Implementation of social media mining for decision making in product planning based on topic modeling and sentiment analysis

M I Irawan¹, R Wijayanto², M L Shahab³, N Hidayat⁴, A M Rukmi⁵
Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
E-mail: mii@its.ac.id

Abstract. The development of information and communication technology, especially on the social media field, brings a wide change in product planning in company. At this time, a company can identify the opportunity for product planning based on customers’ opinion through Twitter’s postings. This opportunity can be used to increase the competitiveness and defend from the competitors. The main method that is used in this research is social media mining based on topic modeling and sentiment analysis. The topic that is widely discussed by the customers is identified as the importance degree and the result of sentiment analysis is identified as the satisfaction degree. We successfully apply topic modeling with Latent Semantic Analysis and K-Means cluster and sentiment analysis with lexicon dictionary from Hutto and Gilbert. Using McDonald’s 2000 tweets, it can be concluded that ”service in the morning” has the highest opportunity degree and it is categorized at under-served thing. It gives us a meaning that ”service in the morning” is one of the service which has a great chance at the market.

1. Introduction
The customers’ opinion is the one of the important things at the product planning based on information retrieval that is necessary to pay more attention [1]. The company can use this data to increase its service and invocation to its released product [2]. Through the information and communication technology development, customers can express their opinions using the online social media, such as Twitter, Facebook, or Reddit easily [3]. One of the most popular social media that customers like to use is Twitter. Through the Twitter, customers can express their opinion using text, picture, emotion sign and soon by mention mode the company at posting tag. The post character limitation is a reason why the scientists choose Twitter as the object to do their research.

Social media mining is the one of the methods can be used in the product planning sourced from social media. The supporting method that also can be used is opportunity mining approach [4]. Text mining theories, topic modeling using Latent Semantic Analysis and K-Means, also sentiment analysis can be used gradually to analyze the customers’ opinion scraped from Twitter.

Topic modeling using Latent Semantic Analysis (LSA) and K-Means will group the opinions with similar topic at the same cluster. From the topic modeling result, can be generated an importance degree. In the other side, the clustered opinion, can be calculated the average of sentiment polarity to get the satisfaction degree. From the two ones, then can be calculated
the opportunity degree for each product part and can be visualized at product landscape map through the three groups, under-served, served-right, and over-served.

To show the application of this approach, it was implemented to social media data, that is Twitter using McDonald’s account.

2. Theoretical Background

2.1. Term Frequency Inverse Document Frequency (TF-IDF)

One of the weighting methods which can be used to transform the text or the collection of sentences as the vector is term frequency inverse document frequency or TF-IDF [5]. This method considers the appearance of features in a certain opinion and in a whole used opinion. Suppose, there are m opinions and n extracted features, the TF-IDF formula follows:

\[ tf_{i,j} = t_{i,j} \]  

(1)

where

- \( tf_{i,j} \) is the term frequency of \( j \)-th feature in the \( i \)-th opinion, \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \)
- \( t_{i,j} \) is the number of \( j \)-th feature in the \( i \)-th opinion, \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \).

Beside the IDF formula follows:

\[ idf_i = \log \left( \frac{m}{doc_j} \right) + 1 \]  

(2)

where

- \( idf_i \) is the inverse document frequency of feature \( j \)
- \( doc_j \) is the number of the opinion containing feature \( j \).

Based on (1) and (2) can be derived the TF-IDF of feature \( j \) at \( i \)-th opinion as follows:

\[ w_{i,j} = tf_{i,j} \times idf_i \]  

(3)

The normalized weight based on (3) to get the sparse matrix containing document term matrix as follows:

\[ W_{i,j} = \frac{w_{i,j}}{\sqrt{\sum_{j=1}^{n} w_{i,j}^2}} \]  

(4)

TF-IDF weight gives a high value for features that often appear in one tweet but rarely appear in other tweets, as well as low scores for high frequency features in a tweet but also often appear in other tweets.

2.2. Topic Modeling

Topic modeling is a method used for recognizing the implied topics in a text collection, including opinion collection. One of the methods that can be used is Latent Semantic Analysis or LSA [6]. Suppose given a sparse matrix \( A_{m \times n} \), LSA uses singular value decomposition theory to decompose \( A \) into \( S \Sigma V^T \). After that, can be chosen \( k \) principal components so that can be gotten a reduced sparse matrix or truncated SVD matrix. The purpose of truncated SVD is to keep the important features existence in an opinion collection and reduce the less important features or less dominant in a collection.

