PROC. INTERNAT. CONF. SCI. ENGIN.
Volume 3, April 2020 | Pages: 749-754

ISSN 2597-5250
E-ISSN 2598-232X

Analysis of Public Opinion on Religion and Politics in Indonesia using K-Means Clustering and Vader Sentiment Polarity Detection

Tanzilal Mustaqim
Computer Science Department, Faculty of Science and Mathematics, Semarang State University
H Building Campus Sekaran Gunungpati Semarang Postal Code: 50229 Tel: (024) 8508092/93 Fax: (024) 8508093/8508082.
Email: tanzilal14@students.unnes.ac.id

Abstract. Religion and politics are two things that are closely related to each other and cannot be separated. Various public responses expressed by various public media such as print media and social media that can be classified as positive, neutral and negative, one of which is using Twitter. Twitter is a microblogging social media that contains many writings with many types from various types of users including posts that contain opinions about religion and politics. This research conducted an analysis process in the form of extraction of hidden insight data, visual analysis and sentiment analysis of public opinion related to religion and politics. The analysis was conducted on 5433 datasets written on Twitter on November 12, 2019. The analysis process began with data pre-processing, data clustering and sentiment analysis. Pre-processing data generates clean data from characters and non-essential data for use in the process of data clustering and sentiment analysis. Data clustering produces extraction of hidden insight data using k-means clustering. Sentiment data analysis uses vader sentiment polarity detection to determine dataset sentiments. The results of tests carried out using jupyter notebook show insight data hidden in the form of 50 unique words that are divided into 5 clusters of 10 words each then the sentiment analysis process is carried out in each cluster. Another result is visual analysis in the form of word cloud and hashtag clustering which shows the dominant words of each piece of data according to sentiment and word count. Also pointed out words that have a frequency of dominant emergence accompanied by word sentiments. The process of analyzing public opinion datasets related to religion and politics using k-means clustering and vader polarity detection sentiments can be done well.

Keywords: k-means clustering, politic, religion, sentiment analysis, twitter, vader sentiment polarity detection

INTRODUCTION

Religion and politics are two things that are closely related to each other and cannot be separated. Both are equally the way of life that affects the daily life of the community. Religion is more directed to the rules made by God for the benefit and goodness of each element of society in the life of this world and the hereafter. Politics aims to regulate the life of the wider community which is discussed and determined by a collection of several people for the benefit of many people. Community life is very much attached to both religious and political activities such as holidays and the election of regional leaders. Various public responses expressed by various public media such as print and social media can be classified as positive, neutral and negative. Public opinion can be analyzed to find out the main discussion topics that are hot and dominant emotional sentiment.

The results of the analysis can be used by related parties to make it easier to determine the right decision in accordance with the conditions of the community. The social media used for analysis is Twitter on the grounds that Twitter is microblogging which contains a limited opinion of 140 characters so that it can facilitate the analysis process. The analysis process for determining the main topic and sentiment analysis is called text mining with the aim of extracting hidden information in the Twitter data set. The process of determining the main topic using the machine learning algorithm method, namely k-means clustering. The use of k-means clustering because it can easily group data similar to the use of memory and low resources.

K-means clustering has also been used to study sentiment analysis of the use of vaccines for children [8]. The process of sentiment analysis in Twitter is often done to find out positive, neutral or negative opinions from the public regarding a particular matter. Examples of research sentiment analysis on Twitter are sentiment analysis about the elections in Jakarta in 2017 and sentiment analysis about the marketplace. There are many methods used for sentiment analysis such as unsupervised learning and supervised learning. unsupervised learning is a method that is not supervised in the sense that no training data is provided beforehand and when the data analysis process is divided into sections into 2, 3 and so on Then for supervised learning is a method that is supervised, the intention is the process of analyzing data using training data that has been provided or has been labeled data. Apart from these two methods, there is also sentiment analysis using a lexical or dictionary approach. Lexical method is a method that does not require training data or data that has been labeled but which is available in a dictionary complete with the sentiment diversity. Vader sentiment polarity detection is an example of a lexical sentiment analysis method.

Twitter data analysis using k-means clustering and sentiment analysis methods can be useful to retrieve important data complete with the emotional sentiment polarity of the twitter dataset. From the above background this research intends to carry out a Twitter data analysis process to find out the main topics of discussion as well as the emotional sentiment polarity of public opinion related to the inauguration of the new Indonesian cabinet in 2019.

