Comparative study between KNN and maximum entropy classification in sentiment analysis of menstrual cup

E Sugiharti¹*, D Fauziah²

¹,²Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia

*Corresponding author: endangsugiharti@mail.unnes.ac.id

Abstract. The rise of reuse, reduce, recycle movement impacts both in pollutants management and waste. Including menstrual waste by using the menstrual cup. However, the use of menstrual cups in Indonesia is not yet popular. Sentiment analysis is needed to see how the public thinks about the menstrual cup. Using 1108 data from Twitter, which was then labeled into positive and negative manually, the sentiment analysis stage was carried out using Maximum Entropy and K-Nearest Neighbor. These two were chosen because maximum entropy works by obtaining the best probability distribution, which closest to reality, and k-nearest neighbors work by classifying new objects based on attribute examples and training data. The implementation program uses PyCharm IDE and python programming language. From the research results, maximum entropy and k-nearest neighbors accuracy each are 84.6% and 83.7% with 1108 tweets. In the next research, a lexicon dictionary can be used to replace the manual labeling process.

1. Introduction

The rise of the movement to reuse, reduce, recycle impacts both in pollutants management and waste. Menstrual waste is no exception. The use of a menstrual cup, a menstrual sanitary device that does not absorb blood, but collects blood [1], allows users to reduce menstrual waste. The litter of this waste ranges up to 10,000 per woman in a lifetime [2]. However, the use of menstrual cups in Indonesia is not yet popular. Only 0.1% of women have heard of menstrual cups [3].

Sentiment analysis is needed to see how the public thinks about the menstrual cup. Sentiment analysis will classify opinions into very positive, positive, neutral, negative, and very negative [4]. Previous studies that examined the issue of analytical sentiment have been published by [5, 6, 7]. Sentiment classification based on machine learning can be formulated by problem-based learners [8]. The study of improving accuracy on sentiment analysis has been studied in [9, 10]. By using 1108 data from Twitter which then labeled into positive and negative manually, the sentiment analysis stage was carried out using Maximum Entropy and K-Nearest Neighbors. The maximum entropy is chosen because the method is able to find a distribution that produces the maximum entropy value. This is because the main principle of maximum entropy is that high entropy is directly proportional to uniformity [11]. Those values are used to obtain the best probability distribution which closest to reality. On the other hand, k-nearest neighbors works by classifying new objects based on attribute examples and training data [12]. Therefore, it does not make any assumptions about the distribution of the existing data, but rather the model is determined by the data.
Research on maximum entropy and k-nearest neighbor have been repeatedly carried out separately. Adiwijaya and Adiwijaya [13] found that by using 1000 positive documents and 1000 negative documents, the accuracy of k-nearest neighbors is 96.8%. These results were obtained by using the Information Gain selection feature. This is because k-nearest neighbor works by measuring the distance between features. The better the features and the optimal number of features, the higher the accuracy results get.

In another study, Putri, et.al [14] compared maximum entropy with support vector machine in sentiment analysis for software reviews. With as many as 4505 reviews consisting of 2736 positive reviews and 1769 negative reviews, obtained from 89.01% for support vector machine and 90.46% for maximum entropy.

Surohman [15] also made a comparison between the Naïve Bayes Classifier and k-nearest neighbor. The data used are reviews of the Dana application with 116 positive and negative reviews respectively. The accuracy produced by the naïve bayes classifier is 84.76% ± 3.93% with a micro average of 84.85%. Meanwhile, k-nearest neighbor produced a value of 82.92% ± 4.87% with a micro average of 82.96%.

Sentiment analysis with the object of the menstrual cup needs to be done to see what the public thinks about the menstrual cup. On the other hand, sentiment analysis by comparing the maximum entropy with the nearest k-neighbors has never been done before. So, this study can be a bridge of knowledge between sentiment analysis and menstrual health.

Based on those backgrounds, this study aims to see the public sentiment towards the menstrual cup by analyzing the sentiment using maximum entropy and k-nearest neighbors, and nothing accurate is produced by the two methods.

2. Methods
This study aims to determine public opinion on menstrual cups and how high the maximum entropy and k-nearest neighbors methods produce text analysis. The whole process of sentiment analysis will be written in Python with Pycharm IDE.

