Clustering Data National Examinations Based on Social Media Using K-Means Method

Chandra Eko Wahyudi Utomo
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
Universitas Jember, Jember, Indonesia
chandra15@mhs.ee.its.ac.id

Mochamad Hariadi
Department of Computer Engineering
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mochar@ee.its.ac.id

Surya Sumpeno
Department of Computer Engineering
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
surya@ee.its.ac.id

***Abstract*— The development of social media as a source of data is now increasingly interesting to study. The social media studied in this research is Twitter. Twitter as one of the top-ranked social media among social media accessed by the people of Indonesia. People's behavior can be learned by collecting and processing data, one of which is people's sentiments or opinions about national examinations in Indonesia. Twitter user behavior in the form of their comments about the national exam in Indonesia. This study aims to analyze the public sentiments of social media users about the National Examination in Indonesia. Data is retrieved by crawling data via the Twitter API. The data needs to be preprocessed first and feature extracted using TF-IDF. However, because the text data on Twitter is unstructured and very diverse data (variety), the grouping stage must be done first. Grouping technique using K-Means Clustering on Spark. Spark clustering techniques are used to overcome the grouping of data on very large and complex amounts of data. From the clustering process using Spark it was found that the grouping process resulted in 3 clusters where elbow detection was found in the third cluster of the number of clusters between 2 and 50. The results of clustering in the form of 3 large groups were further processed (with classification techniques) to get a positive or negative sentiment comparison of social media user comments about the national exam. Furthermore, these results become recommendations and new knowledge about community behavior regarding Social Media-based National Exams.**

Keywords—clustering, sentiment analysis, national exam, social media, K-Means

I. INTRODUCTION

In computer science, one of the fields of study of emotion related to social media is sentiment analysis. Sentiment analysis is a way to obtain public sentiment based on a data processor or machine learning so that later it is useful to assess whether a product is accepted by the community or not [1]. In this case, the sentiment analysis has an impact such as being able to influence the behavior of the people in certain studies, which is to raised public sentiment towards Jokowi's candidacy as a candidate for the President of Indonesia in 2014 [2]. Sentiment on social media is so important because it provides insight into people, supports customer service (in this case the customer service of government agencies). In addition, it can also inform the message of the institution / company. When using sentiment analysis tools such as Hootsuite Insights, public relations (PR) can see when conversations around their brands change negatively. This tool will recognize unusual spikes in conversation volume - and measure tones.

The research that the author did in the topic of sentiment analysis was the public sentiment of Twitter users towards the National Examination in Indonesia. Twitter media was chosen because it is one of the most popular social media and is considered to represent the upper middle class Indonesian community along with the rapid growth of information technology.

Data on social media can be used for research material. One of the many social media users is Twitter. Twitter, with more than 313 million monthly active users and more than 500 million tweets per day [4], is a gold mine for organizations and individuals who have strong social, political or economic interests in maintaining and increasing their influence and reputation. Twitter is a micro-blogging social network that is a very fast emerging platform for users to express their views on politics, sports products etc. This view is useful for businesses, governments and individuals. For this reason, tweets can be a valuable source for mining public opinion [1]. Tweets usually consist of incompleted, noisy and unstructured sentences, irregular expressions, incorrect words and non-dictionary terms. Before feature selection, pre-processing sequences (eg, deleting stop words, removing URLs, replacing negations) were applied to reduce the amount of noise in tweets. [5]

The implementation of the National Examinations in Indonesia in the last few years has experienced pros and cons, both for the organizers of the State and the community. The Minister of Education and Culture of the Republic of Indonesia 2014 - 2019, Prof. Muhajir Efendi briefly threw down the idea of a National Examinations moratorium. This is due to the negative impact of the National Examinations which reduces the nature of education and causes many education actors to be tempted to act dishonestly (perhaps they can include real conditions or sources of information). This is reinforced by the decision of the Constitutional Court which states that the National Examinations cannot be used as a benchmark for student graduation but it is the school that determines it. However, the Vice President of the Republic of Indonesia 2014 -2019 HM Yusuf Kalla expressed his rejection because the National Examinations was very important to be implemented to control the quality of national education and standardization in measuring student achievement. Even in developed countries such as Britain, Japan, China and Singapore, the National Examinations is still the final evaluation material for the development of education in a country. Finally, in the Cabinet Session on 7 December 2016 it was decided that the 2017 National Examination would still be held [6]. It's just that, the National Examination is still a conversation in the community with all its advantages and disadvantages.