MATERIALS AND METHODS

Study Area

The writing method in this research is to find and retrieve information from literature related to religion, politics, k-means clustering and vader sentiment polarity detection. The research data collection was taken by scraping directly from Twitter using the Python Tweepy library. The research data was taken and the analysis process was carried out on November 12, 2019.

Procedures

Research procedure

The work procedure of this study is explained in Figure 1. The dataset for the analysis was obtained from 5433 Twitter tweets with queries relating to "religion" and "politics".

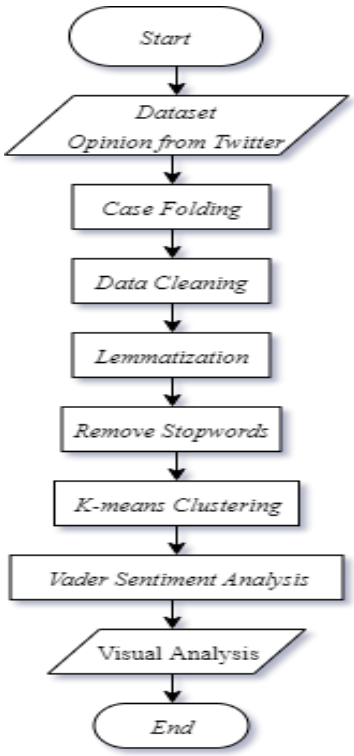

Figure 1. Research Flowchart.

Case Folding

Case folding is a method for turning all the letters in a dataset into capital or all small. This is done to facilitate the process of dataset analysis and reduce the amount of memory usage. An example of a folding case is to change the phrase "Fight For Religion" to "fight for religion". Case folding helps the lemmatization and stemming process to find a match for each data in the dictionary.

Data Cleaning

Every tweet data from social media Twitter usually contains many words and characters that are not useful for the data analysis process. For example there are data tweets such as "RT @fightagain: let's fight for religion together", in the data found words or characters that are not useful as "RT", "@" and "?". Data cleaning methods combined with regex can detect useless characters and are immediately deleted from the main data to improve the quality of the dataset.

Lemmatization

Non-standard words are often used in communicating and interacting with others. Non-standard words are formed as a result of human interaction itself and sometimes far from the standard rules of the original language dictionary. In the sentiment analysis of nonstandard words is very influential on the results of data analysis calculations. To improve the results of sentiment quality, it is necessary to repair or change non-standard words into standard words. Lemmatization is the process of changing nonstandard words into their original languages.

Remove Stopwords

In a text document usually there are words that are not very useful such as prepositions, conjunctions, adjectives, slank words, pronouns and much more. These words usually appear together with the main word so that it is not unique and does not have a specific meaning. A list of words that do not contribute too much to analytical text is called a stopword or stoplist. Stopwords do not have the potential to be indexed documents. A stoplist is unique because each language has its own stoplist. Deleting stopwords in a dataset can improve the quality of analysis sentiment.

K-means Clustering

K-means clustering is a vector quantization method, derived from signal processing, which is popular for cluster analysis in data mining. k-means clustering aims to divide n observations into k clusters where each observation belongs to the cluster with the closest mean, functioning as a prototype of the cluster. This results in partitioning the data space into Voronoi cells. k-Means minimizes in-cluster variance (Euclidean distance squared), but not the usual Euclidean distance, which will be Weber's more difficult problem: on average optimizes quadratic errors, whereas only geometric medians minimize Euclidean distances. Better Euclidean

solutions for example can be found using k-median and k-medoid.

