

# Boyer-Moore Algorithm in Dictionary-based Approach Sentiment Analysis for Marketing Research

Annisa Nurul Azhar - 13515129

*Informatics Undergraduate Program*

*School of Electrical Engineering and Informatics*

*Bandung Institute of Technology, Jl. Ganessa 10 Bandung 40132, Indonesia*

*annisanurulazhar@students.itb.ac.id • annisanurulazhar@yahoo.com*

**Abstract**—Every business' success depends on its marketing. Its marketers should make sure they get the right product at the right place with the right price to the right person. This could only be achieved with good marketing strategy. In order to develop good marketing strategy, we need to do marketing research. Nowadays, lots of businesses use quantitative data only, such as 5-star rating for their marketing research. This paper will further discuss about alternative tool for marketing research which is sentiment analysis and how string matching algorithm (Boyer-Moore) implemented in the analysis.

**Keywords**—string matching, sentiment analysis, marketing research

## I. INTRODUCTION

Not every business owner in this world understands that their marketing plays a great role in their business' success. Most of them still concentrating in how they could develop the quality of their product but they rarely give some efforts for developing their marketing strategy. The fact is, it does not matter how good the product is, if their marketers marketed ineffectively, there is a huge possibility that the business will not run well.

Second problem is some businesses only do quantitative research their market research. This methodology is criticized since opponents describe this as ignoring inherent subjectivity of human social interactions.<sup>[1]</sup> There is always an option to do qualitative research only but this method also criticized because opponents think this method as subjective and lacks structural coherence. In order to address those methodological criticism and improve our research, we should integrate both methods into a comprehensive research design.

Sentiment analysis can be used as complementary research techniques. Actually, sentiment analysis is not new in marketing since marketers have been analyzing sentiments from surveys, customers' comment card, interviews, and focus groups. The problems are the use of these tools is subjected to the researcher presence and small sample size. Sentiment analysis address those problems by systematically collecting data and analyzing sentiments from large sample size data in real time. Dictionary-based approach is one of sentiment analysis' approach which relies on set of opinion words or phrases called dictionary. For classifying sentiments, we could involve string matching

algorithm, more specifically Boyer-Moore algorithm since it is suitable for searching English text.

## II. THEORY

### A. Marketing Research

#### 1. Definition and Purpose

Marketing research is the research of marketing process in a company. According to American Marketing Association, the marketing research is a process or set of processes which connects customers, producers, and end users to the marketers through information. Marketing process consists of several steps such as problem definition, determining needs and data collection method, determining data sample method, collecting and analyzing data, error checking, and creating report.<sup>[1]</sup>The information extracted from marketing research could be used to identify and define marketing opportunities and problems which are very essential for developing good marketing strategy.

#### 2. Methods

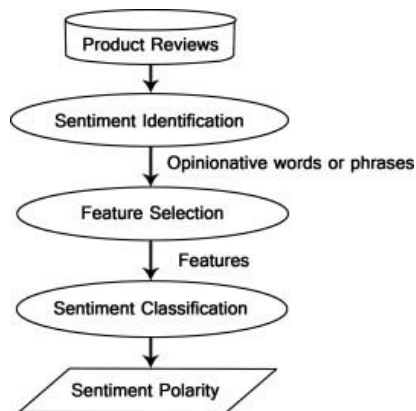
There are two basic approach in research, qualitative and quantitative research. Qualitative research involves feelings and impressions because its way of work is finding what people think and how they feel towards particular marketing actions. Meanwhile quantitative research involves number rather than feelings or impression and it focuses on measuring objective facts.

### B. Sentiment Analysis

#### 1. Definition

In *Sentiment Analysis Algorithms and Applications: A Survey*, Walaa Medhat et all explained, "Sentiment Analysis or Opinion Mining is computational study of people's opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics." However, Sentiment Analysis and Opinion Mining is slightly different. Opinion Mining focuses on extracting and analyzing people's opinion meanwhile Sentiment Analysis are focusing on identifying sentiment in a given text then

analyze it. Figure below shows sentiment analysis process in product review.



