Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes

Achmad Bayhaqy
Master of Computer Science - Postgraduate Programs
STMIK Nusa Mandiri
Jakarta, Indonesia
achmadbayhaqy92@gmail.com

Sfenrianto Sfenrianto
Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management,
Bina Nusantara University, Jakarta, Indonesia 11480
sfenrianto@binus.edu

Kaman Nainggolan
Master of Computer Science - Postgraduate Programs
STMIK Nusa Mandiri
Jakarta, Indonesia
kaman@nusamandiri.ac.id

Emil R. Kaburuan
Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management,
Bina Nusantara University, Jakarta, Indonesia 11480
emil.kaburuan@binus.edu

***Abstract*— Data mining can be used for data analysis of social media users who visit E-Commerce. This study uses data mining techniques aimed at comparing the classification in sentiment analysis from the views of E-Commerce customers who have been written on Twitter. The data set is derived from tweets about E-Commerce in Tokopedia and Bukalapak. Text mining techniques, transform, tokenize, stem, classification, etc. are used to build classification and analysis of sentiment analysis. Rapidminer is also used to assist in making analysis sentiments for comparison by using three different classifications in the dataset with the Decision Tree, K-NN, and Naïve Bayes Classifier approaches to find the best accuracy. The highest result of this study is the Naïve Bayes approach with an accuracy of 77%, precision 88.50% and recall of 64%.**

Keywords*— *classification, e-commerce, data mining, rapid miner, sentiment analysis, twitter

I. INTRODUCTION

Today, most of the internet and social media used has visited online stores or e-commerce. Tokopedia and Bukalapak are e-commerce visited by social media users in Indonesia. One of the most frequently used social media in expressing opinions in transactions on Tokopedia and Bukalapak is Twitter. Twitter is a social media that allows users to send real-time messages [1].

Customer opinion through twitter can conclude which e-commerce online site has the best service. This is due to the many comments on Twitter and even the trending topic on Twitter about the services provided by an e-commerce. Thus it is necessary to analyze the views and sentiments of an e-commerce user.

On the other hand, data mining is a method of finding knowledge in the database to find useful knowledge from data [2]. This is the process of obtaining attractive designs and relationships and can be serviced in large volumes of data [2]. Thus data mining can be used as a sentiment analysis because it has a large amount of twits' data.

This study discusses how to do sentiment analysis of e-commerce customer opinion data on Twitter which is used to determine whether the opinion data entered positive or negative sentiments. The study used three different groups to extract users' thoughts or feelings through their tweets and group them into different categories. Then compare the results to find out the classifier that provides the best results in terms of recall ratio and accuracy using data mining.

II. STUDY OF LITERATURE

Evaluation of online site services has been carried out. Research on the sentiment of analysis from tweets on Twitter using Arabic to analyze the accuracy and predict correct sentiments [3] [4]. Research [5] has proposed an analysis sentiment from English tweets using rapidminer. The tool approach is used for business purpose sentiment analysis [6]. Other studies used Twitter data to do linguistic analysis and then build highly efficient classifiers [7] [8]. Thus, the use of social media twitter for Sentiment Analysis has been carried out.

Sentiment Analysis from Twitter must focus on classification problems [9] [10] [11]. Classification is the process of finding a model (or function) that can explain and distinguish a concept or class of data. Various classification approaches such as Naïve Bayes [12] [4] [13], Support Vector Machine (SVM) [12] [4] [13], and K-Nearest Neighbor (K-NN) [4] has been applied to find the best results [4]. As an example, analysis of sentiment classification from social media twitter has been conducsed by F. Laeeq and N. M. Tabrez [14]. In their research, they used three classifiers namely K-NN, Naïve Bayes and Decision Tree Classifier for sentiment classification and obtained results that showed the accuracy of K-NN, Naïve Bayes, and Decision Tree Classifiers were 77.50%, 80%, and respectively 78%.

Thus, mining can be used for the level of analysis sentiment. Data mining is the process of finding patterns or knowledge [15]. Patterns must be valid, useful, and understandable in a number of steps: pre-processing, data mining, and post-processing [15].

