Sentiment Analysis of User Preferences on Learning Management System (LMS) Platform Data

Boldson Herdianto Situmorang and Andi Chairunnas
Department of Computer Science
Faculty of Mathematics and Natural Sciences
Universitas Pakuan
Jl. Pakuan PO Box 452, Bogor 16143, West Java, Indonesia
boldson.situmorang@unpak.ac.id; andichairunnas@unpak.ac.id

Abdul Talib Bon
Department of Production and Operations, University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

The Covid-19 has resulted in universities shut all across the world. Globally, over 100 million college students are out of classroom. As a result, education has changed dramatically, with the distinctive rise of Learning Management System (LMS), whereby teaching is undertaken remotely and on digital platforms. As online learning has gone mainstream, it has never been more important to choose an educational LMS tailored to institution's mission and goals. Sentiment analysis tells user whether the information about the product is satisfactory or not before they choose it. Sentiment analysis can classify the polarity of the text in sentences or documents to see whether the opinion on the sentences or documents are positive or negative. This research focuses mainly on sentiment analysis of user preferences on Google Classroom data using Improved K-Nearest Neighbor method which is helpful to classify opinions in text into positive or negative categories. From the results obtained, testing with different input values of k results in different accuracy percentage values, for k = 2 of 74.00%, for k = 5 of 79.00%, for k = 10 of 83.00%, for k = 15 of 83.00%, for k = 20 of 84.00%, with the highest accuracy with a value of k = 20. Based on the test results, it can be taken conclusion that the k value has an effect on the accuracy of the classification system.

Keywords
Sentiment Analysis, Improved K-Nearest Neighbor, Learning Management System.

1. Introduction

The Covid-19 has resulted in universities shut all across the world. Globally, over 100 million college students are out of classroom. As a result, education has changed dramatically, with the distinctive rise of Learning Management System (LMS), whereby teaching is undertaken remotely and on digital platforms. As online learning has gone mainstream, it has never been more important to choose an educational LMS tailored to institution's mission and goals.

Customer satisfaction is very important for public service providers, customer satisfaction can be delivered with a survey application or writing criticism that can be used to evaluate and improve service (Muktafin et al. 2020). Though it belongs to the artificial intelligence domain, namely is considered a machine learning technique, sentiment analysis is defined as a text data analysis that provides a deep analysis about opinions, sentiments, appraisals, attitudes, and emotions and it allows us checking how many negative and positive keywords are included in a text message (Păvăloaia et al. 2019). Sentiment analysis tells user whether the information about the product is satisfactory or not before they choose it.

Previous researches have been done till now in the context of sentiment analysis to detect the polarity of the sentiment using text mining techniques classification and clustering. Sentiment analysis of movie to find out positive or negative user sentiment using improved k-nearest neighbor achieved an average accuracy of 96% (Arora et al.

2016). Sentiment analysis on the perceptions learning management system of students and faculty through the used 10 stratified cross validation with a support vector algorithm that is applied in machine learning tools achieved an average accuracy of 91,8182% (Baesa et al. 2019). Sentiment analysis for identifying the business intelligence analysis in GO-JEK using some classifier algorithms such as decision tree, naïve bayes, support vector machine, and neural network, the system shows that the decision tree provides the best performance.

Text classification is one of the problems of text mining. Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics (Sebastiani 2002). In this paper, text classification approach is adopted to characterize the polarity of an opinion in a LMS, namely Google Classroom (GC). This research focuses mainly on sentiment analysis of user preferences on GC data with language of Indonesian which is helpful to classify opinions in text into positive or negative categories. The working principle of framework for text classification based on using improved k-nearest neighbor (ImpkNN) and term frequency-inverse document frequency (TF-IDF) method will be explained in this paper.

2. Method

An outline of the steps and techniques followed for sentiment analysis are depicted in Figure 1.