The step to do a reduction of a sparse matrix is choose \( k \) components, \( 1 \leq k \leq r \) where \( r = \min(m, n) \), then arrange \( S_k \times \Sigma_k \) matrix. \( S_k \) is the matrix containing reduced left singular vector, through keeping all the rows and \( k \) first column of matrix \( S \). Beside \( \Sigma_k \) is the diagonal
matrix with the singular values as the element through keeping $k$ first rows and $k$ first columns of matrix $\Sigma$. Taking $k$ values is based on threshold of sum explained variance ratio, that is minimum 0.3. The formula to calculate this score is as follows:

$$\sigma^2_{\text{trunc}_i} = \sigma^2_i(S_k \Sigma_k)$$

$$\sigma^2_{\text{sum}} = \sum_{j=1}^{m} \sigma^2_j(A)$$

where

- $\sigma^2_{\text{trunc}_i}$ is the variance of matrix $S_k \Sigma_k$, $1 \leq i \leq k$.
- $\sigma^2_{\text{sum}}$ is the sum of the variance of each row of sparse matrix, $1 \leq j \leq km$.

Then, the sum explained variance ratio for $k$ components can be calculated as follows:

$$\sigma^2_{\text{ratio}} = \frac{\sum_{i=1}^{k} \sigma^2_{\text{trunc}_i}}{\sigma^2_{\text{sum}}}$$

The matrix $S_k \Sigma_k$, then used as the input data on K-Means cluster algorithm [7].

2.3. Sentiment Analysis

Sentiment analysis used in this study is a lexicon-based sentiment analysis proposed by Hutto and Gilbert, namely VADER [8]. This sentiment analysis produces polarity values for each tweet based on word dictionary sentiment values determined by Hutto and Gilbert. In measuring the polarity of a tweet, tweets in the form of sentences will be tokenized to the token, then given a score based on the lexicon list. For example, given a tweet that will be analyzed $t$ and $n$ is a lot of words from $t$ tweets. The polarity value of tweets is calculated by:

$$P_{\text{total}}(t) = \sum_{i=1}^{n} P_{\text{word}}(i)$$

where

- $P_{\text{total}}(t)$ is the initial of tweet polarity $t$.
- $P_{\text{word}}(i)$ is the polarity of word $i$.

Hutto and Gilbert define some special conditions in the calculation of polarity as follows:

- If the word $i$ in lexicon is written in capital letters while not for other words, then the word is given emphasis so $P_{\text{total}}$ is added by 0.733 if the word is positive and reduced by 0.733 if the word is negative.
- If the word $i$ is preceded by the word booster then the word is added to the polarity of the word booster (if the word is negative then the polarity of the booster is multiplied by -1).
- If there is a word ”but”, the polarity of the sentence before the word ”but” will decrease the intensity by multiplying 0.5 while the polarity of the sentence after the ”but” is increased by multiplying the intensity by 1.5.
- If before the word $i$ there is the word ”least” then the polarity of the word $i$ is multiplied by -0.74.
- If the sentence contains a command sign (!):
  - If the number of command marks (!) is less than four:
If $P_{total}$ is positive, it is added by $P_{tanda1} = \text{number of signs} \times 0.292$

If $P_{total}$ is negative, it is added by $P_{tanda1} = -\text{number of signs} \times 0.292$

– If the number of command marks (!) is more than or equal to four:

  * If $P_{total}$ is positive, it is added by $P_{tanda1} = 4 \times 0.292$
  * If $P_{total}$ is negative, it is added by $P_{tanda1} = -4 \times 0.292$

• If the sentence contains a question mark (?)