Vader Sentiment Analysis

Vader is an acronym for Valence Aware Dictionary for Social Reasoning which is used as a model for sentiment analysis and is able to determine the diversity of data through the intensity of existing emotional power in accordance with the Lexicon data dictionary available. Vader was introduced in 2014 by C.J Hutto and Eric Gilbert whose method of formation was based on a human-centric approach, combining qualitative analysis and empirical validation using human wisdom and judgment. Vader is able to provide a different polarity between "I like you" and "I don't like you". The polarity assessment combines lexical dictionary features with sentiment scores of 5 additional criteria namely exclamation points, uppercase letters, degree of word order, polarity shift due to the word "but" and uses the tri-gram feature to check the existence of negations. The lexical approach aims to map words into sentiments by constructing a lexicon or 'sentiment dictionary.' We can use this dictionary to assess sentiment of phrases and sentences, without needing to look at anything else. Sentiments can be categorized - such as {negative, neutral, positive} - or can be numerical - such as the range of intensities or scores. The lexical approach looks at the sentiment category or score of each word in a sentence and decides what sentiment category or score the whole sentence is. The strength of the lexical approach lies in the fact that we don't need to train the model using labeled data, because we have everything we need to judge sentiment sentences in the emotional dictionary. Vader is an example of the lexical method.

Visual Analysis

Visual data analysis is an analysis process that delivers processing results using visual graphics such as bar graphs, pie charts and other graphic models. The visual analysis process makes it easy for readers, especially those who are not experts in the field of related sciences, can easily understand and interpret according to the results obtained. Insight or valuable information can be derived from the results of visual analysis such as dominant words that often appear in tweets. This study uses visual analysis to add new insights from public opinion datasets related to religion and politics.

RESULTS AND DISCUSSION

The process of analyzing public opinion related to religion and politics begins with collecting data directly using the tweepy program package that has been linked to Twitter API. Twitter data criteria collected are adjusted to query words related to religion and politics such as "religion" and "government". The total amount of data collected was 5433 tweets. Example 5 raw data tweets taken from Twitter are shown in table 1.

Table 1. Twitter raw data.

No.	Tweet
1	'RT @Nasutn_eD: Tapi kenapa negara Arab yang di salah kan,apa mungkin krn agama Islam berasal dari Arab□□☺☺\r\n#AyoHadiriReuni212 #AyoHadi…'
2	'RT @jiemiardian: Konselor, psikolog, psikiater, dokter, guru, pemuka agama, atau profesi penolong lain; yuk kita kembali ke hakikat profesi…'
3	'RT @HANET__: Tak Selesaikan Kasus Novel, Haris Azhar: Jokowi Presiden Spesialis Bikin Janji\r\n\r\n*Yang sepakat RT lur petelur \r\nhttps://t.co/g…'
4	'RT @jiemiaian: Konselor, psikolog, psikiater, dokter, guru, pemuka agama, atau profesi penolong lain; yuk kita kembali ke hakikat profesi…'
5	"RT @faizrhan33: kimakgirl starter pack :\r\n\r\n- acah reject orang \r\n- post mengemis attention kat socmed \r\n- pretends like she's the one has b…"

Twitter raw data that has been collected is converted into a uniform form and in this study, all characters are changed to lowercase in order to facilitate the analysis process and minimize memory usage. Examples of the results of case folding to lowercase are in table 2.

Table 2. Case folding.

No.	Tweet
1	'rt @nasutn_ed: tapi kenapa negara arab yang di salah kan,apa mungkin krn agama islam berasal dari arab□□☺☺\r\n#ayohadirireuni212 #ayohadi…'

From the case folding process, the Twitter data is then cleared of many characters or pieces of data that have no contribution to the data analysis process such as deleting the rt character, url address, mentioning the user and other unique characters. An example of data cleaning results is shown in table 3.

Table 3. Data cleaning.

No.	Tweet
1	‘tapi kenapa negara arab yang di salah kan,apa mungkin krn agama islam berasal dari arab #ayohadirireuni212 #ayohadi…’

Twitter data are often contained slang words that are often used in writing opinions and different from the rules of writing a large Indonesian dictionary. Twitter conversion data processing process is not appropriate so that the data is more accurate when the analysis process is done. Examples of changing slang words to become more standard are in table 4.

Table 4. Lemmatization.

No.	Tweet
1	‘tapi kenapa negara arab yang di salahkan,apa mungkin karena agama islam berasal dari arab #ayohadirireuni212 #ayohadi…’

The next process is stopword remover with the aim of removing common words that often appear in the text such as deleting the words "and", "the", "because of" and so on. Examples of the application of stopwrod remover are shown in table 5.

Table 5. Stopword Remover.

No.	Tweet
1	‘kenapa negara arab di salahkan,mungkin agama islam berasal arab #ayohadirireuni212 #ayohadi…’

After the Twitter data has been cleared, the core analysis process is carried out, namely k-means clustering and vader sentiment analysis. The process of clustering unique topics from twitter uses the tf-idf selection feature whose function is to find unique words from the Twitter collection. The results of k-means clustering are in table 6.

