Proceedings of the
1st International Conference on Environmental Health, Socioeconomic and Technology 2022
Lamongan, Indonesia, August 10-11 2022
Vol. 1, No. 1 (https://conference.unisla.ac.id/index.php/icehst/)

CLASSIFICATION OF CRITICISMS AND SUGGESTIONS ON PUBLIC SERVICES AT RSI NASHRUL UMMAH LAMONGAN USING COMPETITIVE AUTOENCODER

Danang Bagus Reknadi[1], Yosi Kristian[2], Reddy Alexandro[3]
[1] Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia
[2] Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia
[3] Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia

danz.0907@unisla.ac.id (+62 857-3239-9045) ; yosi@stts.edu (+62 815-5377-9876) ; reddy@stts.edu (+62 817-0333-111)

ABSTRACT

One form of effort to develop an institution is by listening to feedback from its customers. The institution's response to the feedback received has very important benefits to increase customer satisfaction. In addition, to be able to improve and develop services, data about the customer's perceived condition are needed which will be used for performance evaluation. RSI Nashrul Ummah receives criticism and suggestions centrally so that the input received must be forwarded to the relevant division. The number of inputs is quite large so the information cannot be followed up quickly. The main purpose of this study is to classify incoming complaints and suggestions using the K-Competitive Autoencoder for Text (KATE) method to facilitate the process of redirecting complaints and suggestions. This study modifies the model by adding dropouts, and adjusting the threshold value adaptively so that a good representation model is produced. The dataset in this study amounted to 2241, The testing data amounted to 725 data, able to classify complaints and suggestions with F1-measure in the Administration, Facilities, Personnel, and Service classes above 85%, so that if on average it produces an F1-measure of 87.2%. The F1-measure value is calculated using a multi-label classifier using a confusion matrix for each class. The high modeling accuracy indicates that the proposed method can be used to classify reports of criticism and suggestions submitted to RSI Nashrul Ummah Lamongan.

KEYWORDS: *K-Competitive Autoencoder, Text Classification, Multilabel Classification, Hospital Comments*

1 INTRODUCTION

An institution will certainly provide the best facilities and services for every community. However, it is undeniable that every service will never be without imperfections. Every individual involved in the institution will encounter dissatisfaction and complaints about the institution's services. To be able to improve and develop its services, any complaints and concerns must be accommodated as an evaluation material and then followed up. The information system for the complaint box, visitors' criticism, and suggestions is a digital-based system that is useful for assisting visitors in submitting complaints, criticisms, and suggestions to the institution. With these criticisms and suggestions, it becomes an interesting issue to be used as a study on text mining. Text mining can be a solution for solving problems such as processing, organizing, and analyzing large amounts of unstructured text.

Text Mining, also known as Text Mining (Feldman & Dagan, 1995) was first introduced concerning the process of extracting high-quality information from structured text such as RDBMS data; semi-structured such as XML and JSON; and unstructured text resources such as word documents, videos, and images. It broadly covers a large set of topics and related algorithms for analyzing text including information retrieval, natural language processing, data mining, and machine learning. Text mining itself has a definition of mining data in the form of text where the data source is usually obtained from documents. The main purpose of text mining is that summarization is the process of grouping documents according to their categories. The benefits of document classification are very large because the number of documents is increasing every day so the role of document classification is very important.

Classification (Allahyari dkk., 2017) is one of the methods in data mining that aims to define the class of an object whose class is not yet known. In the classification, training and testing will be carried out first. In this process, a dataset whose object class is known will be used. To do the classification, what is needed is a feature. The feature must be able to distinguish one institution from another. Often, to distinguish one institution from another, a feature vector of relatively large dimensions is required. For this reason, there are feature selection methods to select features that are considered "representative" compared to other features. The simplest example of representation learning is the autoencoder.