III. METODOLOGY

In analyzing sentiments, there are several stages that need to be done to get the best test results. The steps consist of Collection and Labelling Data, Pre-processing, and Cof Sentiment Analysis. Figure 1 shows the steps in the analysis of the proposed analysis sentiment classification.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

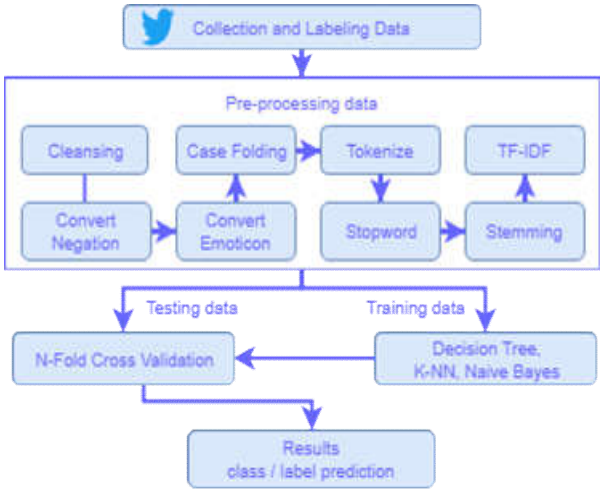


Fig. 1. Classification System Diagram for Analysis Sentiment

A. *Collection and Labeling Data*

In the first stage of carrying out the sentiment analysis process is data collection. The data was taken from Twitter with search queries about Tokopedia and Bukalapak as many as 50 records each using the rapidminer application.

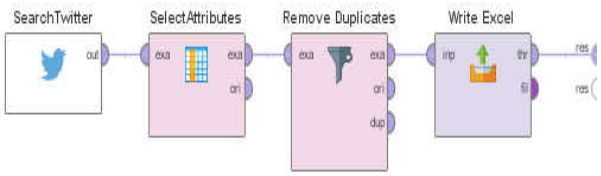


Fig. 2. Classification Data Retrieval

Figure 2 shows the retrieval of data from Twitter using the operator "search twitter" and storing data in an excel file using the operator "Write Excel" by only retrieving text from tweets and deleting all duplicate tweets with the operator "Remove Duplicates". The next stage is giving labeling. Labeling is done to divide the data into several classes of sentiments that will be used. The number of sentiments classes used is two classes, that is negative and positive. The purpose of this labeling process is to divide the dataset into 2 parts, to become data training and data testing. Data training is data that is used to train the system to be able to recognize the pattern being searched for, while data testing is data that is used to test the results of training that have been done. The following is one example of a dataset that has been labeled:

TABLE I. EXAMPLE OF THE LABEL SETTINGS

Text	Sentiment
@Bukalapak @Bukalapak tolong ditanggapi saya kecewa dengn bukalapak	Negative
Min sampe sekarang belum ada tindakan apapun dari admin Tokped ataupun penjual. Saya tidak mau mengulur2 waktu nih	Negative
btw, pas ngetik keywords 'sedotan' di tokopedia, hasil paling atas 'sedotan stainless'. wah, senang. sudah banyak yg makin peduli :)	Positive
Bukalapak Bikin Layanan Pinjaman Tanpa Syarat untuk UKM http://dlvr.it/Qb83n8 pic.twitter.com/8MucIMhUMO	Positive

B. *Pre-Processing*

After labeling the data, the next step is pre-processing. This stage is the stage where the data is prepared to become data that is ready to be analyzed. There are several stages in this preprocessing, including cleansing, convert negation, convert emoticons, case folding, tokenization, filtering stopword and stemming in Indonesian. Figure 3 shows the contents of the "Subprocess" operator. In this case, it is used for "Remove URL", "convert negation", "convert emoticons".

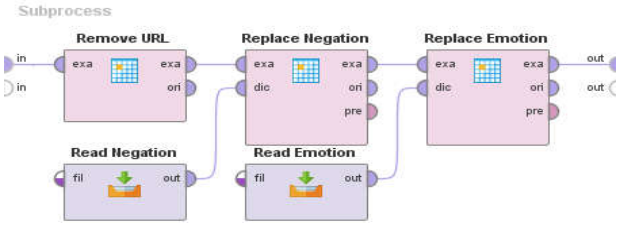


Fig. 3. Contents of the Subprocess Operator

Figure 4 shows the contents of the operator "Process Documents", used by operators "transform cases", "tokenize", "stopword filters" and "stem" in Indonesian.