The benefits of sentiment analysis are very important to determine the extent of the opinion of the social media community on national examinations in Indonesia and used as a tool to see the public response to the implementation of national examinations in Indonesia. In connection with the large amount of comment data, a data analysis process is needed that is able to analyze this problem.

II. Related Works

Previous research which discussed about sentiment analysis has been very much with various fields of study topics. Since Boo Pang and Lilian Lee [7] conducted research on film review as the beginning of a 2004 study of sentiment analysis, this science experienced rapid development. In 2009 Martineu and Finin's research [8] examined Delta TF-IDF (Term Frequency - Inverse Document Frequency) Feature Space for Sentiment Analysis and 2010 research Pak and Paroubek [9] discussed about sentiment analysis on twitter. Based on the IEEE journal, Indonesian language sentiment analysis research began in 2010 by Colbaugh and Grass [10] concerning Jakarta Bombing Intelligence Monitoring. Furthermore, the study of sentiment analysis leads more to the implementation of methods using either the supervised learning or unsupervised learning methods or machine learning or sentiment research. Authors are interested in researching sentiment analysis using spark. This is because when trying to preprocessing data using the normal method, it takes a very long time for large amounts of data.

The limitations in this paper are reviews on pre-processing data and clustering data for automatic labeling processes.

III. Methodology

The complete scheme of this sentiment analysis study can be shown in Figure 1. The stages of the sentiment analysis process begin with preprocessing, clustering and classification. The initial stage of this research is the data collection stage. Data obtained from Twitter through the API is then stored in a database. Then the next step is preprocessing.

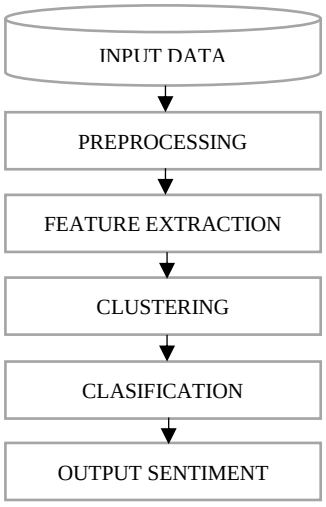

Fig. 1. Sentiment analysis schema

This preprocessing stage runs in Apache Spark. The preparation phase consists of several parts (1) Lowercase, (2) Delete Mention, (3) Remove Link, (4) Delete Hashtags, (5) Delete Retweet, (6) Emoji -> Text, (7) Delete Duplicate Characters , (8) Remove Punctuation.

Then feature extraction is the calculation of the term Frequency value - The frequency of the inverse document (TF-IDF) of each tweet. This TF-IDF process produces features from every tweet. The feature used in the next process is grouping.

The grouping process runs using the K-Means method. This grouping process runs with a number of centroids ranging from 2 to 50. Each iteration is evaluated using the In Set Sum of Squares Errors (WSSSE) generating graphs. The graph is analyzed and then detected with an elbow to get the ideal number of clusters. The results of this ideal cluster are used as a sentiment category.

