Sentiment Analysis of School Zoning System On Youtube Social Media Using The K-Nearest Neighbor With Levenshtein Distance Algorithm

1st Nenny Anggraini, 2nd Muhammad Jabal Tursina
Informatics Engineering Department
Syarif Hidayatullah State Islamic University
Jakarta, Indonesia
nenny.anggraini@uinjkt.ac.id, ummi.masruroh@uinjkt.ac.id, muhammad.jabal14@mhs.uinjkt.ac.id

***Abstract*— The role of the education effect on the progress of each nation and State. The Ministry of education and culture (Kemendikbud) published a regulation of the Minister of education and culture (Permendikbud) No. 14 Year 2018 on the acceptance of New Learners (PPDB). Replace the previous rules, one of which is using a system of zoning for equalization learners. New Learner Acceptance Policy (PPDB) reaps the pros and cons of the zoning system. This research was conducted on the analysis of the sentiments of the public comments against the policy. The data were analyzed taken from youtube as much as 160 comments. This research uses the K-Nearest Neighbor algorithm to the value k = 3, k = 5, k = 7, k = 9 in classifying the test data and the Levenshtein Distance to fix incorrect type. This research used a combination of K-Nearest Neighbor algorithm and Levenshtein Distance that aims to find out the level of accuracy from a combination of the algorithm. Testing was done using the confusion matrix on test data. Conclusions from testing the combination algorithm for K-Nearest Neighbor and Levenshtein Distance can increase the accuracy of classification. Accuracy results from K-Nearest Neighbor algorithm, namely in the amount of 50% to the value k = 3 and k = 7 whereas the results accuracy the combination of K-Nearest Neighbor and Levenshtein Distance of 65.625% with a value of k = 3.**

Keywords— analysis of sentiment, zoning systems, K-Nearest Neighbor, Levenshtein Distance.

I. INTRODUCTION

The Ministry of education and culture (Kemendikbud) Muhadjir Effendy through regulation of the Minister of education and culture (Permendikbud) No. 14 Year 2018. The Ministry of education and culture (Kemendikbud) published a regulation of the Minister of education and culture (Permendikbud) No. 14 Year 2018 on the acceptance of New Learners (PPDB). Replace the previous rules, one of which is using a system of zoning for equalization learners. Child Protection Commission of Indonesia (KPAI) urged the existence of evaluation policy this good from 2018 PPDB Kemdikbud or Office of education throughout the region. System zoning in the process of acceptance of New Learners (PPDB) votes still has a number of weaknesses so that needs to be evaluated further. [1]

Interview with head of sub-division of Law governance and Staffing Mrs. Any Sayekti concluded that School Zoning System established on 2 May 2018 and enacted on May 8, 2018. School zoning system has been implemented throughout Indonesia for all schools except for the SMK. Kemendikbud has a different interpretation of PDSPK (Centre of education and culture Statistical Data) regarding zoning system the zoning system according to her, it's the mileage closer to learners to school while according to PDSPK the contrary, for example, there is one the school became a center of zoning. The school will look for students who are close to the school. PDSPK interpret as zoning system equitable quality of education while according to the Kemendikbud system of zoning as a closer access to students from home to school. School zoning system has several problems including namely the socialization time with implementation budget PPDB year 2018. In socialization to the district town of the province are present not all convey to the public or its stakeholders that there are below. less Government understands them in understanding and translating in the form of the type that in its territory. There are still many areas which do not comply with Regulation No. 14 year 2018 as the PPDB acceptance do not all apply the receipt of at least 90% of the homes closest to the students of the school. Running this policy still receive evaluation and determine the concept of the future so that the system can continue to run based on the evaluation of PPDB each year. Improving the quality of education can be done through policy intervention.