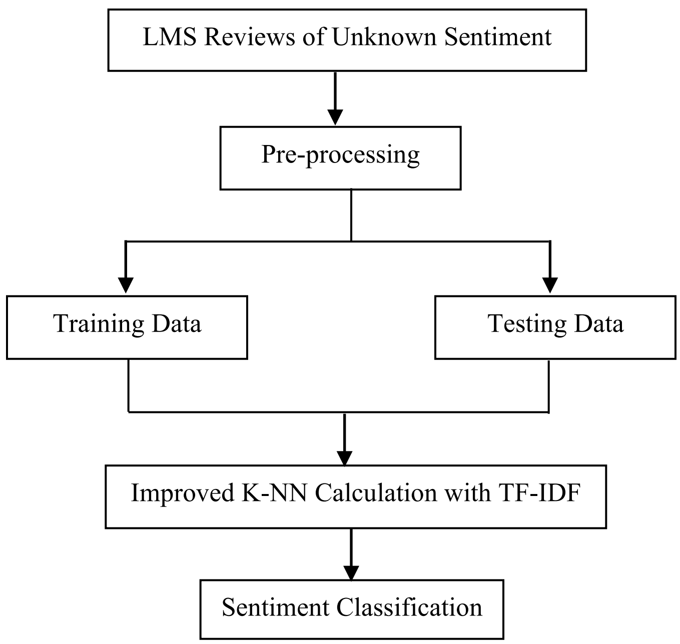

Figure 1: Proposed Framework for Sentiment Analysis

2.1. LMS Reviews

Sentiment classification aims at assigning a category to a document from a predefined set of categories. The predefined category set usually consists of some sentiment classes, e.g. positive or negative, which is the key difference from the topic-based text classification. LMS reviews about GC are obtained through the Google Play Store. Reviews were obtained using mobile application data scrapping technique, which is the process of extracting data from mobile applications. Data reviews from the scrapping process are then stored in CSV format.

2.2. Pre-Processing

The learning algorithms can not deal directly with the texts. That is why a preliminary stage, also called pre-processing, is required. First of all, text information should be processed before starting the classification. Pre-processing transforms documents into a suitable representation ready to go through the classification phase (Kechaou et al. 2010). Text pre-processing is defined as the task of cleaning the unwanted data from the hole and preparing the relevant data for classification. Assorted pre-processing steps are (Prananda and Thalib 2020):

a. Case Folding
   Case folding is the process to change the letter of each word to lowercase.
b. Stop Words Removal
   Stop words removal is the process of removing words that are too general and less important, the characteristic of this word is the frequency of occurrence which is quite a lot compared to other words.
c. Punctuation Removal
   Cleaning data by removing punctuation marks or non-text characters. Special symbols and numbers are not pertinent while observing sentiments. Although punctuation can provide grammatical context which supports understanding, yet it is irrelevant in determining sentiments (Arora et al. 2016).
d. Stemming
   Stemming is the process of changing the words in each sentence to their basic form or removing affixwords.

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

TF-IDF is a numerical statistic that shows the relevance of keywords to some specific documents or it can be said that, it provides those keywords, using which some specific documents can be identified or categorized (Qaiser and Ali 2018). Term frequency (TF) is a method for finding the weight of a document by looking for the number of occurrences of terms in the document. The more often the term appears, it will affect the amount of weight and the suitability value of the document. Inverse document frequency (IDF) is a method for calculating the distribution of terms in a document (Riadi et al. 2019). The TF-IDF method allows documents to be classified into two classes (positive and negative) (Melita, et a. 2018). The formula for calculating the TF-IDF can be written as follows:

$$W_{x,y} = tf_{x,y}\, x \log\left(\frac{N}{df_x}\right) \tag{1}$$

$W_{x,y}$ is the weight of the term (t_y) of document (d_x). While tf_{xy} is the number of ocurrences of term (t_y) in document (d_x). N is the number of documents in the database and df_x is the number of documents containing the term (t_y), there is at least one word, term (t_y).

2.4. Improved K-Nearest Neighbor (ImpkNN)

ImpkNN uses a weight value for each new training sample, which indicates the degree of the importance when classifying the documents. This method improves both the efficiency and the accuracy of text categorization (Joseph and Ramakrishnan 2015). The classification results are strongly influenced by the selected k value. Determination of the correct k value required in order to obtain high accuracy in the process of categorizing test documents. The k value used for each category in K-Nearest Neighbor is the same, it is less effective because each category can have a number of training data different. ImpkNN method is used to solve the problem where this method is used different k value for each category adjusted to the amount of training data owned. The new k value in each category is calculated using the following formula:

$$n = \left\lfloor \frac{k.N(Cm)}{maks\{(Cm)|j = 1 \dots Nc\}} \right\rfloor \tag{2}$$

n is the new k value, k is the preliminary k value, (Cm) is the number of training data of m category, maks {(Cm)|j=1…Nc} is the maximum number of training data in all categories.