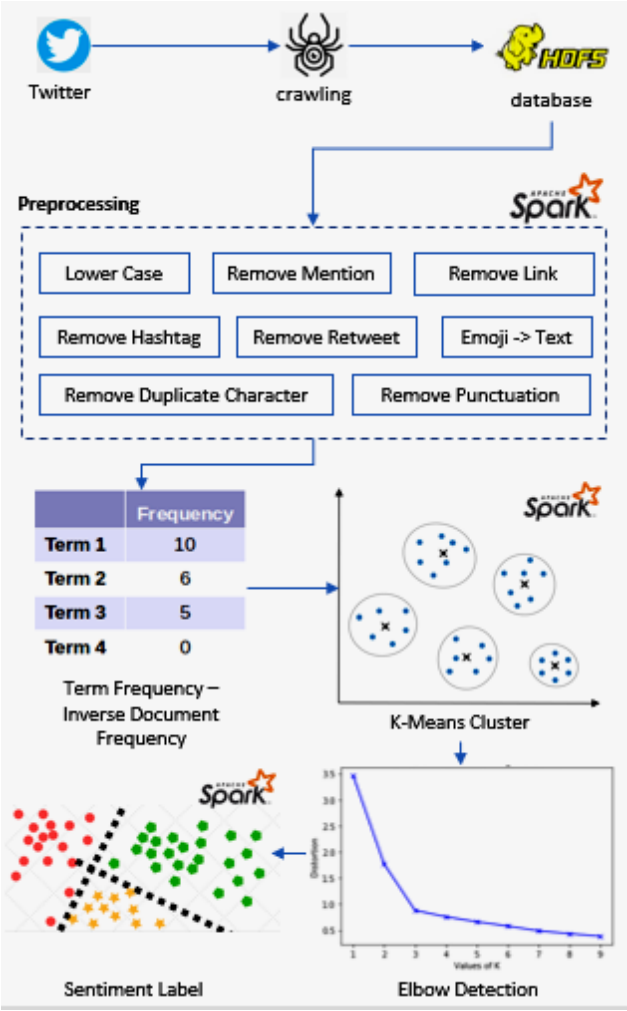

Fig. 2. Overview of system design

A. Term Frequency – Inverse Document Frequency

TF-IDF (Term Frequency - Inverse Document Frequency) is a popular method for weighting words. Word weighting is a mechanism to score the frequency of occurrence of a word in a text document. One of the weighting stages of the feature used is weighted unigram using Term Presence (TP), Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF). Words and symbols are represented in vector form, where each word or symbol is counted as a feature. The weight calculation used is:

1) Feature Term Frequency (TF)

$$\vec{d} := (n_1(d), n_2(d), \ldots n_m(d)) \quad (1)$$

2) Feature Term Presence (TP)

$$n_i(d) = 1, \text{ if feature } f_i \text{ in document d} \quad (2)$$

$$n_i(d) = 0, \text{ if feature } f_i \text{ is not in document d} \quad (3)$$

3) Term Frequency – Inverse Document Frequency (TF – IDF)

$$n_i(d) = df_i . \text{Log } D/df_i \quad (4)$$

B. K-Means Clustering

Clustering method models document collection as vector space model, with the dimension of total word in document collection. Suppose we have a collection of document D= $\{d_i \mid i=1,2,\ldots |D|\}=\{d_1, d_2, \ldots, d_{|D|}\}$ that will be clustered into K cluster. Firstly we parse all documents in the collection into n unique words. The frequency of all words in the document collection then was counted. The collection of document D will be represented as collection of document vector in the dimension of n, where n is the number of unique word. The element of the vector was the word frequency of occurrence in each documents. N order to get more accurate computation of the documents similarity we did normalization using weighting scheme called term frequency (TF) and invers document frequency (IDF). Combine this scheme we get weighting for document feature using TF-IDF normalization weighting:

$$W_{i\varphi} = \frac{(\ln(fij)+1).\log(\frac{N}{ni})}{\sqrt{\left((\ln(fij)+1).\log\left(\frac{N}{ni}\right)\right)^2}} \quad (5)$$

Where f_{ij} represents frequency of word-i in document – j, N represents total number of document, n_i represents total number of document that containts word-i. Using this weighting it can be guaranteed that the length of the vector is one.

C. Elbow Detection

Elbow method is a method which looks at the percentage of variance explained as a function of the number of clusters. This method exists upon the idea that one should choose a number of clusters so that adding another cluster doesn't give much better modellingof the data. The percentage of variance explained by the clustersis plotted against the number of clusters. The first clusters will add much information but at some point the marginal gain will drop dramatically and give an angle in the graph. The correct "k" i.e. number of clusters is chosen at this point, hence the "elbow criterion". The idea is that Start with K=2, and keep increasing it in each step by 1, calculating clusters and the cost that comes with the training. At some value for K the cost drops dramatically, and after that it reaches a plateau when you increase it further. This is the K value you want. The rationale is that after this, you increase the number of clusters but the new cluster is very near some of the existing [11]. This research use Within Set Sum of Squared Errors (WSSSE) [12] as variance calculation.

IV. Experiment

The experiment contains several steps before deliver to sentiment category.

A. Crawling Twitter Data

Data collection is done by mining data about the National Examination on Twitter through the Twitter API. This data is obtained by entering the keyword National Examination or National Exam on the Twitter API (Application Programming Interface). The resulting data is in the form of people's comment data on Twitter about the National Examination in .txt and CSV format.