Based on the data WeAreSocial.net and Hootsuite 2017, growth of internet usage in Indonesia, namely the very rapidly growing 51% within one year. More than 69% of Indonesia community access the internet using their mobile devices. Results of the survey on global web index on internet users in Indonesia in the age range 16-64 years, shows that there are some social media platform that is actively used by the people of Indonesia. The platform is divided into two categories, namely social media social networking media and messenger. YouTube was ranked first with the use of the percentage of 43%, to two Facebook with the percentage of use of 41%, with a percentage of use Whatsapp then amounted to 40%. [2]

In General, opinions can be expressed over what are, for example, product, service, individual, organization, or an event. Term used to indicate the entity object that has been commented upon. An object has a set of components and a set of attributes. Defines that a sentence is a sentence that expresses opinions positive or negative, explicitly or implicitly. [3]

Levenshtein Distance or commonly referred to with the edit distance is a method that can be used to address the occurrence of the misspelling. Spelling errors may occur if the word you typed by the user is not found on the list of Indonesian Language Dictionary. The function of the Levenshtein Distance method to calculate the distance to the closeness of the two pieces of string through the addition of

characters, character conversion, and removal of characters up to second string matching. [4]

Based on the background of the above, the authors make an analysis that serves to categorize someone's comments are included in the category of positive or negative opinions.

II. Related Work

Ernawati and Wati's research entitled "application of the K-Nearest Neighbours Algorithm On the analysis of the Sentiment Review Travel Agent", researchers use data as much as 200 reviews. In addition, researchers used 10-fold cross-validation for a testing model, where each section will be set up in random. Principle 10-fold cross-validation is 1:9, 1 part data into testing and training data into other data, so the opportunity to become a part of the 10 data testing. [5]

On the research of Rizal Setya, entitled "analysis of Sentiment about the opinion of a movie on Twitter Indonesia- speaking Documents Using Naive Bayes with Repairs Not Raw", On this test data used is the original data from the Twitter user about Tweet Indonesia language film opinion. Training data are taken as many as 140 opinions, data that consists of positive opinion and data 70 70 data negative opinion. As for the test data used 60 data, which consists of 30 data opinion positive and 30 negative opinion data. Naive Bayes classification method with improvements not standard can be applied to the process of analysis of sentiment about the opinion of a movie on Twitter Indonesia-language documents. Training data and test data done pre-launch processing process in advance, which processes pre-launch processing there are additional repairs not using raw kamus_katabaku made after case folding. And Repair Word that not standard using normalization of Levenshtein Distance is done after the process of pre-processing. [6]

III. Methodology

In this study, the authors use simulation methods to see sentiment society as objects are examined regarding the Government's policy about the school zoning system using K-Nearest Neighbor algorithm and Levenshtein Distance as Repair Word is not correct. This simulation has several stages, namely:

A. Problem Formulation

The first stage is to identify the problems in the previous research results.

B. Conceptual Model

At this stage, the author does design the concept model for the simulation will be done. The first concept draft text mining processes to be used. The second concept that is identifying the comments that have been obtained for processing the data into training and test, then manually labeled. The third concept is to create a test phase to see the results accuracy results from either a sentiment algorithm using Levenshtein Distance and KNN.

C. Input/Output Data

At this point, the author makes the input and output of what will be processed on the simulation later. Comments data from the Youtube API will be used as input 160 comments, in the form of 128 training data and 32 test data. This data is then processed using the K-NN algorithm and Levenshtein Distance that generates output in the form of accuracy.

D. Modeling Phase

At this stage, the author design scenario model to be created. Modeling made the first scenario i.e. using the K-NN algorithm. The second scenario is a combination of the K-NN algorithm and Levenshtein Distance.

E. Simulation Phase

At this stage, the system was run to simulate the performance of the algorithm according to a predetermined scenario. The simulation is performed with the input dataset, labeling the sentiment of the dataset, do the training on the training data, perform classification using test data. The results of the simulation in the form of a comparison of the accuracy of the algorithms used in this research.

F. Verification, Validation and Experiment

At this stage, the author to verify and validate from the previous stages, so that simulations are ready to run and perform experiments in accordance with model scenarios have been made before

G. Output Analysis

At this last stage, the author does analysis of output resulting from the scenario that's been done, namely to calculate the accuracy of the algorithm used in this study

IV. Implementation and Results

On testing done by as much as 3 times to ensure the accuracy of the system. The first stage in the research of the preprocessing, following the flowchart preprocessing phase to labeling sentiment.