Table 6. K-means clustering.

No.	Cluster
1	Political, religious, politics, radicalism, like, just, right, don, want, god
2	Religion, radical, islam, ministry, minister, don, god, selling, Indonesia, different
3	People, religious, religion, don, like, president, radical, want, pretending, freed
4	President, Jokowi, mr, just, case, paloh, corruption, Bolivia, import, party
5	Greetings, religions, mui, say, muslim, jatim, hello, use, appeal, central

From the unique topics that have been found later in lexicon sentiment polarity detection using vader sentiment analysis. The process of detecting sentiment is based on the polarity values of words that have been determined in accordance with human judgment and then summed based on the many words in a sentence. The results of sentiment analysis are shown in table 7.

Table 7. Cluster data sentiment analysis.

No.	Cluster	Sentiment
1	Political, religious, politics, radicalism, like, just, right, don, want, god	Positive
2	Religion, radical, islam, ministry, minister, don, god, selling, Indonesia, different	Negative
3	People, religious, religion, don, like, president, radical, want, pretending, freed	Positive
4	President, Jokowi, mr, just, case, paloh, corruption, Bolivia, import, party	Negative
5	Greetings, religions, mui, say, muslim, jatim, hello, use, appeal, central	Negative

The core analysis process provides new insights into the public opinion dataset on religion and politics. Visual analysis is also carried out in this study to provide additional information that tends to be easily understood by lay readers. Visual analysis is shown by giving wordcloud, a collection of dominant words that are arranged attractively with a certain color and size according to the frequency of appearance in the dataset. There are three wordcloud appointments namely neutral, negative and positive wordcloud as in Figure 1.

Figure 1. Wordcloud Opinion Dataset.

In addition to wordcloud visual analysis, public opinion datasets related to religion and politics were analyzed with a hash perspective. Twitter data is sometimes supplemented with a hashtag, whose writing starts with a "#" hash mark. Hastag Twitter is usually used by many people to show their opinions regarding events that are currently happening in the community. Hastag visual analysis is grouped into 2 namely positive and negative. The results of the hashtag visual analysis are shown in Figure 2 and Figure 3.

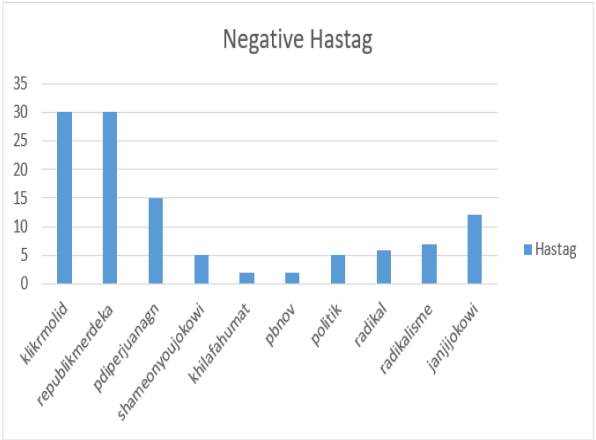

Figure 2. Negative Hastag.

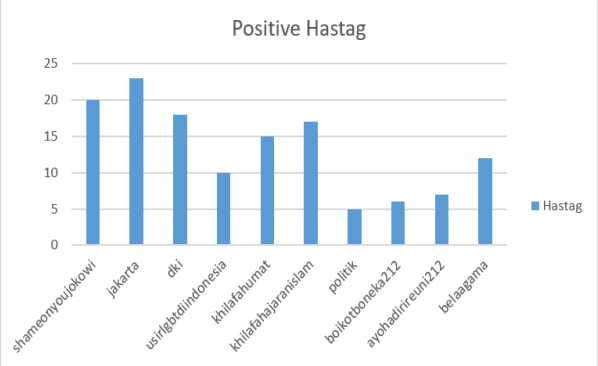

Figure 3. Positive Hastag.

Then the data analysis is done by grouping words based on the dominant frequency of occurrence then in the analysis of the value of sentiment polarity using vader sentiment polarity detection. A list of the top 10 selected words is shown in table 8.

Table 8. Top Ten Words.