TABLE I. Public Opinion in Twitter on National Examinations

No	Student Opinion for National Examinations
1	Semoga nem ujian nasional bagus yaallah udh bagus besar pula amin! o:)(˘∫∏˘)
2	Apakah kolaborasi lembaga bimbingan + sekolah ini yg ikut mempengaruhi agar ujian nasional dihapuskan ? #nanyserius
3	RT @byanmega: Ujian Nasional taun depan 20 paket Hunt
4	Ternyata Nembak cewek itu lebih sulit dari pada ujian nasional...!!!
5	yang kelas 2 jangan seneng dulu mau jadi kelas 3, nanti pas ujian nasional lo kesel sendiri jadi kelas 3 #okesip

B. Preprocessing

Preprocessing consists of several stages including (1) Lower Case, (2) Remove Mention, (3) Remove link, (4) Remove Hashtag, (5) Remove Retweet, (6) Emoji is converted into text, (7) Remove duplicate character, (8) Remove punctuation. The following is an illustration of the results of preprocessing.

TABLE II. Preprocessing Result

Raw Text
RT @StorySMU: *ujian nasional* izin ke toilet sebentar ke pengawas. Padahal kunci jawaban di selipin di atas pintu WC #storySMU 211424704777560000
Preprocessing
USER_MENTION ujian nasional izin ke toilet sebentar ke pengawas. padahal kunci jawaban di selipin di atas pintu wc storysmu
: Lower Case : Remove Mention : Remove Link : Remove Hashtag : Remove Retweet : Emoji -> Text : Remove Duplicate Character : Remove Punctuation

At the preprocessing stage, a folding case is carried out to equalize the letters. Then do the cleaning, to clean from letters other than the alphabet. Then the tokenizing process, breaking the sentence into words. Then the filtering stage to eliminate words that are stopword.

C. K-Means Clustering

The clustering process uses the K-Means method consisting of two stages, namely training and testing. Training is the process of learning patterns from data. The

data used as training data is from tweets that have been crawled. The results of the clustering process are shown in Table III.

TABLE III. CLUSTERING RESULT

Cluster	Tweet (in Indonesian)
0	smpek skarang masih inget bnget wktu minta doa buat ujian nasional sama rikaendingnya kita nanges breng
	USER_MENTION dulu waktu aku lulus ujian nasional tukang pos jeung motorna dipanggul ku urang nepi ka pangandaran saking ku bahagiana sw
	kamu pasti bisa hahaha USER_MENTION kk tanggal aku mau ujian nasional pleaseeee semangaaat nyaa aku gugup mau ngomong apa
1	besok ak liburan USER_MENTION ujian nasional telah usaiUSER_MENTION ujian hidup sedang dimulai
	USER_MENTION brb pindah ke cinaUSER_MENTION cina larang siswi peserta ujian nasional pakai bra LINK
	USER_MENTION brb pindah ke cinaUSER_MENTION cina larang siswi peserta ujian nasional pakai bra LINK
2	kk tanggal aku mau ujian nasional pleaseeee semangaaat nyaa aku gugup mau ngomong apa stres
	just did getop ipa ujian nasional geschool
	yassalam kaya ngadepin soal ujian nasionalo

The clustering technique on Spark for sentiment analysis research on the national exam produces the best 3 clusters from the distribution of the number of clusters between 2 and 50 clusters. Table III shows information that there are 3 groups of clustering results using K-Means on Spark. Cluster data is defined as a group, so basically the cluster analysis will produce a number of clusters (groups). This analysis begins with the understanding that a certain amount of data actually has similarities among its members. Therefore, it is possible to group members that are similar or have similar characteristics in one or more than one cluster. In determining the results of clustering, the researcher used a count of the number of words from each sentence to form 3 clusters. All of this is done in the Spark analyzer tool.