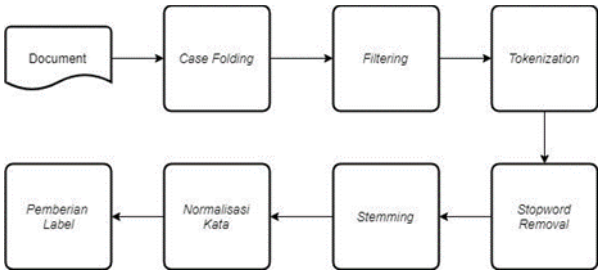

Fig. 1. Preprocessing stage and labeling

1. The first step in preprocessing that is case folding or change all words be lowercase.
2. the second stage, namely filtering to remove character URL links, hashtags, etc.
3. The third stage, tokenization or to break a sentence into words per word.
4. The fourth stage, stopword removal to remove words deemed not important as I, you, he, etc.
5. The fifth stage, stemming to change words into basic words.

6. The sixth stage, namely the normalization of words using the Levenshtein distance, every word of the dataset will be matched with words of KBBI to take the smallest edit to change the value of the word to be changed.
7. The seventh stage, namely labeling sentiment positive and negative form.

- Case Folding

TABLE I. CASE FOLDING

Document Text	Case Folding
Saya Bukannya Anakk ajaibb!!	saya bukannya anakk ajaibb!!

- Filtering

TABLE II. FILTERING

Document Text	Filtering
saya bukannya anakk ajaibb!!	saya bukannya anakk ajaibb

- Tokenization

TABLE III. TOKENIZATION

Document Text	*Tokenization*
saya bukannya anakk ajaibb	| saya | bukannya | anakk | ajaibb |

- Stopword Removal

TABLE IV. STOPWORD REMOVAL

Document Text	*Stopword Removal*
| saya | bukannya | anakk | ajaibb |	| bukannya | anakk | ajaibb |

- Stemming

		b		e		n		c		i	
	0		1		2		3		4		5
b	1		0		1		2		3		4
e	2		1		0		2		3		4
n	3		2		1		0		1		2
c	4		3		2		1		0		1
i	5		4		3		2		1		0
i	6		5		4		3		2		1
i	7		6		5		4		3		2
i	8		7		6		5		4		3

Fig. 2. Calculation Of The Levenshtein Distance

From the example of the word typo "benciiii" being the correct word is "benci" obtained the smallest distance that is 3.

Following are the results of the classification using KNN and Levenshtein Distance.

TABLE V. RESULT KNN

Test	K Value	Accuracy
Testing 1	3	56.25%
	5	53.125%
	7	53.125%
	9	53.125%
Testing 2	3	65.625%
	5	62.5%
	7	62.5%
	9	62.5%
Testing 3	3	62.5%
	5	59.375%
	7	62.5%
	9	62.5%

Results of the KNN and LV distance tests shown in the table above, Generate accuracy based on predefined k values.

TABLE VI. STEMMING

Document Text	*Stemming*
| bukannya | anakk | ajaibb |	| bukan | anakk | ajaibb |

- Normalization of Words

TABLE VII. NORMALIZATION

Document Text	Normalization of Words
| bukan | anakk | ajaibb |	| bukan | anak | ajaib |

Every word of the dataset will be matched with words of KBBI to take the smallest edit to change the value of the word to be changed.

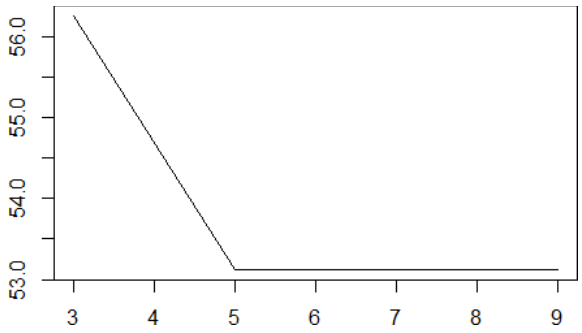

Fig. 3. Calculation Of The Levenshtein Distance

The chart above displays the best accuracy on the first test generated with a value of k = 3. The result is 56.25%.