Autoencoder is a Neural Network that can automatically learn the data representation by trying to reconstruct its input at the output layer. Many variants of the autoencoder have been proposed recently (Kingma & Welling, 2013). Although the purpose of an autoencoder is to minimize reconstruction errors, another goal is to extract meaningful features from data.

The availability of data and the need for information or knowledge to support decision-making to create solutions and support infrastructure in the field of information technology are the forerunners of the birth of data mining technology. The use of data mining techniques is expected to help accelerate the decision-making process, allowing an institution to manage the information contained in the data into new knowledge. Through the knowledge gained, an institution can improve its quality.

Until now, the availability of information on critical reports and suggestions at RSI Nashrul Ummah Lamongan has not been used as much as possible, so the data is not used optimally because no classification system and method can be used to design a decision in the categorization of criticism and suggestions.

2 RELATED WORK

In 2017, KATE or K-Competitive Autoencoder for Text was introduced by Chen & J. Zaki, t.t is a new autoencoder that relies on competitive learning among autoencoding neurons. This approach appears to overcome the weakness of the traditional autoencoder when applied to textual data. In the feedforward phase, only the most competitive k-neurons further combine the activation potential of the inactive neurons. As a result, each hidden neuron becomes better at recognizing certain data patterns, and the entire model can learn the representation of the input data. After training the model, each hidden neuron is different from the others and is not required to be competitive in the testing phase. The autoencoder approach focuses on important patterns in the data by adding constraints to the training phase to address the weaknesses. In competitive learning, neurons compete to respond to a subset of input data and as a result, the specialization of each neuron in the network increases. By introducing competition into the autoencoder, each neuron in the hidden layer can recognize different patterns in the input data. A comprehensive series of experiments shows that KATE can learn better representations than traditional autoencoders including denoising, contractive, variational,and k-sparse autoencoders. This autoencoder model also outperforms deep generative models, probabilistic topic models, and even word representation models (eg, Word2Vec) in terms of document classification.

In previous studies related to the classification of public services on the SP4N LAPOR Portal using the LSTM-RNN method (Rozi dkk., 2020). Data collection was carried out by utilizing community report data found on the report page of the SP4N LAPOR! site in the period January 2015 – December 2019. In the process of forming the classification model, preprocessing was carried out first, then the data were divided into training data and test data. The training data will be calculated using the LSTM RNN method to produce a classification model, while the test data will be used to determine the level of accuracy of the classification model generated by the LSTM RNN method. Long Short-Term Memory (LSTM) is a variation of the Recurrent Neural Network which was created to avoid the problem of a long-term dependence on the Recurrent Neural Network (RNN). LSTM can remember long-term information. The looping RNN network only uses one simple layer, namely the tanh layer. LSTM has 3 types of gates forget gate, input gate, and output gate. The forget gate is the gate that decides which information to delete from the cell. The input gate is a gate that decides the value of the input to be updated in state memory. The output gate is a gate that decides what will be produced according to the input and memory in the cell. To evaluate the LSTM method, k-fold cross-validation was used with each training model using hyperparameter batch size, epochs, optimizer, and activation. The tuning hyperparameter used in this study has resulted in an accuracy of 88.82%. The use of the number of datasets affects the increase in accuracy in this classification.

Another study is about the classification of complaints on online samba using the n-gram method and neighbor weighted k-nearest neighbor (Prasanti, 2017). In this study, distance calculations were performed using cosine similarity and feature extraction using n-grams, then calculating distances using the NW-KNN method. In this study, the most optimal value of the k-nearest neighbor is 3, which can produce the highest f-measure value of 75.25%.

3 LITERATURE REVIEW

3.1 Text Mining

Text mining is a new, developing field that attempts to gather meaningful information from natural language texts. Text Mining is generally used to denote a system that analyzes large amounts of natural language text and detects patterns of lexical or language used in an attempt to extract information that might be useful. (Witten, 2004).

3.2 Preprocessing Text

This is the initial stage in the NLP method for text documents (NLP for Text). Text Preprocessing prepares unstructured text into good data and is ready to be processed.