  – If the number of question marks (?) is more than one and less than or equal to three:
    * If $P_{total}$ is positive, it is added by $P_{tanda1} = \text{number of signs} \times 0.18$
    * If $P_{total}$ is negative, it is added by $P_{tanda1} = -\text{number of signs} \times 0.18$

  – If the number of command marks (!) is one or more than three:
    * If $P_{total}$ is positive, it is added by $P_{tanda2} = 0.96$
    * If $P_{total}$ is negative, it is added by $P_{tanda2} = -0.96$

If it does not contain a command (!) and does not contain a question mark (?), or $P_{total} = 0$ then $P_{tanda1} = P_{tanda2} = 0$. So that the final polarity of opinion can be calculated with the following formula:

$$P_{\text{final}}(t) = P_{total}(t) + P_{tanda1} + P_{tanda2}$$ (9)

To get the opinion polarity value in the range [-1,1], then the polarity in (9) will be normalized with the following formula:

$$P_{\text{norm}}(t) = \frac{P_{\text{final}}(t)}{\sqrt{(P_{\text{final}}(t))^2 + \alpha}}$$ (10)

where $\alpha$ is the normalization parameter. The value of $\alpha$ used in this research is 20.

2.4. Opportunity Mining

To identify parts of the product to be developed, then the degree of importance and degree of satisfaction of the topic are calculated [4]. The degree of importance of the topic is calculated from the number of tweets in one cluster, while the degree of topic satisfaction is measured from the average sentiment polarity of tweets in one cluster. Each degree is then normalized in the interval (0, 10) with the following formulas:

$$\text{importance}_i = 10 \times \frac{CS_i - CS_{\text{min}}}{CS_{\text{max}} - CS_{\text{min}}}$$ (11)

$$\text{satisfaction}_i = 10 \times \frac{SS_i - SS_{\text{min}}}{SS_{\text{max}} - SS_{\text{min}}}$$ (12)

where

• $\text{importance}_i$ is the normalized importance degree topic $i$
• $CS_i$ is the initial of importance degree topic $i$
• $CS_{\text{min}}$ is the minimum value of initial importance degree
• $CS_{\text{max}}$ is the maximum value of initial importance degree
• $\text{satisfaction}_i$ is the normalized satisfaction degree topic $i$
• $SS_i$ is the initial of satisfaction degree topic $i$
• $SS_{\text{min}}$ is the minimum value of initial satisfaction degree
• $SS_{\text{max}}$ is the maximum value of initial satisfaction degree
and the final formula is

$$opportunity_i = importance_i + \max(importance_i - satisfaction_i, 0)$$

(13)

After getting all the components, each topic section will be mapped into the product landscape map diagram as shown in Figure 1. Topic labels are visualizations of the product parts delivered by the customer.

Figure 1. The Product Landscape Map Diagram

Figure 2. Block Diagram of the Proposed Method
3. Results and Discussion

The proposed method for this research is illustrated in a block diagram shown in Figure 2. The data used in this research are the tweet data scraped from Twitter with the keyword @McDonalds from 4-th to 7-th March 2019. There are 2000 tweets. After filtering, tokenizing, and other text mining operations, then we implement the TF-IDF to obtain a sparse matrix. The dimension of the sparse matrix is $2000 \times 7280$. We use $k = 200$ in the truncated SVD stage to obtain sparse matrix reduction. The results of the topic modeling and the selected labels are shown in Table 1.

Next, we implement the sentiment analysis. The result shows that 40.72% of the tweets are positive, 8.08% are neutral, and 51.20% are negative. In addition, using the corresponding confusion matrix, the result show that the accuracy of the system is 93%, the recall is 94%, and the precision is 88%. The result of opportunity mining obtained after normalizing at intervals [0.10] are shown in Table 2.

![Figure 3. The Product Landscape Map](image)

The visualization of the result of opportunity mining on the product landscape map is shown in Figure 3. "Service in the morning" is the one that is under-served. It also has the highest opportunity, that is 14.40. The under-served position in the morning service illustrates that an important part of McDonald’s that should be prioritized, but has a relatively low level of satisfaction compared to its interests in society. It indicates that customers are not satisfied with that. An opportunity value of 14.40 provides another interesting review. Ulwick explained that the product in a market share that has an opportunity value in the range of 12 and 15 will give good opportunity if it is planned well [4].

If you pay attention to the current conditions in the community, sometimes there are many restaurants that do not provide optimal service in the morning or are not even open yet. This is certainly one of the interesting tricks to obtain a better profit compared to other competitors.