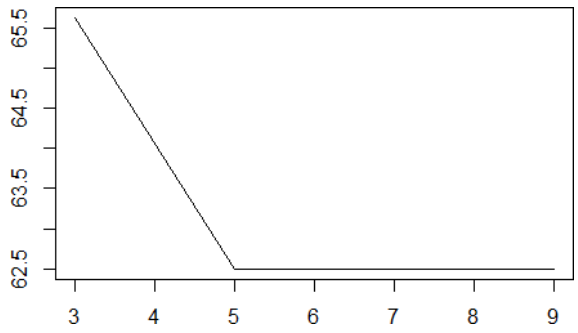

Fig. 4. Accuracy results using KNN and Levenshtein Distance on the second testing

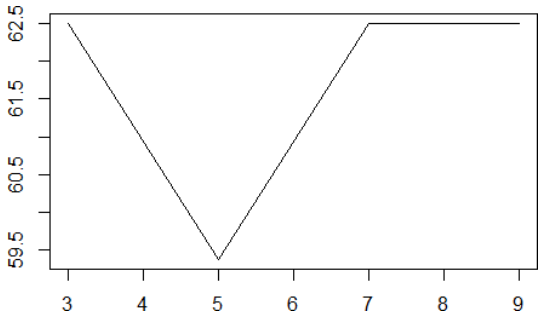

Fig. 5. Accuracy results using KNN and Levenshtein Distance on the second testing

The chart above displays the best accuracy on the second test generated with a value of k = 3. The result is 65.625%.

Based on testing the combination of k-nearest neighbor and Levenshtein distance is done above the highest accuracy of the results obtained in the second test when the value k = 3 is 65,625%. The increase in the value of accuracy caused by the existence of the word typo normalized so as to increase the frequency of the word.

V. Conclusion And Suggestions

Based on the discussion that is done, then it can be inferred that:

1. A combination algorithm for k-nearest neighbor and Levenshtein distance can be applied to the analysis of sentiment with the highest accuracy from a combination of K-Nearest Neighbor and Levenshtein Distance is indicated at the time of the value k = 3 is 65,625% when the second test.
2. Combination algorithm for K-Nearest Neighbor and Levenshtein Distance can increase accuracy in classifying analysis of sentiment in social media youtube.
3. The conclusion that can be taken is increased accuracy results with a combination of K-Nearest Neighbor and Levenshtein Distance caused by the repair of inappropriate words or typo into words that correspond to KBBI.

Suggestions that should be done, as follows:

1. Increase the number of training data that are used in the process of classification.
2. Use cross validation to find the optimal value of accuracy.
3. Use other spell checking algorithms to see how the resulting accuracy than Levenshtein Distance algorithm.

References

[1] B. P. Nugroho, "Ramai soal PPDB, begini aturan aistem aonasi sekolah," 4 July 2018. [Online]. Available: https://news.detik.com/berita/d-4097504/ramai-soal-ppdb-begini-aturan-sistem-zonasi-sekolah.[Accessed 28June 2018].

[2] K. Data, "Inilah media sosial dengan pengguna aktif terbesar di indonesia," Kata Data, 13 September 2017. [Online]. Available: https://databoks.katadata.co.id/datapublish/2017/09/13/inilah-. media-sosial-dengan-pengguna-aktif-terbesar-di-indonesia. [Accessed 13 June 2018].

[3] S. Adi, "Perancangan klasifikasi tweet berdasarkan sentimen dan fitur calon gubernur dki jakarta 2017," perancangan klasifikasi tweet berdasarkan sentimen dan fitur calon gubernur dki jakarta 2017, vol. III, pp. 10-16, 2018.

[4] P. Antinasari, R. P. Setya and M. A. Fauzi, "Analisis aentimen tentang opini film pada dokumen twitter berbahasaanalisis aentimen aentang opini film aada aokumen twitter berbahasa andonesia menggunakan naive bayes dengan perbaikan kata tidak Baku," Analisis Sentimen Tentang Opini Film pada Dokumen Twitter Berbahasa Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku, vol. I, pp. 1733-1741, 2017.

[5] S. Ernawati and R. Wati, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel," Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel, vol. VI, 2018.

[6] M. D. A. Putri, A. Syukur, A. Prihandono and D. Rosal, "Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes," Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes, 2018.