3. Result and Discussion

In order to classify the document, it uses the pre-processing technique. Table 1 and Table 2 show the example of opinion documents in Indonesian before and after pre-processing.

Table 1. Sample positive and negative documents of opinion before pre-processing

LMS Review Documents (D) Before Pre-Processing		
Code	Opinion	Class
Training Data		

D1	Bagus	Positive
D2	Sangat membantu dalam proses perkuliahan saya, saya dapat melihat tugas & juga submit tugas hanya melalui aplikasi ini.	Positive
D3	Saya kasih bintang 1 aja karena saya tidak bisa login	Negative
D4	eror mulu !	Negative
D5	Tidak bisa buka berkas yg dikirim	Negative
Test Data		
D6	sistem error ketika saya hendak mengirim tugas sistemnya tidak mengizinkan file saya untuk terkirim	?

Table 2. LMS review documents after pre-processing

LMS Review Documents (D) Before Pre-Processing		
Code	Opinion	Class
Training Data		
D1	bagus	Positive
D2	bantu proses kuliah tugas submit tugas aplikasi	Positive
D3	kasih bintang aja login	Negative
D4	eror mulu	Negative
D5	buka berkas yg kirim	Negative
Test Data		
D6	error kirim tugas sistemnya izin file kirim	?

3.1. ImpkNN Calculation Using TF-IDF

This study uses the ImpkNN algorithm with k=3 and the term weighting uses TF-IDF. The TF-IDF calculation is applied to 900 training data and 100 test data after pre-processing. The next step is calculating the number of matches labeling results from the prediction of 100 testing data compared with manual labeling using confusion matrix. The confusion matrix divides the results of the predictions into four categories, they are: 1) true positive (TP) shows the number of data with positive classes and true predictive results, 2) true negative (TN) shows the number of data with negative classes and true predictive results, 3) false positive (FP) shows the number of data with positive classes and false, and 4) false negative (FN) shows the number of data with negative classes and wrong prediction results (Santra and Christy 2012).

Table 3. Confusion matrix

Prediction Results	Number of Documents
True Positive (TP)	65
True Negative (TN)	1
False Positive (FP)	23
False Negative (FN)	11

The prediction results in Table 3 show that 65 documents with positive classes and true predictive results, 1 document with negative class and true predictive result, 23 documents with positive classes and false, and 11 documents with negative classes and wrong prediction results.

The results of the confusion matrix are performed calculation to get the value of accuracy, precision, recall, and F1-measure (Kechaou et al. 2010). Accuracy is the percentage of TP and TN perceptions that is correctly classified manually by the researchers over the total number of perception opinions; the formula for accuracy can be seen in equation (3). Precision is the percentage of correctly classified as TP perception over the number of TP perceptions and FP perceptions; the formula for precision can be seen in equation (4). Recall is the percentage of correctly classified as TP perception over the number of TP perceptions and FN perceptions; the formula for recall can be seen in equation (5). F-measure is the percentage of precision multiply by two over the sum of the precision and recall; the formula for F1-measure can be seen in equation (6).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$Precision\ (P) = \frac{TP}{FP + TP} \tag{4}$$

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - measure = \frac{2.P.R}{P + R} \tag{6}$$

Table 4. Labeling test results

Measurement	Measurement Results
Accuracy	66.00%
Precision	73.86%
Recall	85.53%
F1-measure	79.27%

The measurement results in Table 4 show that labeling using ImpkNN and TF-IDF produces an accuracy 66.00%, precision 73.86%, recall 85.53% and F1-measure 79.27%. Overall the test results showing positive results show machine learning has good performance.

As can be noticed, the experimental results indicate that GC performs good for sentimental terms selection and sxhibits and exhibits good performance for sentiment classification in terms of precision, recall, and F-measure. As for the accuracy, it does not appear to be so good due to several problems including login, file upload, download, and update. Among these problems, the most common problem faced by users is the problem with login.

3.2. Testing of the k Value Effect

The test has been carried out five times to determine the effect of the proportion of test data in each category and the k value on the effectiveness of the classification system. There are 1.000 datasets consisting of 900 documents as training data and 100 documents as test data. The results of system testing based on k value using 100 test data can be seen in Table 5.