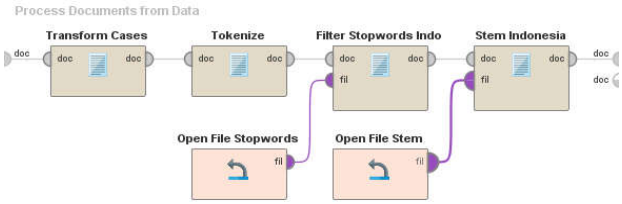


Fig. 4. Fill in the Process Documents Operator

The following is a detailed description of the pre-processing stages above:

- Cleansing

  Cleansing is a stage where characters and punctuation that are not needed are removed from the text [28]. It works to reduce noise in the dataset. Examples of characters that are omitted such as URLs, tags (#), punctuation such as dots (.), Commas (,) and other punctuation marks. This is an example of data cleansing sentences, input "Bukalapak Bikin Layanan Pinjaman Tanpa Syarat untuk UKM http://dlvr.it/Qb83n8pic.twitter.com/8MucIMhUMO, Then the output "Bukalapak Bikin Layanan Pinjaman Tanpa Syarat untuk UKM".

- Convert Negation

  In Indonesian, there are words "no", "no", "no", "less", "no" which are called the word negation which is a word that can reverse the meaning of the actual word [22]. This is an example of sentence convert negation, input "Saya tidak mau mengulur2 waktu nih", output "Saya tidak_mau mengulur2 waktu nih".

- Convert Emoticon

  Emotion is a facial expression represented by a combination of letters, punctuation, and numbers. Users usually use emoticons to express the mood they

are feeling. Convert Emoticons is one way to express expressions of feelings textually. This is an example of sentence convert emoticon, input "wah, senang. sudah banyak yg makin peduli :)", Then the output "wah, senang. sudah banyak yg makin peduli senang". Here is one example of the word convert emoticon:

TABLE II. EXAMPLE CONVERT EMOTION

Before	After
:(:'(:[;(:/): x(--" :# :-@ :c :f ;(:v :x :s)": *_*	sedih
:) :] (^_^) ^^v <3 ^^ ^_^ 0:) <:} :* (^.^) =) :3	senang

- Case Folding

  In the writing a tweet, there must be a different form of letters, this stage is a process of uniformity of letters, both of them are lower cases or upper cases. This is an example of case folding sentence, input "Bukalapak Bikin Layanan Pinjaman Tanpa Syarat untuk UKM", and output "bukalapak bikin layanan pinjaman tanpa syarat untuk ukm".

- Tokenization

  A process carried out to cut or break sentences into parts or words. The result of this deduction is called a token. In some cases, the tokenization process is also carried out by removing punctuation that is not needed. There are several tokenization models that can be used, namely unigram, bigram, trigram, and n-gram. This is an example of tokenization sentences, input "bukalapak bikin layanan pinjaman tanpa syarat untuk ukm", output "bukalapak, bikin, layanan, pinjaman, tanpa, syarat, untuk, ukm".

- Filtering

  Filtering is the stage of eliminating words that appear in large numbers but is considered to have no meaning (stopwords). Basically, the stopwords list is a set of words that are widely used in various languages. The reason for deleting stop words in many application programs related to text mining is because its usage is too general, so users can focus on other words that are far more important. This is an example of sentence stopwords, input "bukalapak bikin layanan pinjaman tanpa syarat untuk ukm", output "bukalapak bikin layanan pinjaman tanpa syarat ukm". Here is one example of a word from Stopwords:

TABLE III. EXAMPLE OF STOPWORDS

ada	di	kalau	pada	yaitu
aku	dia	kami	saja	bila
bapak	ini	lalu	tentu	hari
berbagai	itu	lewat	untuk	masa
cara	jadi	meski	yang	tapi
cuma	juga	oleh	wah	hal