4. Conclusion

Based on the results, we know that the of McDonald’s products that needed to be planned and developed to the maximum was the service "service in the morning". This conclusion is obtained from the topic modeling that is well implemented by utilizing LSA with truncated SVD and reduction to 200 principal components and clustering it to 42 clusters. The sentiment analysis is also used to obtain an average polarity value and the social media mining can be implemented properly.
Acknowledgement
This research was supported by the Department of Mathematics, Institut Teknologi Sepuluh Nopembe (ITS), Ministry of Research, Technology, and Higher Education (KEMENRISTEKDIKTI). In accordance with the funding agreement of Penelitian Dana Departemen, contract number: 1603/PKS/ITS/2019.

References
[1] Silverstein D, Samuel P and DeCarlo N 2013 The innovator’s toolkit: 50+ techniques for predictable and sustainable organic growth (John Wiley & Sons)
[2] Park H and Yoon J 2015 Service Business 9 115–135
[3] Brooks S 2015 Computers in Human Behavior 46 26–37
[4] Pinegar J S 2006 Journal of Product Innovation Management 23 464–466
[5] Manning C, Raghavan P and Schütze H 2010 Natural Language Engineering 16 100–103
[6] Landauer T K, McNamara D S, Dennis S and Kintsch W 2013 Handbook of latent semantic analysis (Psychology Press)
[7] Han J and Kamber M 2012
[8] Hutto C J and Gilbert E 2014 Eighth international AAAI conference on weblogs and social media
| Cluster | Number of Tweet | Keywords                                                                 | Label                                      |
|---------|----------------|--------------------------------------------------------------------------|--------------------------------------------|
| 1       | 70             | service, customer, customer service, manager, order                      | Manager and customer service               |
| 2       | 57             | coffee, iced, iced coffee, work, sugar                                   | Coffee drink                               |
| 3       | 49             | chicken, sandwich, nuggets, chicken nuggets, food                        | Chicken nugget                             |
| 4       | 51             | window, order, drive, thru, car                                          | The drive thru service for car             |
| 5       | 12             | food, drive, thru, drive thru, work                                      | The service at the drive thru              |
| 6       | 152            | mcdonalds, food, service, people, staff                                 | Consument service                          |
| 7       | 637            | food, service, breakfast, mcd, eat                                      | Service at the morning                     |
| 8       | 7              | drive, order, avoid, mcdonald, hard                                     | Delivery service                           |
| 9       | 156            | drive, thru, drive thru, order, mcdonald                                | Food order using drive thru                |
| 10      | 8              | mcchicken, spicy, ordered, spicy mcchiken, dollar                       | Spicy mcchicken                           |
| 11      | 5              | cruelty, supply, cruelty supply, chicken cruelty supply, chicken cruelty| Chicken supply                            |
| 12      | 16             | thru, drive thru, drive, order, drive thru order                         | Order system in drive thru service         |
| 13      | 129            | mcdonald, food, service, people, eat                                    | Food service                               |
| 14      | 15             | happy, meal, happy meal, toys, meal toys                                | Children ground                           |
| 15      | 14             | pounder, quarter, quarter pounder, cheese, double quarter               | Food portion                              |
| 16      | 11             | sandwiches, menu, chicken, sandwich, chicken sandwiches                 | Chicken sandwich                          |
| 17      | 9              | dirty, kids, hope, staff, table                                         | Playground                                |
| 18      | 19             | job, hate, friendly, bag, employees                                     | Employee friendliness                      |
| 19      | 16             | chickens, chickens deserve, deserve lives, lives living chicken         | Chicken supplier                          |
| 20      | 180            | order, time, food, times, mcdonald                                      | Food order duration                        |
| 21      | 4              | close, mcdonald, dirty, experience, eating ice cream, machine, cream, ice, ice cream machine | Place cleanliness |
| 22      | 13             | ice cream, machine, cream, ice, ice cream machine                       | Ice cream and ice cream machine            |
| 23      | 8              | dead, man, raccoon, waited, video                                       | Food hygiene                               |
| 24      | 20             | coke, diet, diet coke, drink, soda                                      | Soda drink                                |
| 25      | 17             | coffee, cup, manager, milk, job                                         | Coffee portion                            |
Continuation of Table 1