- The normalization stage is used to convey the idea of changing the format of the text to meet certain requirements to understand the meaning of the data presented
- The stemming stage is the stage where you find the root of the word from the filtering results or the formation of the basic word by removing the word affixes.
- The Stopword stage is the stage of taking important words using a stop list algorithm, which is removing less important words, or a word list, which is storing important words.
- Tokenizing stage is the stage of cutting the input string based on each word that composes it
- TF-IDF is the stage of assigning weights to each keyword in each category to find the similarity of keywords to the available categories.

3.3 K-Competitive Autoencoder for Text

Algorithm 1 KATE: K-competitive Autoencoder

procedure TRAINING
    Feedforward step: $\boldsymbol{z} = tanh(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$
    Apply k-competition: $\hat{\boldsymbol{z}} =$ k-competitive_layer($\boldsymbol{z}$)
    Compute output: $\hat{\boldsymbol{x}} = sigmoid(\boldsymbol{W}^T\hat{\boldsymbol{z}} + \boldsymbol{c})$
    Backpropagate error (cross-entropy) and iterate

procedure ENCODING
    Encode input data: $\boldsymbol{z} = tanh(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$

Figure 1. Pseudocode K-Competitive Autoencoder

The pseudo-code for the k-competitive autoencoder KATE is shown in Algorithm 1. KATE is an autoencoder with one competitive hidden layer, with each neuron competing to respond to a specific set of input patterns. For example:

$\boldsymbol{x} \in \mathbb{R}^d$	:	d-dimensional input vector which is also the desired output vector
$h_1, h_2, \ldots, h_m$	:	m hidden neuron
$\boldsymbol{W} \in \mathbb{R}^{dxm}$	:	the weight matrix that connects the input layer to the hidden layer neurons
$b \in \mathbb{R}^m$ dan $c \in \mathbb{R}^d$	:	respectively be biased for hidden and output neurons

In KATE, each text document is represented as a log normalized word count vector x∈R^d where each dimension is represented as

$$xi = \frac{\log(1 + n_i)}{max_{i\epsilon V}\ \log\,(1\ +\ n_i)}\ , for\ i\ \epsilon\ V \tag{1}$$

V is the vocabulary and ni is the word count I in the document. For example, ˆx is the KATE output given at input x. Binary cross-entropy is used as a loss function, which is defined as

$$l(x, 7x)^n = -\sum_{i\epsilon V}^{n} x_i \log(7x_i) + (1 -\ x_i) log(1 -\ 7x_i) \tag{2}$$

Where xî is the reconstructed value. In the feedforward phase, after calculating the activations of z for a given input x, the k neurons are chosen as the most competitive while the rest are rendered inactive. KATE uses the tanh activation function for hidden layers. division of neurons into positive and negative based on their activation. The most competitive k neurons are those with the greatest absolute activation values. However, the selection of the k/2] of the largest positive activations as positive winners. is a hyperparameter. Activation setting of all losers to zero.

Algorithm 2 K-competitive Layer

function K-COMPETITIVE-LAYER(z)
  sort positive neurons in ascending order $z_1^+ \dots z_P^+$
  sort negative neurons in descending order $z_1^- \dots z_N^-$
  if $P - \lceil k/2 \rceil > 0$ **then**
    $E_{pos} = \sum_{i=1}^{P-\lceil k/2 \rceil} z_i^+$
    for $i = P - \lceil k/2 \rceil + 1, \dots, P$ **do**
      $z_i^+ := z_i^+ + \alpha \cdot E_{pos}$
    for $i = 1, \dots, P - \lceil k/2 \rceil$ **do**
      $z_i^+ := 0$
  if $N - \lfloor k/2 \rfloor > 0$ **then**
    $E_{neg} = \sum_{i=1}^{N-\lfloor k/2 \rfloor} z_i^-$
    for $i = N - \lfloor k/2 \rfloor + 1, \dots, N$ **do**
      $z_i^- := z_i^- + \alpha \cdot E_{neg}$
    for $i = 1, \dots, N - \lfloor k/2 \rfloor$ **do**
      $z_i^- := 0$
  return updated $z_1^+ \dots z_P^+, z_1^- \dots z_N^-$