- Stemming

  Stemming is a stage to make the word suffixes into basic word according to the correct Indonesian rules. This is an example of sentence stemming, input "bukalapak bikin layanan pinjaman tanpa syarat untuk ukm", and the output "bukalapak bikin layan pinjam syarat ukm". The following is one example of the word from stemming:

TABLE IV. EXAMPLE OF WORD STEMMING

Before	After
adanya	ada
akhiri	akhir
sebelum	belum
diberikan	beri
secukupnya	cukup
dipergunakan	guna

- Weighting Word

  Weighting Word is a mechanism to give a score on the frequency of occurrence of a word in a text document. One of popular method for weighting words is TF-IDF (Term Frequency-Inverse Document Frequency). Term Frequency-Inverse Document Frequency is a weighting method that combines two concepts, namely Term Frequency and Document Frequency. Term Frequency is the concept of weighting by looking for how often (frequency) the appearance of a term in one document. Because each document has a different length, it can happen that a word appears more in a long document compared to short documents. Thus, term frequency is often divided by the length of the document (the total words in the document).

  While Document Frequency is the number of documents where a term appears. The smaller the frequency of occurrence, the smaller the weight value. When calculating the term frequency, all the words in it are considered as important. However, there are words that are actually less important and need not be taken into account such as "di-", "ke-", "dan", etc. Therefore, these less important words need to be reduced in weight and add weight to other important words. This is the basic idea of why stopword is needed. Thus it is needed to calculate TF-IDF, that scores can be obtained using Equation.

$$tf - idf_{t,d} = tf_{t,d} * idf_t \quad (1)$$

  Term frequency (tf) is the frequency of occurrence of term (t) in the document (d).

C. *Classification of Sentiment Analysis*

After pre-processing the data, the next step is the classification of sentiment analysis. This stage is the stage to provide training and implement various data mining algorithms. Figure 5 shows the contents of the "Cross Validation" operator in the rapidminer application. In this case, using three different classification operators for

comparison, the classification operator "Decision Tree", the classification operator "K-NN" and the classification operator "Naïve Bayes".

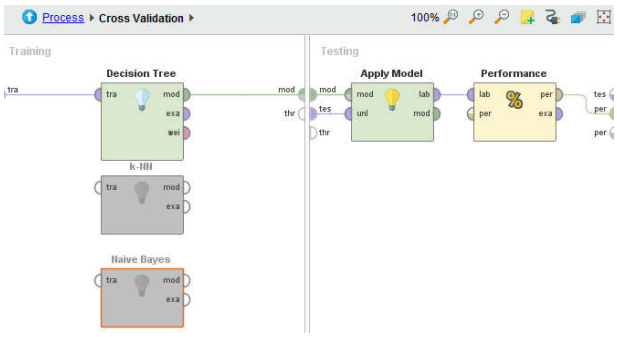


Fig. 5. Contents of the Cross Validation Operator

More detailed understanding for comparison the classification of Decision Tree, K-NN and Naïve Bayes are as follows:

- Decision Tree

  A decision tree is a hierarchical model for supervised learning where local regions are identified as a series of recursive separations through decision nodes in the test function. A decision tree is a method that is quite efficient in creating classifiers from data. Descriptions of Decision trees are most widely used with logic methods. The Decision Tree is a flowchart structure that resembles a tree, where each internal node (not leaf node) tests an attribute, each branch represents the results of the test, and each leaf node (or terminal node) states the class label. While the node at the top of the decision tree is the root node. Here's the data equation in tuple *D*.

$$Info\ (D) = \sum_{i=1}^{n} -p_i\ log_2(p_i) \quad (2)$$

  p_i is the probability of tuple in *D* which becomes the class *C_i* assuming |*C_(i,D)*|/|*D*|. Info (*D*) or also called entropy of D is the average information needed for identification of tuples in *D* [16].

- K-NN

  K-Nearest Neighbor (K-NN) algorithm is a method for classifying objects based on learning data that are closest to the object [16]. Therefore, to make predictions with K-NN, we need to define a metric to measure the distance between the query point and the case from the example sample. One of the most popular choices for measuring this distance is known as Euclidean.

$$D\ (x, p) = \sqrt{(x-p)^2} \quad (3)$$

  Where *x* and *p* are query points and examples of example cases, respectively.