Source: Walaa Medhat et al. 2014. Sentiment Analysis Algorithms and Applications: A Survey  
Figure 1 Sentiment Analysis Process of Product Review

## 2. Feature Selection in Sentiment Classification

First step in sentiment classification problem is to extract and select text features. These are some of the current features.

### 2.1 Terms presence and frequency

These features either give the words one-zero value, for example one if the word exist and zero if does not or using the frequency to indicate the importance of the features.

### 2.2 Parts of Speech

These features finding adjectives since they are great indicators for opinion.

### 2.3 Opinion Words and Phrases

These features are words which is commonly used to express opinions towards particular condition such as love, hate, good, bad, etc. Meanwhile some phrases do express opinions without using opinion words.

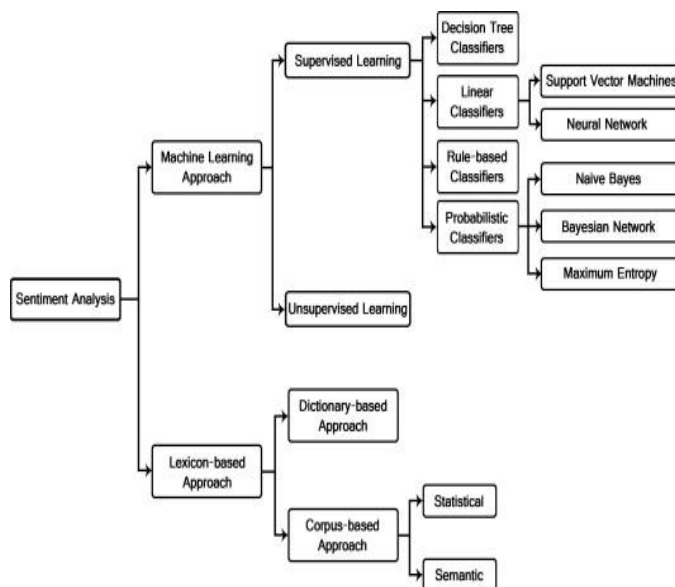
### 2.4 Negations

The existence of negation words such as *not* may change opinion orientation.

## 3. Sentiment Classification Techniques

Sentiment classification techniques can be divided into machine learning approach, lexicon-based approach, and hybrid approach. Machine Learning approach uses machine learning algorithm and linguistic features. Lexicon-based approach uses a collection of sentiment terms. On the other side, Hybrid approach combined both machine learning approach and lexicon-based approach.

Further, both machine learning approach and lexicon based approach could be divided into more specific classification as shown in Figure 2 below.



Source: Walaa Medhat et al. 2014. Sentiment Analysis Algorithms and Applications: A Survey  
Figure 2 Sentiment Classification Techniques

Later in the implementation, we would go with Dictionary-based approach. This approach relies on set of opinion words or phrases called dictionaries which can be created manually or automatically by using seed words to expand list of words (Turney, 2002).

## C. String Matching

### 1. String Concept

Suppose  $S$  is a string with length  $m$ .

$$S = x_1x_2 \dots x_{m-1}$$

A prefix of  $S$  is *substring*  $S[0..k]$ .

A suffix of  $S$  is *substring*  $S[k..m-1]$ .

Where  $k$  is an index between 0 until  $m-1$ .

Examples:

S	T	I	M	A
---	---	---	---	---

All possible prefixes of STIMA are “S”, “ST”, “STI”, “STIM”. Meanwhile all possible suffixes of STIMA are “A”, “MA”, “IMA”, “TIMA”.

### 2. String Matching Problem

Given:

- a. text, long string which length is  $n$  character
- b. pattern, string with length  $m$  character ( $m < n$ ) which will be searched in text.

Find or locate the first location inside text that match with pattern.