Figure 2 Algoritma K-Competitive layer

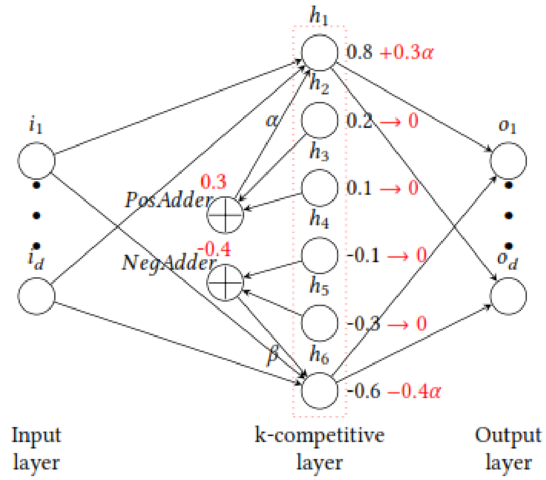

Figure 3 Competition between neurons

The description of this new k-competitive autoencoder, KATE explicitly enforces competition among neurons in the hidden layer by selecting the highest k-neuron activation as the winner. Although it uses a basic model with one hidden layer, this model outperforms the various methods on many different text analysis tasks. In this regard, across tasks such as document classification, multi-label classification, regression, and document retrieval, KATE outperformed competing methods or came close to the best results.

3.4 Multilabel Classification

Multilabel Classification is a classification to design a model that functions to label each label independently in a set of many labels (Herrera dkk., 2016). The Multi-Label classification is different from the single-label classification. In Multi-Label Classification, each dataset (institution) or collection of attributes in the dataset is associated with several labels consisting of several binary classes. Whereas in single-label classification, each dataset (institution) or collection of attributes in the dataset is only related to one label, be it binary class (binary-class) or multi-class (multi-class).

4 DATASET AND METHODOLOGY

In this study, three main processes will be carried out by the system, namely the input process, preprocessing, and the classification process using the K-Competitive autoencoder method. Generally, the workflow of the text classification system is represented in pada Figure 4.

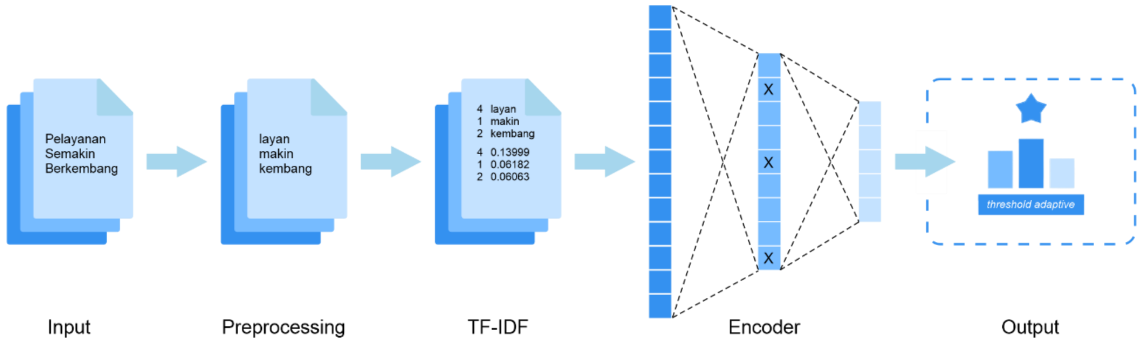

Figure 4 System flow

In this study, the data used were text data from reports and complaints from visitors or patients at RSI Nashrul Ummah, which amounted to approximately 2200 complaints. The dataset was obtained directly from the institution that became the object of study. In the class search process, each data requires several people to do the labeling in determining the majority vote. Next, preprocessing steps will be carried out, namely letter matching or case folding, tokenization, stemming, stopword removal, and weighting using TF-IDF.