  Because the K-NN prediction is based on the intuitive assumption that objects that are close in distance are potentially the same, it makes sense to distinguish between the closest neighbors *K* when making predictions. Let the closest point between *K*'s closest neighbors have more noise in influencing the results from the point of demand. This can be achieved by introducing a set of weights *W*, one for each of the closest neighbors, which is determined by the relative proximity of each neighbor by paying attention to the point of demand.

$$W(x, p_i) = \frac{\exp(-D(x, p_i))}{\sum_{i=1}^{K} \exp(-D(x, p_i))} \quad (4)$$

  Where *D* (x, pi) is the distance between the *x* query point and the case with the example sample p_i. The weights defined in this way will be satisfying:

$$\sum_{i=1}^{k} W(x_0, x_i) = 1 \quad (5)$$

  So, for classification problems, the maximum *y* is taken for each class variable [5].

$$\max(y = \sum_{i=1}^{K} W(x_0, x_i)(y)) \quad (6)$$

- Naïve Bayes Classifier

  Naïve Bayes is a machine learning that uses probability calculations that use the concept of a Bayesian approach. The use of Bayes theorem in the Naïve Bayes algorithm is by combining prior probability and conditional probability in a formula that can be used to calculate the probability of each possible classification [5].

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad (7)$$

D. *Evaluation of Sentiment Analysis*

After the sentiment analysis classification process is complete, one more step is needed to determine the quality of the process that has been carried out, namely evaluating the results. At this stage, the performance of the calculations that have been carried out will be tested with parameters of accuracy, precision, and recall.

Accuracy (*A*) is the number of documents classified correctly, both True Positive and True Negative. Calculating the accuracy value can use the equation:

$$A = \frac{(TP + TN)}{(TP + FP + TN + FN)} x\ 100\% \quad (8)$$

Precision (*P*) is how much processing results are relevant to the information you want to search. In other words, precision is a classification of True Positive and all data predicted as positive classes. Calculating precision values can use the equation:

$$P = \frac{TP}{(TP + FP)} x\ 100\% \quad (9)$$

While Recall (*R*) is how many relevant documents in the collection are generated by the system. In other words, recall is the number of documents that have the True Positive classification of all documents that are really positive (including False Negative). Calculating recall values can use the equation:

$$R = \frac{TP}{(TP + FN)} x\ 100\% \quad (10)$$

Variables such as *TP*, *TN*, *FP,* and *FN* come from the confusion matrix. *TN* stands for True Negative, negative data classified as negative. *TP* stands for True Positive, positive

data classified as positive. *FN* stands for False Negative, positive data classified as negative. *FP* stands for False Positive, negative data classified as positive. For a more detailed explanation:

TABLE V. CONFUSION MATRIX

	Prediction Yes	**Prediction No**
True Yes	*TP*	*FN*
True No	*FP*	*TN*

After the data is collected, the data will be divided into data training and data testing. The data division will be done using the N-fold cross validation method to eliminate word bias. N-fold cross validation divides documents into *n* sections. In a set of experiments will be carried out n pieces of document classification experiments with each experiment using one part as data testing, $(n-1)/2$ parts as labeled documents, and $(n-1)/2$ other parts as unlabeled documents that will be exchanged every experiment as many as n times. A collection of documents that are owned first are randomly sorted out before being inserted into a fold. This is done to avoid grouping documents from one particular category on a fold.

IV. RESULTS AND PERFORMANCE ANALYSIS

This section describes the results of the experimental results and analyzes their performance.

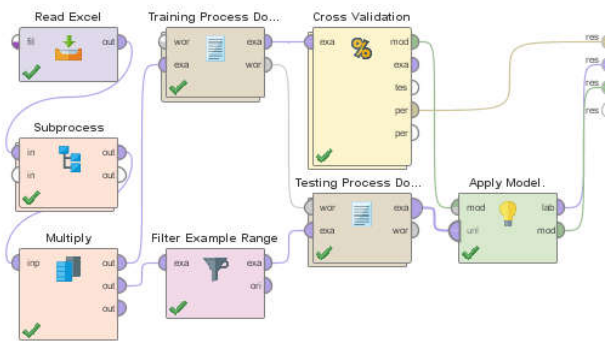
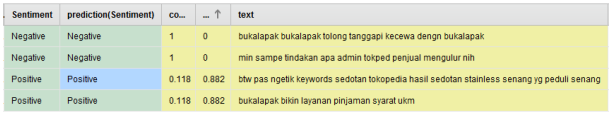


