaah’ which comes from group Sadness so that this tweet is classified into represent sadness.

On third tweet, there is found pattern of ‘yey’ which slightly match with ‘yeay’ that comes from group Hapiness, but since it isn’t matched with anything from keyword in dictionary, so that this tweet is classified into have no emotions or neutral.

On fourth tweet, there is found pattern of ‘apa sih’ which comes from group Anger so that it is classified into represent anger.

On last tweet, there is found pattern of ‘?!’ and ‘egois’ which comes from group Anger also pattern of ‘wow’ from group of Happiness. Although this tweet comes up matched with keywords from different group, each group is rated. Anger has score of 2 and Happiness has score of 1. Conclude with the biggest one, therefore this tweet represents anger.

# V. Conclusion

As conclusion, pattern matching could be used to match keyword which usually comes out when people express their feeling in textual form in a rough way. Boyer-Moore Algorithm, matching in more effective way than to match character one by one, is really helpful to pattern matching process.

Unfortunately, it doesn’t work in effective way which has to match every single word from dictionary onto single tweet, although using Boyer-Moore algorithm, the efficient method in pattern matching. Besides that, other detection problem is keyword in dictionary are not fully representative the behaviour of usage word in everyday tweets. One specific word could be used in those three different emotions, but always based on its context. This situation makes prediction isn’t accurate.

This kind of case could be develop furthermore in text classification using machine learning algorithm for example in Semantic Analysis or Hybrid Classification in Fine Grained Method. It is needed large datasets in bag of keywords so pattern choice is wider then the emotion detection that appear to a tweet could be more accurate. It is needed a large trained dataset first which is preprocessed and classified into labeled features.

From trained dataset, it can be known people’s behaviors more significantly, just like which word tends to be used for expressing happiness or sadness or anger. Besides that, we can also see the context more wider, not only specific emotional keywords in adjective verb, but also the other supportive words that is used simultaneously with emotional keywords.

Furthermore, using machine learning algorithm that used labeled dataset, there can be more group of classification that comes up with. Not only happiness, sadness, or anger, also other Ekman’s emotions like disgust, fear, and probably surprise which is not used to use in general classification because of its factor.

# VI. Acknowledgment

Maharani Devira Pramita, as the author of this paper would like to express gratitude towards Dr. Masayu Leylia Khodra, S.T., M.T. as the lecturer of IF2211 “Strategi Algoritma” and to my family and friends for their given supports so that I could finish this paper.

# References

1. Boyer, Robert S. and Moore, J Sthroter, “A Fast String Searching Algorithms”. in Communications of the ACM, vol 20 Issue 10, Oct 1977, pp. 762-772 W.-K. Chen, *Linear Networks and Systems* (Book style)*.* Belmont, CA: Wadsworth, 1993, pp. 123–135.
2. Klipelainen, Pekka (2005). Lecture 3 : Boyer-Moore Matching [PDF Document]. Retrieved from www.cs.uku.fi/~kilpelai/BSA05/lectures/slides03.pdf B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.
3. Shaheen, Shaheen and El-Hajj, Wassim and Hajj, Hazeem and Elbassuioni, Shady, “Emotion Recognition from Text Based on Automatically Generated Rules” *IEEE Data Mining Workshop.*, 2014.
4. S.N. Shivhare “Emotion Detection from Text” Department of CSE and IT, Maulana Azad National Institute of Techonology*. India*, 2012.
5. Munir, Rinaldi, Diktat Kuliah IF2211: Strategi Algoritma. Bandung: Institut Teknologi Bandung, 2007

# Pernyataan

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 14 Mei 2018

******

Maharani Devira Pramita - 13516142