Table 5. System test results based on k value

k	N (new k value)		Accuracy
	Positive	Negative	
2	2	1	74.00%
5	5	2	79.00%
10	10	6	83.00%
15	15	8	83.00%
20	20	11	84.00%

Based on the results of the system test in Table 5, it can be seen that the highest accuracy is at the k value = 20 of 84.00% and the lowest is at the k value = 2 of 74.00%. So it can be concluded that if the k value is getting higher, then it is likely the test data correctly classified the higher it will be, but the category of nearest neighbor that used is not yet same as the category you are looking for, thus causing the counting result opportunities are fickle.

4. Conclusion

In this paper we present a framework for text classification based on ImpkNN algorithm and the TF-IDF method. Classification using the ImpkNN and TF-IDF algorithm can result in better of sentiment analysis. Testing with confusion matrix produces a value 66.00% of accuracy, 73.86% of precision, 85.53% of recall, and 79.27% of F-measure.

Tests carried out with different input k values resulted in different accuracy percentage values, for k = 2 of 74.00%, for k = 5 of 79.00%, for k = 10 of 83.00%, for k = 15 of 83.00%, for k = 20 of 84.00%, with highest accuracy with a value of k = 20. Based on the results of testing and analysis that has been done, it can be taken conclusion that the k value has an effect on the accuracy of the classification system is getting greater k value, the higher the chance of the test data correctly classified.

References

Păvăloaia, V-D., Teodor, E-M., Fotache, D., and Danileț, M., Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences, *Sustainability* 2019, 11, 4459; doi:10.3390/su11164459, pp. 1-21, 2019.

Arora, T., Dhawan, S., and Singh, K., Sentiment Analysis of Online Movie's Reviews Using Improved k-Nearest Neighbor Classifier, *Advances in Computer Science and Information Technology (ACSIT)*, vol. 3, issue 4, pp. 242-245, 2016.

Sebastiani, F., Machine Learning in Automated Text Categorization, *Consiglio Nazionale delle Ricerche*, 2002.

Muktafin, E. H., Pramono, and Kusrini, Sentiment analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach, *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 1, pp. 146-154, 2020.

Baesa, E. M., Bigueras, R. T., Cruz, J. D., Maligat Jr, D. E., and Torio, J. O., Sentiment Analysis of LMS Users Using Support Vector Algorithm, *Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering*, Hong Kong, 15-17 June, 2019 , pp. 288-294.

Prananda, A. R., and Thalib, I., Sentiment Analysis for Customer Review: Case Study of GO-JEK Expansion, *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 1, pp. 1-8, 2020.

Riadi, I., Sunardi, S., and Widiandana, P., Mobile Forensics for Cyberbullying Detection using Term Frequency-Inverse Document Frequency (TF-IDF), *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol.5, pp. 68-76, 2019.

Qaiser, S., and Ali, R., Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents, *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25-29, 2018.

Joseph, F., and Ramakrishnan, N., Text Categorization Using Improved K Nearest Neighbor Algorithm, International Journal for Trends in Engineering & Technology, vol. 4, issue 2, pp. 61-64, 2015.

Li, B., Yu, S., and Lu, Q., An Improved k-Nearest Neighbor Algorithm for Text Categorization, *Proceedings of the 20th International Conference of Computer Processing of Oriental Language*, Shenyang, China, June 16, 2003.

Kechaou, Z., Ammar, M. B., and Alimi, A. M., *Improving e-learning with sentiment analysis of users' opinions, Proceedings of IEEE Global Engineering Education Conference (EDUCON)*, pp. 10321038, Amman, Jordan, April 4-6, 2010.

Santra, A. K., and Christy, C. J., Genetic Algorithm and Confusion Matrix for Document Clustering, *IJCSI International Journal of Computer Science Issues*, vol. 9, issue 1, pp. 322-328, 2012.

Biographies

Boldson Herdianto Situmorang is a lecturer in the Department of Computer Sciences, Faculty of Mathematics and Natural Sciences, Universitas Pakuan. He teaches in Information System, System Analyze and Design, Software Testing, and Software Engineering. He leads research group of Decision Support System (DSS) and Multimedia.

Andi Chairunnas is a lecturer in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Pakuan. He teaches in Software Engineering, Interface and Microcontroller System and also actives on sensors and robotics research.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.