D. Elbow Detection

The final stage of Clustering is testing the clustering algorithm using WSSSE (Within Set Sum of Squared Errors). The purpose of this method is to find out how many optimal "clusters" for a dataset. The script of the testing shown in Fig 3.

```
import scala.collection.mutable.ListBuffer

var listBufferWSSSE = new ListBuffer[(Int,Double)]()

for( numCluster <- 2 to 50){
  var model = KMeans.train(vectors, numCluster, numIterations)
  var WSSSE = model.computeCost(vectors)
  println("Within Set Sum of Squared Errors untuk "+ numCluster +" Class = "+WSSSE)
  var value = (numCluster, WSSSE)
  listBufferWSSSE += value
}

val listWSSSE = listBufferWSSSE.toList
val WSSSEDf = listBufferWSSSE.toDF("Num of Cluster", "WSSSE")

display(WSSSEDf)
```

Fig. 3. Script of clustering testing

The Fig 3 shown that calculation of variance in this process uses Within Set Sum of Squared Errors (WSSSE). This data grouping uses the K-Means method which is repeated iteratively with n values ranging from 2 to 50. Then the optimal n value is searched using the elbow method. Calculation of variance is performed at each iteration of cluster construction. For each iteration, the WSSSE value is calculated. Clusters are made from clusters with centroids totaling 2 to 50.

The calculation of the variance forms the graph shown in Fig 4. The graph shows that the formation of elbows on clusters with centroid 3, so that the number 3 is chosen as the number of centroids in making cluster of sentiment labels.

E. Sentiment Label

The results of clustering are sentiment labels that are obtained automatically. The label is assumed to be a sentiment class divided into 3 classes. The three classes are derived from 3 clusters that have been formed from the results of clustering using the K-Means method. The Fig 4 shown that the elbow method interprets and validates consistency in cluster analysis to determine the optimal number of clusters. From the results of the elbow method used in the range of clusters 2 to 50, the optimal cluster is obtained in the 3rd cluster.

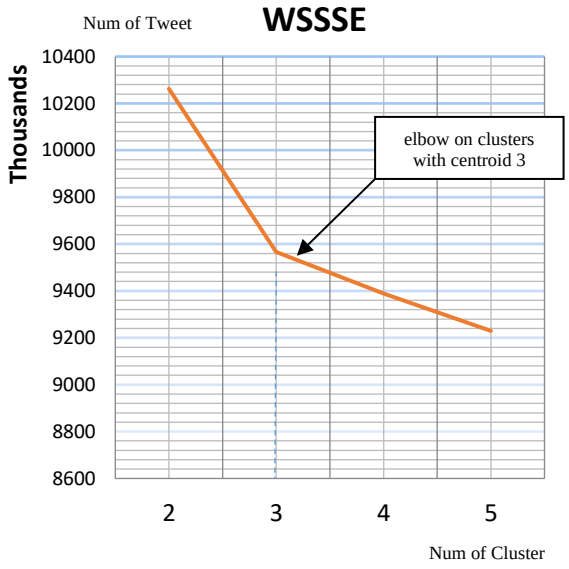

Fig. 4. The formation of elbows on clusters with centroid 3

F. Rule-based Sentiment Score

The rule-based sentiment score process is the process of labeling tweets automatically according to sentiment based on existing keywords. Rule-based Sentiment Score is part of sentiment analysis to understand, extract, and process textual data automatically to get sentiment information contained in an opinion sentence. In this study, it refers to research conducted by Rahmat [13], which uses a previously developed rule-based. In this study, in classifying sentiments, using 3 classes, namely positive, negative, and neutral. To do sentiment analysis, it is necessary to have a word sentiment dictionary and rule data. The word sentiment dictionary is used as a word sentiment reference and the rule is used as a sentiment calculation technique. The following explains the steps for calculating sentiment using rule-based: (1) Word Sentiment Dictionary. The word sentiment dictionary is used to assign values to

each word. This value is in the form of numbers 1, -1 and 0 where 1 is positive, -1 is negative, and 0 is neutral. In addition to sentiment values, the word sentiment dictionary also stores word types to make it easier for the system to create an opinion sentiment assessment rule later. The Fig 5 shown that a dictionary of sentiment that has been developed from research conducted by Rahmat [13]; (2) Labeling words. Before an opinion mining process is carried out with the implementation of the impression rule, the preprocessing words must be labeled, based on the word sentiment dictionary. Due to the limited vocabulary in the sentiment word dictionary, some words that are not in the word sentiment dictionary database will get an unknown word type and sentiment value is 0 / neutral; (3) Sentiment Rule. After the sentimental word dictionary database, and labeling the words, the rule design process begins. Rule is used to provide rules for commenting sentiment assessment. This process does not use a special algorithm, but rather an impression technique. This technique is simpler than using an algorithm. Impression techniques are more inclined to analyzing the wording of a sentence [5].