Making a complaint classification model requires data that has been labeled as a source of learning data. The dataset contains two variables, namely training data and testing data.

In the training process, 2652 neurons feature was obtained. The input comes from the number of vocabularies in the training data as many as 1516 data that have gone through the preprocessing stage. From these inputs, the model reduces from 2652 to 2000 neurons in the first layer. To avoid overfitting, we added a dropout of 0.54. Finally, in K-CompetitvCompetitive, the model reduces from 2000 neurons to 1000 neurons.

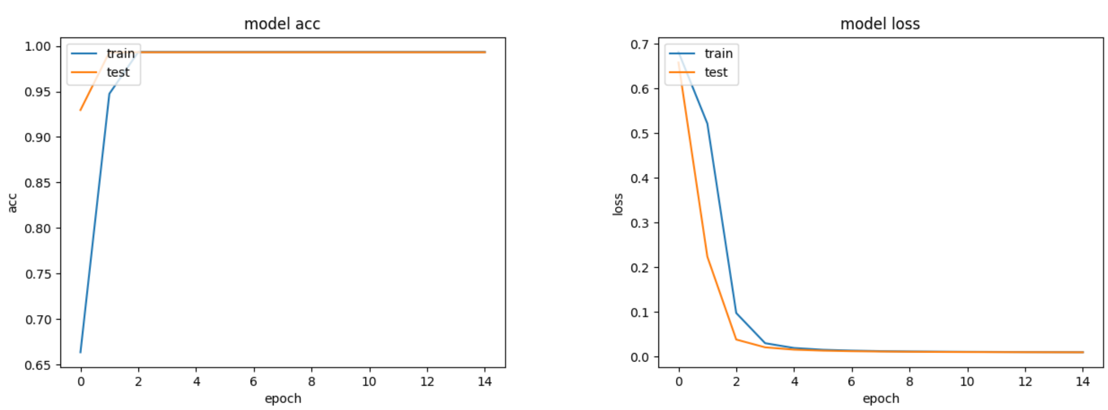

Figure 5 Graphic Performance Accuracy and Loss in the training process

Figure 5 is a graph that shows the model very well. In this training process, Adama optimization is used with epoch = 15 and batch size = 25.

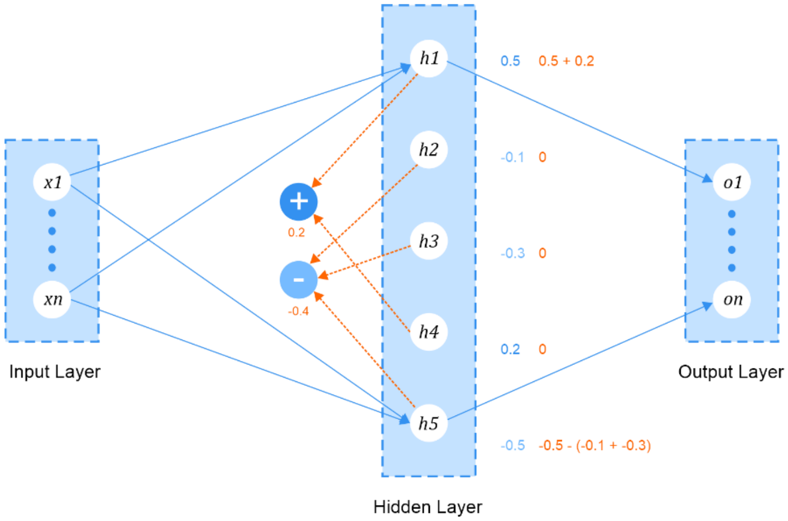

Figure 6 Architecture KATE

The following in Figure 6 is the KATE Architecture where the number of neurons in the input layer is determined from the number of vocabs in the training process with the activation function tanh in the hidden layer. In each feedforward step, there are positive and negative neurons in the hidden layer that are useful for capturing positive and negative patterns from input data. If each neuron in the hidden layer is interpreted as a class, positive neurons represent the existence of a topic in the input document, while negative neurons represent the absence of a topic.