#### D. 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. The idea of this algorithm is doing the comparison between string and pattern from the right side of pattern to the left side. This resulting in more information gained through the process. Boyer-Moore algorithm is very suitable for a text or strings which have large alphabet and when the pattern is long. Boyer-Moore algorithm may be specified as follows.

```

//Preprocessing
Compute R(x) for each x ∈ 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;

```

Unlike the brute-force algorithm for string matching, Boyer-Moore algorithm has shifting rules. These shifting rules should be applied in the right cases. There are three main cases in Boyer-Moore algorithm. Suppose we have a pattern P with length m and a text T with length n. Mismatch occur in  $T[i]=x$  and the character in pattern  $P[j]$  is not the same as  $T[i]$ .

Case 1, x exists in pattern P and last occurrence of x in P is before j position. Last occurrence of x is computed by the last occurrence function L(). Then shift P to the right to align x in pattern P with  $T[i]$ .

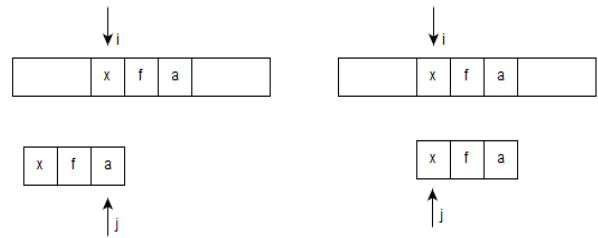


Figure 3 Case 1 in Boyer-Moore Algorithm

Case 2, x exists in pattern P and last occurrence of x in P is after j position. Then shift P to the right by 1 character to  $T[i+1]$ .

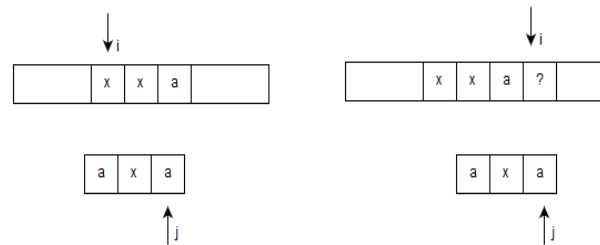


Figure 4 Case 2 Illustration in Boyer-Moore Algorithm

Case 3, x does not exist in pattern P. Then shift P to the right and align  $P[0]$  with  $T[i+1]$ .

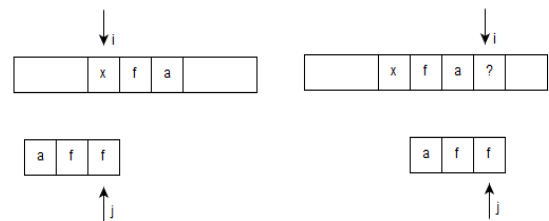


Figure 5 Case 3 Illustration in Boyer-Moore Algorithm

The figure below shows how Boyer-Moore Algorithm works.

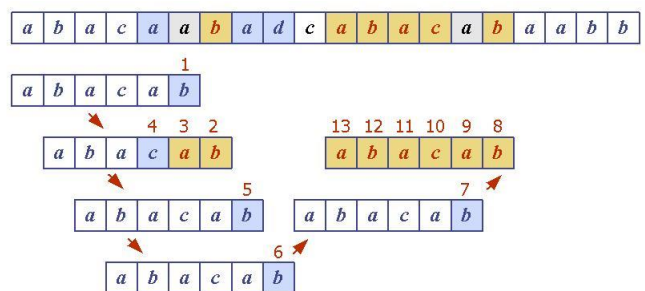


Figure 6 Illustration of How Boyer-Moore Algorithm works.

Time complexity of Boyer-Moore algorithm is  $O(nm)$  for best and average case. For worst case, the time complexity is  $O(nm + A)$ , where  $A$  is the size of alphabet.

### III. PROBLEM SOLVING ANALYSIS

Suppose that there is a marketing action from a certain company in form of a campaign which was held at car-free-day area. The marketing manager wants to know what do people think and feel towards the campaign. So, he tries to find out using sentiment analysis. The flowchart below describes the process from collecting data until finally got the conclusion.