| Cluster | Number of Tweet | Keywords | Label                           |
|---------|-----------------|----------|---------------------------------|
| 26      | 20              | cheese, bacon, egg, egg cheese, biscuit, commitment, chicken commitment, chicken, store, expect | Bacon and cheese egg |
| 27      | 11              | commitment, chicken commitment, chicken, store, expect | Other |
| 28      | 18              | ice cream, ice, cream, cone, cream cone | Ice cream cone and ice cream portion |
| 29      | 23              | egg, mc m m m, sausage, egg mc m m m, sausage egg | Mc m m m |
| 30      | 15              | fish, filet, filet fish, cheese, fillet | Filet and fish |
| 31      | 7               | align actions, align, rhetoric, trust rhetoric, align, rhetoric align | Other |
| 32      | 9               | apple, pies, apple pies, pie, apple pie | Apple pie |
| 33      | 3               | burger, king, veggie, burger king, wanted | Vegetable burger |
| 34      | 86              | fries, cold, french, french fries, food | French fries |
| 35      | 19              | burger, king, cheese, burger king, bun | Cheese burger |
| 36      | 4               | dinner, messed, fries, messed order, friend | Dinner menu |
| 37      | 11              | wrap, chicken, missing, lettuce, hope minutes, waited, waiting, waited minutes, order | Packaging |
| 38      | 37              | | Service time |
| 39      | 21              | free, cheap, food, fries, buy | Price |
| 40      | 29              | sauce, asked, nuggets, sour, sour sauce | Sauce |
| 41      | 2               | window, coffee, hope, wait, mcdonalds | Worker discipline |
| 42      | 10              | order, mess, mess order, messed, staff | Order recording |
Table 2. The Result of Opportunity Mining

| Cluster | Importance | Satisfaction | Opportunity |
|---------|------------|--------------|-------------|
| 1       | 1.07       | 3.55         | 1.07        |
| 2       | 0.87       | 6.46         | 0.87        |
| 3       | 0.74       | 5.44         | 0.74        |
| 4       | 0.77       | 3.32         | 0.77        |
| 5       | 0.16       | 5.85         | 0.16        |
| 6       | 2.36       | 5.27         | 2.36        |
| 7       | 10         | 5.6          | 14.4        |
| 8       | 0.08       | 2.28         | 0.08        |
| 9       | 2.43       | 5.1          | 2.43        |
| 10      | 0.09       | 5.4          | 0.09        |
| 11      | 0.05       | 1.34         | 0.05        |
| 12      | 0.22       | 4.09         | 0.22        |
| 13      | 2          | 6.87         | 2           |
| 14      | 0.2        | 9.29         | 0.2         |
| 15      | 0.19       | 4.97         | 0.19        |
| 16      | 0.14       | 4.66         | 0.14        |
| 17      | 0.11       | 2.09         | 0.11        |
| 18      | 0.27       | 5.08         | 0.27        |
| 19      | 0.22       | 9.36         | 0.22        |
| 20      | 2.8        | 3.37         | 2.8         |
| 21      | 0.03       | 5.88         | 0.03        |
| 22      | 0.17       | 4.8          | 0.17        |
| 23      | 0.09       | 0            | 0.19        |
| 24      | 0.28       | 5.51         | 0.28        |
| 25      | 0.24       | 5.55         | 0.24        |
| 26      | 0.28       | 4.14         | 0.28        |
| 27      | 0.14       | 10           | 0.14        |
| 28      | 0.25       | 5.8          | 0.25        |
| 29      | 0.33       | 7.18         | 0.33        |
| 30      | 0.2        | 7.64         | 0.2         |
| 31      | 0.08       | 8.81         | 0.08        |
| 32      | 0.11       | 4.42         | 0.11        |
| 33      | 0.02       | 6.66         | 0.02        |
| 34      | 1.32       | 5.43         | 1.32        |
| 35      | 0.27       | 5.35         | 0.27        |
| 36      | 0.03       | 5.93         | 0.03        |
| 37      | 0.14       | 5.8          | 0.14        |
| 38      | 0.55       | 3.26         | 0.55        |
| 39      | 0.3        | 7.3          | 0.3         |
| 40      | 0.43       | 5.23         | 0.43        |
| 41      | 0          | 9.69         | 0           |
| 42      | 0.13       | 3.3          | 0.13        |