Indonesian Lexicon-Based Sentiment Analysis of Online Religious Lectures Review

Rahmad Kurniawan
Department of Informatics Engineering Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru, Indonesia
https://orcid.org/0000-0002-0957-9480

Fitra Lestari
Department of Industrial Engineering Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru, Indonesia

Abdul Somad Batubara
Visiting Professor, Universiti Islam Sultan Sharif Ali,
Bandar Seri Begawan, Brunei Darussalam

Mohd Zakree Ahmad Nazri
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
Bangi, Selangor, Malaysia

Khairunnas Rajab
Faculty of Psychology, Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru, Indonesia

Rinaldi Munir
School of Electrical Engineering & Informatics, Institut Teknologi Bandung, Bandung, Jawa Barat,
Indonesia

2021 International Congress of Advanced Technology and Engineering (ICOTEN) | 978-1-6654-1224-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOTEN52080.2021.9493530

***Abstract*— Online videos platforms such as YouTube is the most popular social media platform in terms of user numbers in Indonesia. YouTube is also one of the most popular online platforms for accessing religious lectures. Users can provide feedback on the videos through comments, likes, and shares. On the other hand, sentiment analysis in the Indonesian language is getting popular, but few have tapped the vast unstructured data source on YouTube. Comments and reviews from viewers are valuable feedbacks for improvements. The review on YouTube is an essential resource to be analyzed by a preacher. However, manual analysis of YouTube reviews is complicated due to a large amount of review data. Therefore, this study aims to analyze sentiment on YouTube video reviews. In this paper, we employed the Lexicon and *Latent Dirichlet Allocation* (LDA) to analyze a total of 2575 review data. In this case study, we mined YouTube user's review to understand the netizen's opinion on a famous Islamic Preacher in South East Asia, namely Ustadz Abdul Somad (UAS). We employed the Google Apps Script (GAS) with Javascript coding language to crawl YouTube review data. Based on the results, the lexicon method successfully analyzed sentiments with an accuracy of 70%. Furthermore, 98% of YouTube users gave positive reviews on the UAS videos lecture. This study is a stepping stone for more complex sentiment analysis regarding text pre-processing and algorithm robustness.**

Keywords*— *Lexicon-based, Latent Dirichlet Allocation, sentiment analysis, Ustadz Abdul Somad, Natural Language Processing, Islamic preacher

I. Introduction

Like any other country globally, Indonesia's preacher has going digital to engage with the community of believers and enhance their knowledge with various topics. YouTube is a social media platform where users can upload videos to be viewed by many people. YouTube is currently the most popular social media platform in terms of user numbers [1]. Users can provide feedback on videos through comments, likes, and shares on YouTube [2]. Users will use this function to see what the community thinks about the video they have submitted.

According to Statista, YouTube is Indonesia's most successful social network, with a subscriber of about 94% [3]. All recently well-known social media places, such as WhatsApp and Instagram, have a strong invasion rate in Indonesia, making it one of the world's largest social media markets.

However, the incoming data flow on the YouTube server is 191.1 Gigabytes per minute, making manual analysis difficult [2]. Additionally, several reviews are lengthy and include just a few sentences expressing the user's thoughts on the UAS videos. On the other hand, AI technology is increasingly progressing in the direction of unstructured data [4]. Feedback, critiques, and other comments made online by internet users make up social media data. These remarks can represent feelings not often captured by conventional data collection methods like filling out a survey questionnaire. As a result, social media data provides a rich source of knowledge that can be adequately analyzed and comprehended.

II. Sentiment Analysis

Sentiment analysis, also known as opinion mining, uses text analysis techniques to understand and identify the emotions (positive, negative, and neutral) expressed in language. In general, the researchers examined are online writing to determine the author's emotional tone. Social media were used to get a source of knowledge from what is shared by its users [5]. There have been several studies regarding sentiment analysis on social media [6]. Two of them have used deep learning [7], [8]. Based on the literature review, the studies on sentiment analysis tend to develop merely in English. The exclusive reliance on social media data in the English language is a significant weakness in most current research. Recently, researchers have been tried using another language, such as using the Chinese language [9], [10]. This study gap must be filled by developing appropriate sentiment analysis methods and data approaches in non-English languages.