2.1. Data Collecting
Data Collecting is a step to get all of the data needed. Data collection in text mining is an important step because data is the initial foothold in research. It is known that this step takes the most time.

2.1.1. Twitter crawling. Twitter crawling is a step to get the data from Twitter database. This is done by requesting access to Twitter with registering Twitter API in return for API Key.

2.1.2. Manual labelling. Manual labeling is a process to label the documents we get from Twitter, crawling into positive and negative manually. Weighted Majority Voting is used in this process to decide what final label is, because this process needed at least 3 validators to validate the label. The equation for weighted majority voting can be seen in equation 1.

$$f(x) = \arg \max_i \sum_{j=1}^{m} w_j X A (C_j(x) = i)$$

(1)

2.2. Text Preprocessing
Text preprocessing is an activity to change data to be more structured. This is a process to organize data which consists of transform case, tokenization, stopword removal, and stemming. Those steps can be inverted on the data.

2.3. Split Validation
Split validation is a technique of dividing data into training data and testing data. The training data will be used to in the sentiment analysis model. Meanwhile, testing data will be used to make predictions. The comparison between training data and testing data is randomly selected. However, this process based on the Paretto Principle, which is an 80:20 ratio for training data and testing data.
2.4. Word Vectorization
This is a process to vectorize documents based on their weight. Term-Frequency-Inverse Document Frequency is used to be the method of this process. Term Frequency (TF) means the number of repetitions of words in the text/document while Inverse Document Frequency (IDF) is an algorithm that functions to calculate the inverse probability of finding words in the text [16]. The equation for Term-Frequency-Inverse Document Frequency can be seen in equation 2.

\[ tfidf = tf_{ij} \times \log \frac{D}{tf_{ij}} \]  

(2)

2.5. Feature Selection
This process aims to select best-weighted features amongst others. Information Gain method used in this process. Information Gain works by sorting the attributes from highest to lowest and reduce noise from irrelevant features. In other words, this method detects the most features with specific characteristics and classes [17]. Information gain is calculated by curating the total entropy, feature entropy, and Information Gain respectively, can be seen in equation (3), (4), (5) as follows.

\[ Entropy(S) = \sum_{i=1}^{k} (P_i) \log 2(P_i) \]  

(3)

\[ Entropy(S,A) = -\sum_{i=1}^{n} \left( \frac{|Sv_i|}{S} \times Entropy(Sv_i) \right) \]  

(4)

\[ Information Gain(S, A) = Entropy(S) - Entropy(S, A) \]  

(5)

2.6. Classification
This process classifies the document into positive and negative based on previous steps. The classification consists of 2 methods, Maximum Entropy and K-Nearest Neighbor.

2.6.1. Maximum entropy. Maximum entropy is a classification method that could find the maximum entropy value from the distribution \( p = (a|b) \) for getting the best probability distribution [14]. The equation for Maximum Entropy can be seen in equation 6.

\[ p = arg \max -\sum_{(a,b) \in (A,B)} p(a, b) \log(a, b) \]  

(6)

2.6.2. K-Nearest neighbor. This algorithm works by classifying objects based on their attributes and training data [3]. The equation for K-Nearest Neighbor can be seen in equation 7.

\[ Cos(a, b) = \frac{\sum b_{(a,b)} \sum b_{d}d_{f}}{\sqrt{\sum b_{(a,b)}^2} \sqrt{\sum b_{d}^2d_{f}^2}} \]  

(7)

3. Results and Discussion
This research is a accuracy comparison between Maximum Entropy and K-Nearest Neighbors in Sentiment Analysis. The used data is directly crawled from Twitter with menstrual cup as keyword. This research consists of data collection, data processing, feature selection, classification using the Maximum Entropy and K-Nearest Neighbors, and comparison of accuracy results.