TABLE IV. DICTIONARY WORD IN INDONESIAN

Word	Type	Value	Word	Type	Value
Berlanjut	Verba	1	Dibaca	verba-di	1
Berlawanan	Verba	-1	Dibahas	verba-di	1
Berlebihan	Verba	-1	Dibalas	verba-di	1
Berlindung	Verba	1	Dibalik	verba-di	0
Bermain	Verba	0	Dibanding	verba-di	0
Bermakna	Verba	1	Dibandingkan	verba-di	0
Bermanfaat	Verba	1	Dibangun	verba-di	1
bermasalah	Verba	-1	Dibantai	verba-di	1
Bermobil	Verba	0	Dibarengi	verba-di	0
Bermuara	Verba	0	Dibatalkan	verba-di	-1
Bernama	Verba	0	Dibatasi	verba-di	-1
Berontak	Verba	-1	Dibawa	verba-di	0
berorientasi	Verba	0	Dibawah	verba-di	-1
Berpengalaman	Verba	1	Dibayang	verba-di	-1
Berpengaruh	Verba	1	Dibela	verba-di	1

This technique analyzes the location of adjectives, verbs, and prepositions in a sentence. Prepositions are words that compose words or parts of sentences and are usually followed by nouns or pronouns, for example no, not yet, very, etc. The table V shown that word labeling based on dictionary.

TABLE V. WORD LABELING BASED ON DICTIONARY [5]

Word Preprocessing Result (in Indonesian)	Type	Value Sentiment
Diperparah	Verb_di	-1
Jokowi	Noun	0
Mengeluarkan	Verb	-1
kebijakan bebas visa	Noun	0

This technique analyzes the location of adjectives, verbs, and prepositions in a sentence. Prepositions are words that compose words or parts of sentences and are usually followed by nouns or pronouns, for example no, not yet, very, etc. In this study using research conducted by Rahmat [13]. The data label referred to in this research is the data label based on the Indonesian language data dictionary. In research conducted by Rahmat [13], using 51 rules, which are divided into 3 categories as follows.

TABLE VI. RULE CATEGORY

Category Name	Information
Active Verb	Word combination rules that make active verbs their focal point
Passive Verb	Word combination rules that make passive verbs their focal point
Adjective	Word combination rules that make adjective their focal point

The results of automatic labeling can be seen in Table VII. In automatic labeling using the rule based score on Spark obtained 3 groups of clusters marked with numbers 0, 1 and 2.

TABLE VII. SENTIMENT LABEL

Tweet (in Indonesian)	Cluster
USER_MENTION waaa o USER_MENTION yang kelas jangan seneng dulu mau jadi kelas nanti pas ujian nasional lo kesel sendiri jadi kelas	0
waaa o USER_MENTION yang kelas jangan seneng dulu mau jadi kelas nanti pas ujian nasional lo kesel sendiri jadi kelas okesip	0
hahecie sp USER_MENTION dapet like dari orang itu sangat sulit. sulitnya melebihi ujian nasional	1
USER_MENTION itu bio nya kok ada alumni emang udah lulus udah ujian nasional	1
just did getop ipa ujian nasional geschool	2
cuman diinformatika nilai ujian nasional matematika dapet x_x o	2

V. RESULT AND DISCUSSION

The data crawling process successfully collected as many as 124.612 tweet data. Then the data is processed in clustering, so as to get an elbow value of 3. The results of the clustering with 3 centroids will be used as sentiment labels. These series of processes can run well on the big data platform, thus saving processing time. However, the results obtained have not been able to show the accuracy of label sentiment. The process of separating clusters tends to be based on the contents of the tweet and the short length of the tweet. As an example is cluster 2, tweets classified as cluster 2 have tweets that tend to be short. While cluster 1 and 0 have user_mention, the difference is that cluster 1 has shorter content, while cluster 0 has longer content.