Therefore, the use of the Indonesian language in sentiment analytic studies also needs to be considered because Indonesia

is one of the largest countries of social media users in the world. Indonesian researchers also aimed to use Indonesian in analytical sentiment research. A sentiment analysis using Indonesia tweets was conducted using Lexicon [11]. User opinions for the Indonesian language in Google Play and App Stored were also performed [12], [13].

Opinion analysis may be accomplished in two ways: via the use of a lexicon-based method or through the use of machine learning. In the context of sentiment analysis, one of the two primary ways is using a lexicon, which entails deriving the sentiment from the semantic orientation of words or phrases that appear in a piece of writing. These methodologies are used in opinion analysis to distinguish between opinions that are positive, negative, or neutral with respect to a certain issue. Support Vector Machines, the Naive Bayes classifier, Maximum Entropy, Deep Learning, and k-Nearest Neighbors are some of the most popular machine learning algorithms for sentiment analysis. They are used in many applications.

In this study, Ustadz Abdul Somad (UAS), who is one of the most famous preachers in South East Asia, is chosen as the subject matter or context for this research. UAS is well respected and considered a reference point in Indonesia, Brunei, Singapore, and Malaysia [4]. UAS followers on YouTube have reached 1.83 million subscribers [14]. The proposed method is tested on the collected comments from the UAS channel to understand the viewer's sentiment regarding Ustadz Abdul Somad's lecture videos quickly based on the Lexicon method. The lexicon method was chosen because it is simple and proven effective in determining positive and negative opinions in a sentence [11]. It was furthermore owing to a lack of lexicon method which is developed in the Indonesian language. Few studies have been conducted to examine using a lexicon-based sentiment analysis approach for the Indonesian language [12].

Furthermore, we also need to find out the distribution of words in user opinion data. A cluster of words in a collection of documents can be found while parsing all the documents to identify a potential topic. The *Latent Dirichlet Allocation* (LDA) algorithm assumes that a text comprises several unknown topics. The global distribution of unknown topics in the corpus and the distribution of each document in the corpus can be detected using LDA on a set of documents.

III. Materials and Methods

The primary source of the data was obtained from YouTube. We accessed four videos from YouTube containing UAS lectures. The footage was selected based on the latest and most popular videos; thus, it was expected that the new videos would receive many reviews and comments from YouTube.

A total of 2575 reviews were achieved from four videos. Furthermore, the review data were scrubbed using text pre-processing techniques in machine learning.

Fig. 1 concisely explains the methodology that has been implemented.

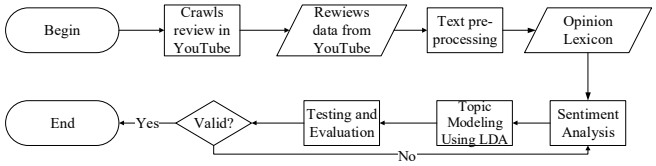

Fig. 1. The methodology of lexicon-based sentiment analysis on YouTube Videos

A. Crawls Reviews in YouTube Videos

We employed the Google Apps Script (GAS) with Javascript coding language to obtain YouTube reviews data.

The brief *pseudocode* for getting a user review of Ustadz Abdul Somad Videos on YouTube as follows:

```
function Crawls Reviews on UAS Videos (){
  var result=[['Name','Comment']];
  while{
      var data = UAS Videos
      for (var row=0; row<data.items.length;
row++) {
       result.push([data.items[row].snippet.topLev
       elComment.snippet.authorDisplayName}
       }
```

B. Text Pre-Processing

1) Transformation

Transformation is a step in the process of removing unnecessary words from the text. Transformation is needed to minimize noise and improve classification accuracy. *Numbers, icons, URL links, and hash tags* are among the characters that were deleted. The transformation was also needed to convert all the letters in a text to lowercase letters. Several terms can be used for this stage, such as the following.

- All text was turned to lowercase and removed all diacritics or accents in the text. Example: *Ābdul Somad → abdul somad*.
- Parse HTML to be detected HTML tags and parsed out text only. *<a href. . . >Ustadz Abdul Somad → Ustadz Abdul Somad*
- URL to be removed from the text. Example: *Video ini menarik dibahas dan ditulis di web http://somadmorocco.blogspot.com/2016/06/shalat-witir-2-1-adakah-dalilnya.html → Video ini menarik dibahas dan ditulis di web.*

2) Tokenizing

Tokenizing is a stage that converts a sentence into a single word. Several terms can be used for this stage, such as the following.