In positive neurons, in the feedforward step, there are two positive hidden neurons, h1 and h4. The topmost competitive neuron wins, in this case, h1. The remaining neurons of h4 are the losing neurons.

Here the most competitive neuron is defined as the neuron that has the largest absolute value. More precisely, this method introduces competition by allowing the winning neurons to take the energy of the losing neurons. Energy is defined as the sum of the absolute values of activation for a given set of neurons.

In this example, the losing neuron's energy is 0.2 (of h4). By amplifying and reallocating the energy of losing neurons among winners using amplification connections. So, the new value of h1 is 0.5 + 0.2 and all losing neurons are rendered inactive or 0. Likewise for h5 which is a negative winner and it combines the amplified energy of negative losing neurons h2 and h3. h2 and h3 are deactivated in this feedforward step.

As a result, each hidden neuron that is strong in digging for certain data patterns can learn diverse patterns and meaningful representations of the input data.

When doing backpropagation, the gradient will first flow through the winning neurons in the hidden layer and then the losing neurons via the amplification connection (hyperparameter). When = 0 then no gradient will flow through the losing neuron because it has been rendered inactive in the feedforward. When > 2 / k then the gradient signal flowing will increase through the losing neuron. Here k is the total number of pre-determined winning neurons including positive and negative winners.

In the output stage of this study an adaptive threshold feature is added, that is, if the number of words in the input is more than 20 words, a threshold will be given with a certain value. In the experiment, 10 tests were conducted, with different threshold values for each test as shown in Table 1.

Table 1 Threshold value

Testing	Threshold	
	Min	Max
1	0.075	0.125
2	0.080	0.130
3	0.085	0.135
4	0.090	0.140
5	0.095	0.145
6	0.100	0.150
7	0.105	0.155
8	0.110	0.160
9	0.115	0.165
10	0.120	0.170

5 RESULT

Based on 10 experiments using the KATE method with this adaptive threshold variation, for the evaluation of this multilabel classification, we managed to get precision & recall values as shown in Table 2.

Table 2 Precision and Recall results for each class in 10 tests

Testing	Administrasi		Fasilitas		Kepegawaian		Pelayanan		lain-lain	
	Presisi	Recall	Presisi	Recall	Presisi	Recall	Presisi	Recall	Presisi	Recall
1	0.49	0.89	0.84	1.00	0.45	0.91	0.85	1.00	0.69	1.00
2	0.52	0.87	0.84	1.00	0.48	0.90	0.85	1.00	0.69	1.00
3	0.55	0.85	0.83	0.99	0.53	0.90	0.85	1.00	0.69	0.99
4	0.60	0.84	0.83	0.98	0.60	0.89	0.85	1.00	0.69	0.99
5	0.65	0.82	0.84	0.97	0.65	0.89	0.85	1.00	0.69	0.99
6	0.71	0.82	0.84	0.97	0.71	0.88	0.85	1.00	0.70	0.99
7	0.77	0.80	0.85	0.96	0.75	0.85	0.85	1.00	0.71	0.99
8	0.84	0.78	0.87	0.95	0.79	0.83	0.85	1.00	0.71	0.99
9	0.91	0.77	0.87	0.93	0.85	0.81	0.85	1.00	0.71	0.99
10	0.98	0.76	0.90	0.91	0.95	0.78	0.85	1.00	0.72	0.97

From Table 2, in the 10th testing using a threshold of 0.120 – 0.170 it was able to produce a higher precision value while the recall value showed lower results than other tests. This is because the greater the threshold value that is applied, the greater the probability value that appears for each class.