TABLE VIII. TWEET ANALYSIS

Cluster	Tweet (in Indonesian)	Word
0	smpek skarang masih inget bnget wktu minta doa buat ujian nasional sama rikaendingnya kita nanges breng	16
	USER_MENTION dulu waktu aku lulus ujian nasional tukang pos jeung motorna dipanggul ku urang nepi ka pangandaran saking ku bahagiana sw	21
	kamu pasti bisa hahaha USER_MENTION kk tanggal aku mau ujian nasional pleaseeee semangaaat nyaa aku gugup mau ngomong apa	19
1	besok ak liburan USER_MENTION ujian nasional telah usaiUSER_MENTION ujian hidup sedang dimulai	12

	USER_MENTION brb pindah ke cinaUSER_MENTION cina larang siswi peserta ujian nasional pakai bra LINK	14
2	kk tanggal aku mau ujian nasional pleaseeee semangaaat nyaa aku gugup mau ngomong apa stres	15
	just did getop ipa ujian nasional geschool	7
	yassalam kaya ngadepin soal ujian nasionalo	6

Based on the labeling results using the clustering technique with the K-Means method, it is found that positive sentences when processed with Spark can be labeled negative. This is interesting to study because it is different from conventional sentiment analyzers, especially those using the semantic lexicon approach. Based on previous studies, a sentence can be labeled (positive or negative) at least by some linguists to ensure that the sentence has a positive, negative or neutral sentiment. Spark's automatic labeling is based on the number of words in a certain sentence. The Table VIII shown that certain sentences can be labeled positive because the sentence structure has complete sentence structure requirements (consisting of Subject, Predicate, Object and Description). However, the definition of a word is negative, but because the number of words is large, the sentence is positive. With clustering, the algorithm only uses the TF-IDF feature value. So, the labeling is only based on the TF-IDF value which is actually computed by the program.

VI. Conclusion

In this research, it can be concluded that the preprocessing of the National Examination sentiment data can actually be done by using Spark on the unstructured data model, labeled automatically. Because the National Examination sentiment data is in the form of unclassified sentences, the grouping is done by the K-Means method. From the results of the grouping is equipped with automatic labeling. Automatic labeling uses the rule based score method on Spark and produces 3 clusters, namely cluster 0, cluster 1 and cluster 2. In terms of speed, unstructured data grouping and automatic labeling techniques using the Spark analyzer tool is much faster than using manual techniques.

Acknowledgment

The authors would like to thank to Universitas Jember for giving the school the opportunity and supporting this research, and we also very grateful to Kemenristekdikti which has provided scholarships.

References

[1] A. P. Jain, “Application of Machine Learning Techniques to Sentiment Analysis,” pp. 628–632, 2016.

[2] H. T. Gemilang, A. Erwin, and K. I. Eng, “Indonesian president candidates 2014 sentiment analysis by using Twitter data,” *Proc. - 2014 Int. Conf. ICT Smart Soc. “Smart Syst. Platf. Dev. City Soc. GoeSmart 2014”, ICISS 2014*, pp. 101–104, 2014.

[3] J. Messias *et al.*, “An evaluation of sentiment analysis for mobile devices,” *Soc. Netw. Anal. Min.*, vol. 7, no. 1, p. 20, 2017.

[4] Z. Jianqiang and G. U. I. Xiaolin, “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis,” vol. 5, 2017.

[5] A. Krouska, C. Troussas, and M. Virvou, “The effect of preprocessing techniques on Twitter sentiment analysis,” pp. 1–5, 2016.

[6] R. Koordinasi, “Persiapan UN dan USBN,” 2016.

[7] B. Pang, L. Lee, H. Rd, and S. Jose, “Thumbs up ? Sentiment Classification using Machine Learning Techniques,” 1988.

[8] J. Martineau and T. Finin, “Delta TFIDF : An Improved Feature Space for Sentiment Analysis,” no. May, 2009.

[9] Pak, A. and Paroubek, P. (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the 7th International Conference on Language Resources and Evaluation, 1320-1326.

[10] Glass K and Colbaugh R. Estimating the sentiment of social media content for security informatics applications. Security Informatics 2012; 1(3).

[11] P. Bholowalia, A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN", 2014.

[12] Spark, A.: Clustering - spark.mllib (2016). http://spark.apache.org/docs/latest/mllib-clustering.html. Accessed 05 November 2019.

[13] Rahmat Heru Kurniawan, Real Time Opinion Mining of Social Media about Indonesian Government Policy , Tugas Akhir Sarjana Terapan Politeknik Elektronika Negeri Surabaya, Surabaya, 2017.