3.1. Results
The first results we obtained from Twitter crawling process in *csv extension. We get 1108 tweets from April 26th 2020 up until May 25th 2020. The results from Twitter crawling can be seen in Table 1.
Move onto the manual labeling. This process was validated by 3 validator, a writer, both a co-writer and a lecturer, and an English teacher professional. In our experiment, we use weighted majority voting to determine what labels will be used in this study if there are differences in the labeling process. The results from manual labeling shown in Table 2.

| Text | Validator 1 | Validator 2 | Validator 3 | Final Label |
|------|-------------|-------------|-------------|-------------|
| i wouldn't have known the existence of a menstrual cup if dani hadn't left the band and open up abt mature talks onâ€¦ | positive | positive | positive | positive |
| soooo, I thought lá€™ve did give the menstrual cup a whirl. Saving the environment and all that good stuff! *honest &amp; detaiâ€¦ | positive | positive | positive | positive |
| I wish i could give every woman i love a menstrual cup. SUCH Aa game changer | positive | positive | positive | positive |
| δŸ’ėDid you know that period panties are an excellent partner for your menstrual cup? Especially in the early days, whâ€¦ | positive | positive | positive | positive |
| To all my tampon wearing ladies, I cannot stress this enough: buy a menstrual cup | positive | positive | positive | positive |

The final label will be used as a label column in the document for the next process, data processing. The first step is text preprocessing. This consists of 4 stages. First, transform the case. This is a step to transform all the text into the lowercase letter so it will be score equally for each letter. Second, tokenization. A step to divide the text into each token. Third, stopword removal. A step to remove stopword from the text so the text will be composed by main tokens. The last, stemming. A process to change token into its original word/word stem. The result of text preprocessing presented in Table 3.

| Text | Text Preprocessing |
|------|--------------------|
| i wouldn't have known the existence of a menstrual cup if dani hadn't left the band and open up abt mature talks onâ€¦ | “will” “know” “exist” “menstrual” “cup” “left” “band” “open” “mature” “talk” |
| soooo, I thought lá€™ve did give the menstrual cup a whirl. Saving the environment and all that good stuff! *honest &amp; detaiâ€¦ | “think” “give” “menstrual” “cup” “whirl” “save” “environment” “good” “stuff” “honest” |
| I wish i could give every woman i love a menstrual cup. SUCH Aa game changer | “wish” “can” “give” “woman” “love” “menstrual” “cup” “game” “change” |
| δŸ’ėDid you know that period panties are an excellent partner for your menstrual cup? Especially in the early days, whâ€¦ | “do” “know” “period” “panty” “excellent” “partner” “menstrual” “cup” “early” |
| To all my tampon wearing ladies, I cannot stress this enough: buy a menstrual cup | “tampon” “wear” “lady” “can” “stress” “enough” “buy” “menstrual” “cup” |
After text preprocessing is done, then word vectorization is next. Using Term Frequency-Inverse Document, this process starts by dividing the document into two main data, training data and testing data, with ratio 20% for testing data and 80% for training data. Afterward, the data is converted into numeric data from the previous string data. Term Frequency-Inverse Document works by vectorizing weighted features with maximum feature chosen known as max_features. In this case, 1000 has been picked as max_features. This process is closely related to feature selection. Information Gain is used as feature selection, and the authors select features based on the top k value. The top k value here is k = 900. The results from feature selection can be seen in Table 4.

### Table 4. Feature selection results

| Top 4 feature selection results | Bottom 4 feature selection results |
|--------------------------------|-----------------------------------|
| (0 to 840) 0.15842658687749753 (771 to 479) 0.04367586624694716 | (771 to 710) 0.22775355993191196 |
| (0 to 701) 0.29118361006169224 (771 to 442) 0.04467250786807992 | (771 to 710) 0.04367586624694716 |
| (0 to 550) 0.21012049441224545 (771 to 170) 0.20670702772092264 | (771 to 710) 0.20670702772092264 |

The last stage is classification. The data from feature selection then be classified using Maximum Entropy and K-Nearest Neighbor. Experiments were carried out with differences in the amount of C in Maximum Entropy and N in K-Nearest Neighbor until optimal results were obtained for each classifier. The results of these experiments can be seen in Table 5.