Fig. 6. Main Process on Rapidminer

Figure 6 shows the main process in the rapidminer application. The "Read Excel" operator is used to read data in the Excel file. "Subprocess" operators and "Process Documents" operators are used for pre-processing. "Cross Validation" operators are used for classification and evaluation of sentiment analysis with experiments conducted ten times (10-fold cross validation).



Sentiment	prediction(Sentiment)	co...	...	text
Negative	Negative	1	0	bukalapak bukalapak tolong tanggapi kecewa dengn bukalapak
Negative	Negative	1	0	min sampe tindakan apa admin tokped penjual mengulur nih
Positive	Positive	0.118	0.882	btw pas ngetik keywords sedotan tokopedia hasil sedotan stainless senang yg peduli senang
Positive	Positive	0.118	0.882	bukalapak bikin layanan pinjaman syarat ukm

Fig. 7. Prediction results from Rapidminer

Figure 7 shows the results of the prediction in the rapidminer application. Looks at the sentiment and also prediction labels from the same as sentiment label.

The following described the results of each algorithm's confusion matrix from rapid miner:

TABLE VI. CONFUSION MATRIX EACH ALGORITHM

Metode	***TP***	***FP***	***TN***	***FN***
Decision Tree	42	12	38	8
K-NN	35	7	43	15
Naïve Bayes	32	5	45	18

From the configuration matrix in table VI, the average value of accuracy, precision and recall are shown in table VII with the calculation using formulas (8), (9), (10).

TABLE VII. VALUE OF ACCURACY, PRECISION AND RECALL AVERAGE USING FORMULA

Metode	**Accuracy**	**Precision**	**Recall**
Decision Tree	80%	78%	84%
K-NN	78%	83%	70%
Naïve Bayes	77%	86%	64%

There are differences in the results of the average value of precision using the rapidminer application as shown in the table below:

TABLE VIII. VALUE OF ACCURACY, PRECISION AND RECALL AVERAGE USING RAPIDMINER

Metode	**Accuracy**	**Precision**	**Recall**
Decision Tree	80 %	79.96 %	84 %
K-NN	78 %	85.67 %	70 %
Naïve Bayes	77 %	88.50 %	64 %

These results show the Accuracy of Decision Tree, K-NN, and Naïve Bayes by 80%, 78%, and 77% respectively. Results for Precision of Decision Tree, K-NN, and Naïve Bayes amounted to 79.96%, 85.67%, and 88.50% respectively. While the results for Recall from the Decision Tree, K-NN, and Naïve Bayes were 84%, 70%, and 64% respectively. So it can be seen that the Naïve Bayes classifier is the best classifier for use with social media datasets because it provides more accurate and precise predictions. There are differences in the results of the previous research, showing the accuracy of K-NN, Naïve Bayes, and Decision Tree Classifiers of 84.66%, 50.72%, and 64.42% [6]. The difference is due to the characteristics of different datasets and processes.

CONCLUSION

In this study, an attempt was made to classify the analysis sentiments from tweets on Twitter on various customer opinions about several online marketplace sites in Indonesia. To summarize customer views about online marketplace in Indonesia the text mining techniques is used, and data mining using three different classifiers namely Decision Tree, K-NN, and NaïveBayes. The three classifiers predict the labels in the dataset. The results show the Accuracy of

the Decision Tree, K-NN, and Naïve Bayes of 80%, 78%, and 77%. Results for Precision of Decision Tree, K-NN, and Naïve Bayes amounted to 79.96%, 85.67%, and 88.50%. The results also show that Recall from Decision Tree, K-NN, and Naïve Bayes is 84%, 70%, and 64%. So it can be concluded that the Naïve Bayes classifier is the best classifier for use with social media datasets because it provides more accurate and precise predictions.