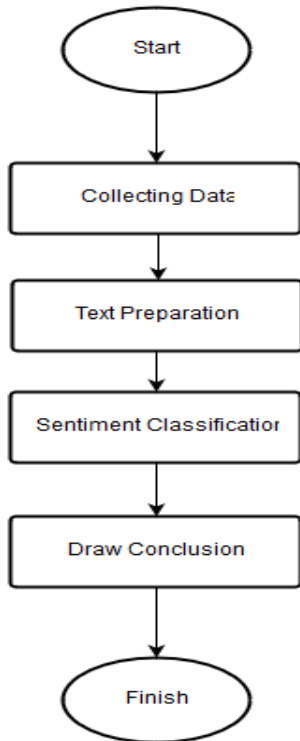


Figure 7 Flowchart

#### A. Data Collection

Opinions and comments from customers regarding particular marketing actions which is campaign at car-free-day area is stored in an external file that will be our input file.

#### B. Text Preparation

Before we analyze the opinions or comments, we have to “clean” the texts first. We should identify and remove non-textual contexts and any information which are not relevant to the course of analysis.

#### C. Sentiment Classification

After that, let us classify the opinions into three different groups; positive, negative, and neutral. Since we use dictionary-based approach, first we should make dictionaries of positive words and negative words. Table below shows sample of positive words and negative words.

Positive	Negative
Fun	Disturbing
Attractive	Noisy
Creative	Annoying
Entertaining	Bad
Love	Boring

Table 1 Sample of Positive Words and Negative Words

Then we use those words as patterns and search whether the text contains those patterns or not using Boyer-Moore algorithm. If it does contain the words from positive words dictionary then increment counter for positive and if it contains the words from negative words dictionary then increment counter for negative.

#### D. Conclusion

We could draw a conclusion by comparing the sum of positive words with the sum of negative words in an opinion or comment. If the number of positive is greater than the number of negative then the opinion classified as positive, else if the number of positive is less than the number of negative then the opinion classified as negative, and if the number of positive and negative are the same then the opinion classified as neutral.

By counting the number of positive opinion and negative opinion, we could find out whether the marketing action received more positive feedback or more negative feedback from customer. We could visualize the result into a pie-chart to make it clearer. Then the result could guide us to make the right decisions in determining the next marketing actions we should do, do we want to keep doing campaign or try another marketing action, for example.

#### IV. IMPLEMENTATION AND ANALYSIS

##### A. Sample Data

Comment/Opinion:
1. I love the campaign yesterday. It was so entertaining.
2. This need to stop. Yesterday's campaign was disturbing for me.
3. The campaign was attractive but sometimes it was very noisy.
4. The campaign sucks. I wish they could make it more entertaining in the future.

##### B. Result

Opinion ID	Positive	Negative	Conclusion
1	2	0	Positive
2	0	1	Negative
3	1	1	Neutral
4	1	0	Positive

Table 2 Result from Sentiment Analysis Process

##### C. Analysis

Let us analyze the first opinions. We can see that the opinion contains 'love' and 'entertaining' thus the sum of positive words is two. This first opinion does not contain any negative word. So, we can conclude this opinion as positive and if we checked semantically, it really is a positive one. Next, the second opinion contains 'disturbing' which is in the dictionary of negative words. So it increases the sum of negative words. Second opinion also does not contain any positive words. This opinion is negative and semantically, it is also a negative one. The third opinion contains 'attractive' which exist in our dictionary for positive words and 'noisy' which exist in our dictionary for negative words. Since the sum of positive words and negative words are the same so we could conclude this opinion as neutral. As for the fourth opinion, we know that the word 'sucks' is a negative word but it does not exist in our negative words library so it is not considered as a pattern and of course it does not cause any change to the sum of negative words. Meanwhile, our system which use string matching algorithm, finds positive word 'entertaining' in the fourth opinion. We know that in this context the positive words are being used to express the customer's hope towards the campaign in the future, not about what the customer think and feel

towards yesterday's campaign. But still, because of the pattern exist in the opinion so the sum of positive words is increases. The sum of positive words is greater than the sum of negative words so this opinion concluded as positive.