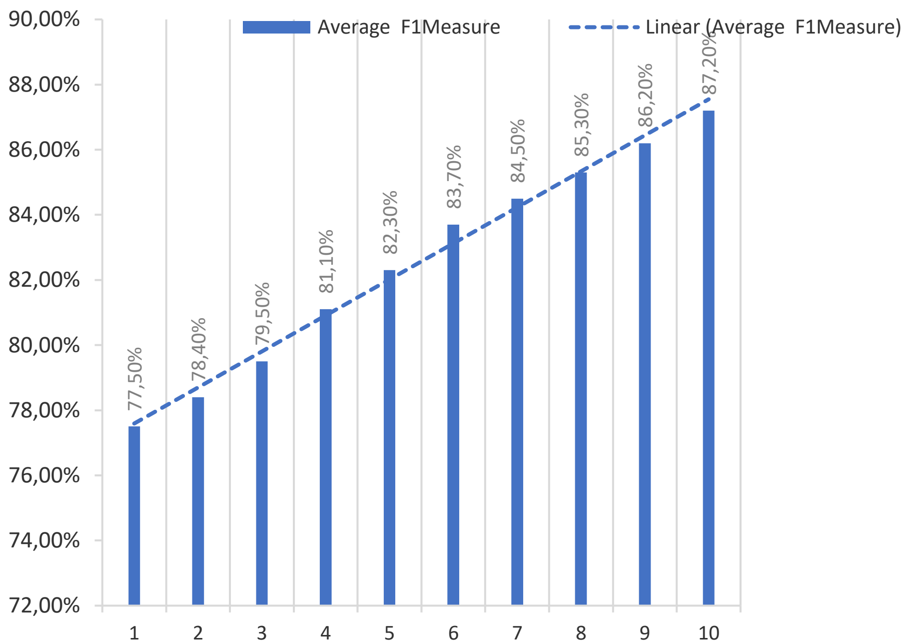

Figure 7. F1-Measure Average in 10 tests

From Figure 7, The F1 measure has increased significantly, it can be seen from the 1st test (77.50%) to the 10th test with an average F1-measure of 87.20%. This is due to the decrease in false positives in the administration class, facilities, staffing so that the precision increases.

6 CONCLUSION

This study shows that the variation of the threshold applied to the system has a large influence on the classification results. This statement is supported by the results of study on the 10th testing which shows that with a threshold value of 0.120 – 0.170 it produces a precision above 80% for the Administration, Facilities, Personnel, and Service classes. While the average f-measure reaches 87.20% compared to other threshold values. Based on the results of testing in this study, it can be concluded that the K-Competitive Autoencoder method with a combination of adaptive thresholds can classify complaints text very well.

REFERENCES

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering, and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Chen, Y., & Zaki, M. J. (2017). KATE: K-Competitive Autoencoder for Text. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 85–94.

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual database (KDT). *Proceedings of the first ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 112–117.

Herrera, F., Charte Ojeda, F., Rivera, A. J., & Del Jesus, M. J. (2016). *Multilabel classification: Problem analysis, metrics and techniques*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Prasanti, A. A. (2017). *Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N-Gram Dan Neighbor Weighted K-Nearest Neighbor (NW-KNN)* [Sarjana, Universitas Brawijaya]. http://repository.ub.ac.id/id/eprint/1571/

Prasetyo, E. (2012). *Data mining konsep dan aplikasi menggunakan MATLAB*. CV Andi Offset.

Rozi, I. F., Wijayaningrum, V. N., & Khozin, N. (2020). Klasifikasi Teks Laporan Masyarakat Pada Situs Lapor! Menggunakan Recurrent Neural Network. *Sistemasi: Jurnal Sistem Informasi*, *9*(3), 633–645.

Witten, I. H. (2004). *Computer Science, University of Waikato, Hamilton, New Zealand email ihw@cs.waikato.ac.nz*. 23.