In the future, we should use a larger and more complex dataset with an increased number of labels and more e-commerce reach and can include non-standard Indonesian.

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," Assoc. Comput. Linguist., pp. 30–38, 2011 [Online]. Available: http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf

[2] Priyadharsini.C and D. A. S. Thanamani, "An Overview of Knowledge Discovery Databaseand Data mining Techniques," Int. J. Innov. Res. Comput. Commun. Eng., vol. 2, no. 1, pp. 1571–1578, 2014 [Online]. Available: http://www.rroij.com/open-access/an-overview-of-knowledge-discovery-databaseand-data-mining-techniques.pdf

[3] S. Al-Osaimi and M. Badruddin, "Sentiment Analysis Challenges of Informal Arabic Language," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 2, pp. 278–284, 2017 [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080237

[4] M. A. Ibrahim and N. Salim, "Sentiment Analysis of Arabic Tweets: With Special Reference Restaurant Tweets," Int. J. Comput. Sci. Trends Technol., vol. 4, no. 3, pp. 173–179, 2013 [Online]. Available: http://www.ijcstjournal.org/volume-4/issue-3/IJCST-V4I3P28.pdf

[5] P. Tripathi, S. K. Vishwakarma, and A. Lala, "Sentiment Analysis of English Tweets Using Rapid Miner," in 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 668–672 [Online]. Available: https://doi.org/10.1109/CICN.2015.137

[6] M. Shoeb, J. A.- work, and undefined 2017, "Sentiment Analysis and Classification of Tweets Using Data Mining," Irjet.Net, pp. 4–7, 2017 [Online]. Available: https://irjet.net/archives/V4/i12/IRJET-V4I12267.pdf

[7] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Proc. Seventh Conf. Int. Lang. Resour. Eval., pp. 1320–1326, 2010 [Online]. Available: http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf

[8] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Found. Trends® InformatioPang, B., Lee, L. (2006). Opin. Min. Sentim. Anal. Found. Trends® Inf. Retrieval, 1(2), 91–231. doi10.1561/1500000001n Retr., vol. 1, no. 2, pp. 91–231, 2006 [Online]. Available: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[9] M. Singh, "Sentiment Analysis and Similarity Evaluation for Heterogeneous-Domain Product Reviews," Int. J. Comput. Appl., vol. 144, no. 2, pp. 16–19, Jun. 2016 [Online]. Available: http://www.ijcaonline.org/archives/volume144/number2/singh-2016-ijca-910112.pdf

[10] A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6332 LNAI, 2010, pp. 1–15 [Online]. Available: http://link.springer.com/10.1007/978-3-642-16184-1_1

[11] J. Isabella, "Analysis and evaluation of Feature selectors in opinion mining," Indian J. Comput. Sci. Eng., vol. 3, no. 6, pp. 757–762, 2013 [Online]. Available: http://www.ijcse.com/docs/INDJCSE12-03-06-033.pdf

[12] T. O'Keefe and I. Koprinska, "Feature selection and weighting methods in sentiment analysis," Australas. Doc. Comput. Symp., pp. 67–74, 2009 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.5694&rep=rep1&type=pdf#page=77

[13] K. Bhuvaneswari, R. Parimala, and A. Professor, "Sentiment Reviews Classification using Hybrid Feature Selection," Int. J. Database Theory Appl., vol. 10, no. 7, pp. 1–12, Jul. 2017 [Online]. Available: http://dx.doi.org/10.14257/ijdta.2017.10.7.01

[14] F. Laeeq and N. M. Tabrez, "Sentimental Classification of Social Media using Data Mining," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, p. pp 546-549, 2017 [Online]. Available: https://doi.org/10.26483/ijarcs.v8i5.3360

[15] B. Liu, Web Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011 [Online]. Available: http://link.springer.com/10.1007/978-3-642-19460-3

[16] S. Wahyuningsih, D. R. Utari, U. B. Luhur, D. Tree, and K. Validation, "Perbandingan Metode K-Nearest Neighbor , Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit," Konf. Nas. Sist. Inf. 2018, pp. 8–9, 2018 [Online]. Available: http://jurnal.atmaluhur.ac.id/index.php/knsi2018/article/view/424/349