Finally from the analysis we have known that from four different opinions there are two positive opinions, one negative opinions, and one neutral opinions. Let us make a pie-chart based on the information that we got from our research to make it clearer.

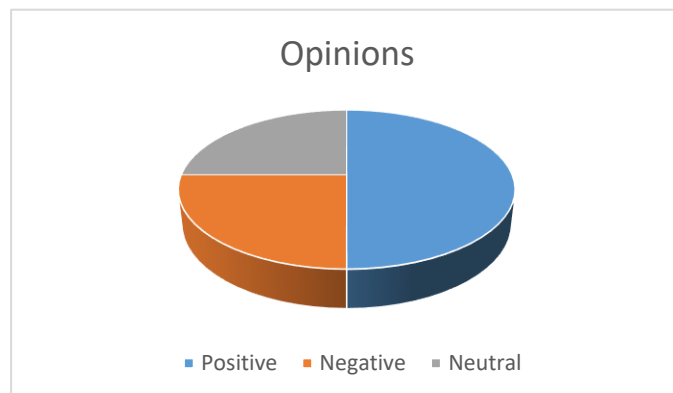


Figure 8 Pie-chart

Because our data sample size is small, the pie-chart does not seem to do much in helping us to draw conclusion. However, if our data sample size is large, the pie-chart or any other visualization techniques would be really helping us to extract the information.

In the pie-chart above we could see that the slice of positive opinions is larger than negative opinions or neutral opinions. This information could be use to decide whether we should keep doing the campaign in a car-free-day are or not. In this case, because the positive slice is bigger, we may tend to keep doing the campaign. Thus, we have made the decision which is keep doing the campaign for the marketing actions.

#### V. CONCLUSION

In this paper, we do simple marketing research with sentiment analysis as our research techniques. We use dictionary-based approach for classifying sentiments and we implement Boyer-Moore algorithm in sentiment analysis process as our string matching algorithm. We finally got the information which is needed to make decisions for determining our future marketing actions. However, due to lack of sample data and words in dictionaries, we found several opinions which classified into wrong categories. Therefore, I suggest the readers to do further research on another marketing actions by collecting larger data sample size and learning more about

marketing research and sentiment analysis so the information extracted from the research would be more reliable.

#### ACKNOWLEDGMENT

Annisa Nurul Azhar, as the author of this paper wants to express her gratitude towards Dr. Ir. Rinaldi Munir, M.T., Dr. Masayu Leylia Khodra, S.T., M.T., and Dr. Nur Ulfa Maulidevi, S.T., M.Sc., as the lecturers of IF2211 “Strategi Algoritma”. Special thanks for my beloved family and friends who have given their support so that I could finish this paper.

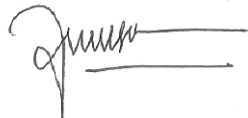
#### REFERENCES

- [1] R. Meena and G. Joao, “Marketing Research: The Role of Sentiment Analysis,” FEP Working Papers, April 2013
- [2] M. Walaa et al, “Sentiment Analysis Algorithms and Applications: A Survey,” in Ain Shams Engineering Journal , vol 5 Issue 4, Dec 2014, pp. 1093-1113
- [3] 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
- [4] Klipelainen, Pekka (2005). Lecture 3 : Boyer-Moore Matching [PDF Document]. Retrieved from [www.cs.uku.fi/~kilpelai/BSA05/lectures/slides03.pdf](http://www.cs.uku.fi/~kilpelai/BSA05/lectures/slides03.pdf)
- [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, 19 Mei 2017



Annisa Nurul Azhar, 13515129