- Word & punctuation is to keep punctuation symbols and divide the text by words. For instance: *Abdul Somad. → (Abdul), (Somad), (.)*
- Whitespace has divided the text only by whitespace. For instance: *Abdul Somad. → (Abdul), (Somad.)*
- Sentence will be separated the text by full stop, retaining only full sentences. For instace: Syaikh UAS. *Ustadz Abdul Somad. → (Syaikh UAS.), (Ustadz Abdul Somad.)*
- Regular expression will be separated from the text by provided *regex*. It splits by words only by default (omits punctuation).

3) Normalization

We applied stemming and lemmatization to words. Stemming is a method of determining the origin of an affixed noun. Stemming is the process of removing affixes from a phrase, which can be *prefixes, infixes, suffixes,* and *co-fixes* (a mix of *prefixes* and *suffixes*). *UDPipe Lemmatizer* [15] for the Indonesian language was applied to text normalization.

4) Filtering

Filtering (stop removal) was applied to a common word or conjunction used in various sentences, especially in the meaningless conjunction, and has no bearing on the meaning. Several examples of conjunctions are *and, at, to, from, there, with, will, etc*. Stop word removal's primary goal is to reduce the number of words in a document to speed up word processing.

C. Indonesian Lexicon-Based Sentiment Analysis

The Sentiment analysis stage is applied for predicts sentiment for each document in a corpus that will be received. We have employed the lexicon-based sentiment analysis algorithm for the Indonesian language. Indonesia sentiment lexicons were used from the *Data Science Lab* [16]. The lexicon-based method calculates the polarity of the reviews on the YouTube video as a corpus using Formula (1) for the positive sentiment (P_s) and Formula (2) for the negative sentiment (N_s).

$$P_s = \sum_{i\,\varepsilon\,t}^{n} P\ score_i \qquad (1)$$

$$N_s = \sum_{i\,\varepsilon\,t}^{n} N\ score_i \qquad (2)$$

$$S_r = \begin{cases} P_r\ if\ P_s > N_s \\ N_r\ if\ P_s < N_s \end{cases} \qquad (3)$$

The Lexicon method [17] will be calculated the positive and negative terms in each user review on YouTube. The Lexicon-based method needs to be calculated the variance between positive and negative sentiment to obtain the Sentiment review (S_r) in Formula (3). The conditions are if $P_s > N_s$ the review of sentiment is Positive review (P_r) while if$P_s < N_s$ the was obtained Negative review (N_r).

Besides, we also need to search for a brief topic from user reviews on YouTube. The *Latent Dirichlet Allocation* LDA algorithm was modeled in determining the topic due to is easier to interpret. LDA establishes a model as the following command [18].

```
lda_model_tfidf =
gensim.models.LdaMulticore(corpus_tfidf,
num_topics=5, id2word=dictionary, passes=2,
workers=4)for idx, topic in
lda_model_tfidf.print_topics(-1):
    print('Topic: {} Word: {}'.format(idx, topic))
```

D. Testing and Evaluation

The following are the Formula (4) used to determine the performance of each classification modeling [19].

$$Accuracy = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

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

$$F\text{-}Measure = 2.\ \frac{Precision\ .\ Recall}{Precision+\ Recall} \qquad (6)$$

The number of actual positive cases in the data is Condition Positive (*P*). Negative Condition (*N*) is the number of actual negative instances found in the data. Meanwhile, True Positive (*TP*) is a positive instance in the classifier's data correctly labels. *TP* denotes the number of true positives.

IV. Results and Discussion

Clean data was obtained at the text pre-processing stage. Then we conducted a sentiment analysis on user review data from YouTube. Sentiment analysis was employed to look at both positive and negative reviews. Based on the results, we obtained information about the four videos with the most comments, 98% have positive reviews, and 2% have negative as Fig.2.