No.	Top Ten Words	Frequency	Polarity Value	Sentiment Result
1	Agama	572	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
2	Presiden	351	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
3	Politik	224	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
4	Islam	150	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
5	Radikal	145	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
6	Salam	132	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
7	Jokowi	123	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
8	MUI	110	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
9	Indonesia	104	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral
10	Ulama	102	'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0	Neutral

CONCLUSIONS

The results of this study indicate that various types of sentiment polarity from public opinion are positive, negative and neutral. The results of clustering using k-means clustering machine learning algorithm produces 3 negative sentiments and 2 positive sentiments. The polarity of negative sentiment tends to result from the appearance of the words "radical" and "hate" then the polarity of positive sentiment tends to result from the appearance of the words "like" and "god". Words that have the frequency with which a dominant word appears are more likely to produce a neutral sentiment polarity.

ACKNOWLEDGEMENTS

The author would like to thank the Faculty of Science and Mathematics at the Semarang State University for providing facilities to support the research process and my partner Aprilia Dewi Ardiyanti for her assistance and enlightenment.

REFERENCES

A. Sholikin, "Pemikiran Politik Negara dan Agama Ahmad Syafii Maarif," *J. Polit. Muda*, vol. 2, no. 1, pp. 194–203, 2012.

Abdullah Zawawi, "Politik Dalam Pandangan Islam," *Ummul Quro*, vol. 5, no. Jurnal Ummul Qura Vol V, No 1, Maret 2015, pp. 85–100, 2015.

A. F. Bahary, Y. Sibaroni, and M. S. Mubarok, "Sentiment analysis of student responses related to information system services using Multinomial Naïve Bayes (Case study: Telkom University)," in *Journal of Physics: Conference Series*, 2019, vol. 1192, no. 1.

B. G. Julian, I. Budi, and D. Tanaya, "Performance of DKI Jakarta Governor and Vice Governor on 2017-2018 based on Sentiment Analysis using Twitter and Instagram Data," in *Proceedings of 2019 2nd International Conference on Data Science and Information Technology (DSIT'19). Seoul, Republic of Korea*, 2019, pp. 122–127.

C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).","*Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, 2014.

E. Gunawan, "Relasi Agama dan Negara (Perspektif Pemikiran Islam)," *J. al-Hikmah*, vol. 15, no. 2, pp. 185–200, 2014.

F. Rahutomo, A. Retno, and T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 1, p. 41, 2019.

H. Du *et al.*, "Twitter vs News: Concern Analysis of the 2018 California Wildfire Event," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 2019, vol. 2, pp. 207–212.M. Noori and J. Maktoubian, "Business Improvement Approach Based on Sentiment Twitter Analysis : Case Study," *EAI Endorsed Trans. Cloud Syst.*, pp. 1–7, 2019.

J. P. Pinto and V. Murari, "Real Time Sentiment Analysis of Political Twitter Data Using Machine Learning Approach," *Int. Res. J. Eng. Technol.*, vol. 6, no. 4, pp. 4124–4129, 2019.

J. Garay, R. Yap, and M. J. Sabellano, "An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment

analyzer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 482, no. 1, 2019.

N. T. Romadloni, I. Santoso, and S. Budilaksono, "Perbandingan Metode Naive Bayes , Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line," *J. IKRA-ITH Inform.*, vol. 3, no. 2, pp. 1–9, 2019.

N. F. Rozi, F. Arianto, and D. P. Hapsari, "Analisis Sentimen Pada Opini Pengguna Maskapai Penerbangan Sentiment Analysis on Passenger Opinions At Airlines Company," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, pp. 321–326, 2019.

R. A. Plunz *et al.*, "Landscape and Urban Planning Twitter sentiment in New York City parks as measure of well-being," *Landsc. Urban Plan.*, vol. 189, no. April, pp. 235–246, 2019.

S. Kumar, M. Yadava, and P. P. Roy, "Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction," *Inf. Fusion*, vol. 52, pp. 41–52, 2019.

S. Yang and H. Zhang, "Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis," *Int. J. Comput. Inf. Eng.*, vol. 12, no. 7, pp. 525–529, 2018.

U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis," *Sci. J. Informatics; Vol 6, No 1 Mei 2019*, vol. 6, no. 1, pp. 138–149, 2019.