### Table 5. The experiments results

| Amount of C | Amount of N | Accuracy of Maximum Entropy (%) | Accuracy of K-Nearest Neighbor (%) |
|-------------|-------------|----------------------------------|------------------------------------|
| 1           | 3           | 82.8                             | 83.7                               |
| 1           | 2           | 82.8                             | 76.1                               |
| 1           | 1           | 82.8                             | 81.5                               |
| 2           | 3           | 82.8                             | 83.7                               |
| 3           | 3           | 84.6                             | 83.7                               |
| 4           | 3           | 84.2                             | 83.7                               |

The highest accuracy results are shown in dashed red-box Table 5 with 84.6% for Maximum Entropy with C=3 and 83.7% with N=3. On the contrary, the lowest accuracy results are 82.8% for Maximum Entropy with C=1 or 2 and 76.1% for K-Nearest Neighbor with N=2.

### 3.2. Discussion

The results obtained from Maximum Entropy algorithm are optimum with C=3. With 1 or 2 as C, the accuracy drops by 1.4% to 82.8% and when C hits 4, the accuracy drops by 0.2% to 84.2%. It is clearly stated that C influences the number of accuracy Maximum Entropy gets. C is the inverse of regularization power. The smaller the value, the stronger regularization gets. Meanwhile, the results obtained from the K-Nearest Neighbor algorithm are with the number of 3 in N. When it hits 1 and 2, the accuracy decrease. N is the number of neighbors use in this algorithm for k-neighbors queries. With N=3, then each document is using 3 other documents as a reference.

With these result level of accuracy, this model is expected to be able to analyze the sentiment data of the menstrual cup well and able to compare the results between Maximum Entropy and K-Nearest Neighbor. Apart from sentiment analysis towards menstrual cup is somewhat few, comparison between those 2 algorithms has never been done alone before.
4. Conclusion
The accuracy of Maximum Entropy and K-Nearest Neighbor are 84.6% and 83.7% in sentiment analysis using menstrual cup data. Those results were obtained by first crawling data from Twitter and labeled them manually. Next is preprocessing, word vectorization using Term Frequency-Inverse Document Frequency, feature selection using Information Gain, and last classification stage. Based on the results, Maximum Entropy algorithm produces higher results by 0.9% than K-Nearest Neighbor.

References
[1] L. Johansson and H. Hellstrom 2018 http://urn.kb.se/resolve?urn=urn:nbn:se:sh:diva-35612
[2] Stewart K, Powell M and Greer R 2009 J. of Obste. and Gynae.. 29 49
[3] Davis J, Macintyre A, Odagiri M, Suriastini W, Cordova A, Huggett C, Paul A. Agius, Faiqoh, Budiyani A E, Quillet C, Cronin A A, Diah N, Triwahyunto A, Luchters S 2018 Tropi. Medic. and Int Healt. 23 12
[4] Chen Y and Skiena S 2014 Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. 2 383
[5] Larasati U I, Muslim M A, Arifudin R, and Alamsyah 2019 Sci. J. Inform. 6 138
[6] Duong H and Nguyen-Thi T 2021 Comput. Soc. Netw. 8 1
[7] Fransiska S, Rianto R and Gufroni A I 2020 Sci. J. Inform. 7 203
[8] Desai M and Mehta M A 2016 Int. Conf. Comput. Commun. Autom. p 149
[9] Tiffani I E 2020 J. Soft Comput. Explor. 1 1
[10] Suslistiana, Muslim M A 2020 J. Soft Comput. Explor. 1 8
[11] Patel D, Saxena S and Verma T 2016 Int. J. of Innov. Resea. in Scien. Engin. and Tech.. 5 5
[12] Istia S S and Purnomo H D 2018 Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. p 84
[13] Daeli N O F and Adiwijaya A 2020 J. of Data Scien. and Its App. 3 1
[14] Putri B A D, Khasanah A U and Azzam A 2019 2019 Int. Conf. on Infor. and Comm. Tech. p 468
[15] Surohman, Aji S, Rousyati and Wati F F 2020 Evolusi: J. Sains dan Manaj. 8 93
[16] Hakim A A, Erwin A, Eng K I, Galinium M and Muljadi W 2014 6th Int. Conf. on Infor. Tech. and Elect. Engin. p 29
[17] Utama H S, Rosiyadi D, Aridarma D and Prakoso B S 2019 J. Pilar Nusa Mandiri 15 247