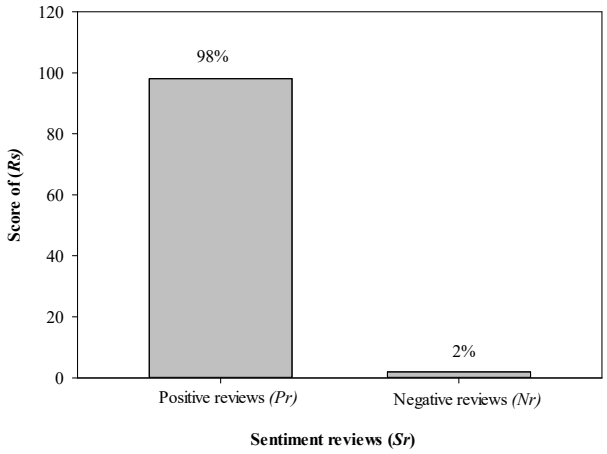

Fig. 2. The Lexicon-based sentiment analysis results toward the Ustadz Abdul Somad lecture videos on YouTube

To evaluate the sentiment analysis results based on the lexicon method, we need to compare it with actual. Based on Formula (4), we have evaluated the results obtained, i.e., *accuracy*, *recall,* and *F-Measure*. Table I shows the evaluation results on three parameters.

TABLE I. The evaluation results of lexicon-based sentiment analysis on three parameters

Evaluation Model	Accuracy	Recall	F-Measure
A comparison of the lexicon-based results and actual	70%	62.5%	76%

Furthermore, we also counted the number of words that appeared after text pre-processing. In general, we found data that are upbeat sounds (English translated) from a total of 2492 words, as shown in Figure 3 below.

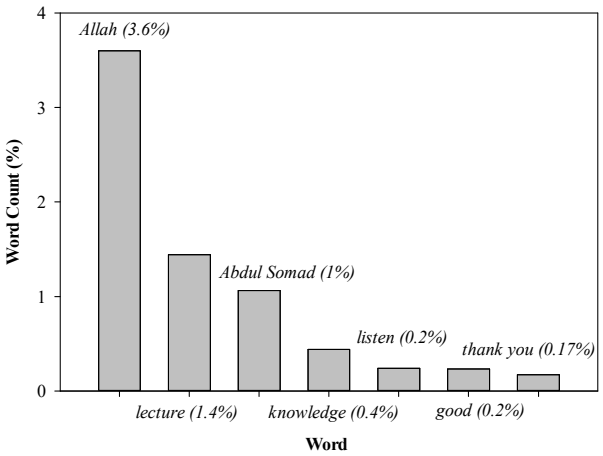

Fig. 3. The most frequent words appeared from a total of 2492 words

In this study, we obtained the topic on YouTube opinion. The *Latent Dirichlet Allocation* (LDA) topic modeling technique was employed to find concise topics in the corpus based on each document's word groups and their respective frequency. Table II shows the top five topics (English translated) were obtained from YouTube opinions.

TABLE II. THE TOP FIVE TOPICS WERE OBTAINED FROM USER REVIEW DATA ON UAS LECTURE VIDEOS

Topic	Topic Keywords
1	*Alhamdulillah, lecture, Indonesia, continues, good*
2	*Healthy, one, Takbir, only*
3	*Ustadz, I, happy*
4	*Present, healthy, always, listening*
5	*lecture, useful, abdul, somad*

V. CONCLUSIONS

There has not been significant research on the use of the Indonesian language in sentiment analysis studies. Furthermore, distinct from Twitter data, the review data on YouTube are not widely used as analysis sources. However, this study analyzed user sentiment towards the video on YouTube using Lexicon. Based on the sentiment analysis results, about 98% of users have positive UAS Lecture Videos reviews. Furthermore, based on the evaluation results, a score of 70% was obtained for accuracy, recall 62.5%, and F-measure 76%. A higher F-measure score indicated that the lexicon-based sentiment analysis could be used in unbalanced class data.

Based on experimental testing, we have observed that the lexicon-based is not robust to analyzing the unstructured sentences, i.e., containing mixed language and questions. However, the use of the lexicon method is a possibility for conducting a sentiment analysis quickly. This method is simple but promising for reaching overall conclusions in sentiment analysis.

Furthermore, the *Latent Dirichlet Allocation* (LDA) summarized the five topics most related to YouTube user reviews. We also counted the number of words that appeared after text pre-processing. Generally, we found seven most words that are positive sound.

A Preacher needs the sentiment analysis results as feedback. The topics modeling technique summarized the user's opinions on YouTube comments to provide helpful input for future lectures. Nevertheless, this research is also a stepping stone for more robust sentiment analysis, especially in text pre-processing and accuracy.

ACKNOWLEDGMENT

Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau and Universiti Kebangsaan Malaysia support this work.

REFERENCES

[1] Statista, "Most used social media 2020 | Statista." https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed Mar. 08, 2021).

[2] C. Dabas, P. Kaur, N. Gulati, and M. Tilak, "Analysis of Comments on Youtube Videos using Hadoop," in *Proceedings of the IEEE International Conference Image Information Processing*, Nov. 2019, vol. 2019-November, pp. 353–358, doi: 10.1109/ICIIP47207.2019.8985907.

[3] Statista, "Indonesia: social media penetration 2019 | Statista." https://www.statista.com/statistics/284437/indonesia-social-network-penetration/ (accessed Mar. 28, 2021).

[4] K. Jamal, R. Kurniawan, A. S. Batubara, M. Z. A. Nazri, F. Lestari, and P. Papilo, "Text Classification on Islamic Jurisprudence using Machine Learning Techniques," in *Journal of Physics: Conference Series*, Jul. 2020, vol. 1566, no. 1, p. 12066, doi: 10.1088/1742-6596/1566/1/012066.

[5] K. Rajab, *Psikoterapi Islam*, 1st ed. Jakarta: Bumi Aksara, 2019.

[6] Z. Wang, V. Joo, C. Tong, and D. Chan, "Issues of social data analytics with a new method for sentiment analysis of social media data," in *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*, Feb. 2015, vol. 2015-February, no. February, pp. 899–904, doi: 10.1109/CloudCom.2014.40.

[7] L. C. Cheng and S. L. Tsai, "Deep learning for automated sentiment analysis of social media," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, Aug. 2019, pp. 1001–1004, doi: 10.1145/3341161.3344821.

[8] R. P. Kusumawardani and M. W. Maulidani, "Aspect-level Sentiment Analysis for Social Media Data in the Political Domain using Hierarchical Attention and Position Embeddings," Aug. 2020, doi: 10.1109/ICoDSA50139.2020.9212883.

[9] Y. J. Su, H. W. Huang, and W. C. Hu, "Using idiomatic expression for Chinese sentiment analysis," Oct. 2017, doi: 10.1109/UMEDIA.2017.8074108.

[10] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model," *IEEE Access*, vol. 8, pp. 138162–138169, 2020, doi: 10.1109/ACCESS.2020.3012595.

[11] Kusrini and M. Mashuri, "Sentiment analysis in twitter using lexicon based and polarity multiplication," in *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, Mar. 2019, pp. 365–368, doi: 10.1109/ICAIIT.2019.8834477.

[12] B. T. Pratama, E. Utami, and A. Sunyoto, "A comparison of the use of several different resources on lexicon based Indonesian sentiment analysis on app review dataset," in *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, Mar. 2019, pp. 282–287, doi: 10.1109/ICAIIT.2019.8834531.

[13] E. W. Pamungkas and D. G. P. Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," in *Proceedings - 2016 6th International Annual Engineering Seminar, InAES 2016*, Jan. 2017, pp. 28–31, doi: 10.1109/INAES.2016.7821901.

[14] Youtube, "Ustadz Abdul Somad Official - YouTube," 2021. https://www.youtube.com/channel/UClvc6c04-xEYKFFyeP3yjKA (accessed Mar. 08, 2021).

[15] I. of F. and A. Linguistics, "UDPipe 1 | ÚFAL." https://ufal.mff.cuni.cz/udpipe/1 (accessed Mar. 18, 2021).

[16] D. S. Lab, "Multilingualsentiment - Data Science Lab." https://sites.google.com/site/datascienceslab/projects/multilingual sentiment (accessed Mar. 18, 2021).

[17] M. Ahmad, M. Ferdy Octaviansyah, A. Kardiana, and K. Fadli Prasetyo, "Sentiment Analysis System of Indonesian Tweets using Lexicon and Naïve Bayes Approach," Oct. 2019, doi: 10.1109/ICIC47613.2019.8985930.

[18] T. Medium, "Topic Modeling and Latent Dirichlet Allocation (LDA) in Python | by Susan Li | Towards Data Science." https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 (accessed Apr. 01, 2021).

[19] Akbarizan, R. Kurniawan, M. Z. A. Nazri, S. N. H. S. Abdullah, S. Murhayati, and Nurcahaya, "Using Bayesian Network for Determining The Recipient of Zakat in BAZNAS Pekanbaru," in *2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI)*, Oct. 2018, pp. 12–17, doi: 10.1109/ICon-EEI.2018.8784142.