Twitter Sentiment Analysis of Juvenile Behaviour Deviations using LSA (Latent Semantic Analysis)

Rochma Muthasima¹, Surya Sumpeno¹ and Yoyon Kusnendar Suprapto¹

¹ Department of Electrical Engineering, Institute Teknologi Sepuluh Nopember Surabaya, Indonesia

Email: rochma16@mhs.ee.its.ac.id

Abstract. Sentiment analysis is an effort to mine data or can be called text mining. One of the uses from sentiment analysis is to analyze public opinion on a particular topic or product. Research related to sentiment analysis has been carried out with various methods and different accuracy results. The approach method in analyzing sentiment is divided into 2, namely with statistical approach and semantic approach. Research with a statistical approach has been developed previously. While the semantic approach is still quite rare. Based on previous study, The LSA (Latent Semantic Analysis) approach seems effective in analyzing data and further more for semantic understanding of sentiment orientation. In this study, we build two model, the first is statistic approach using TF-IDF for feature extraction then classified into SVM (Support Vector Machine) algorithm. Then compare it with semantic approach using TF-IDF and SVD (Singular Value Decomposition) for feature extraction then put into LSA (Latent Semantic Analysis) model. The LSA (Latent Semantic Analysis) model will convert the original feature space to a new lower-dimension space by calculate cosine similarity between sentence vectors with vector topics to produce a score for each sentence by measuring how closely the semantic meaning between the sentence and all the words on the topic. Then divided into three collection of words that sentiment oriented including its vector similarity value. The results of the study prove that the research with statistical approach using SVM (Support Vector Machine) algorithm gives results of precision average about 0.56 and recall about 0.54. While with semantic approach using LSA (Latent Semantic Analysis) gives a slight increase in result of precision about 0.57 and recall about 0.56.

Keywords—text mining, sentiment analysis, SVM (Support Vector Machine), LSA (Latent Semantic Analysis)

1. Introduction

Text mining is the development of data mining. Data consisting of extracted texts from information from one another connected to form a new pattern to be explored further. Text mining is commonly used for the analysis of public sentiments in social media such as Twitter, Facebook, Instagram. For example in business operations such as product analysis, warranty analysis, health record analysis, and others.

Sentiment analysis is also called opinion mining which is a branch of knowledge from text mining in analyzing sentiments and emotions through blog posts, comments, reviews, statuses and tweets that describe community feedback about a particular topic or product. This includes building systems,
collecting data, and testing opinions about certain products. For example in the field of business can help assess the success rate of an advertisement depending on the version of the product or service used. In general, sentiment analysis is divided into 3 levels, including the level of documents, sentences, aspects, and entities [1].

In this study, we will explore data or opinions from one of the most popular social media in the community, namely twitter. Data related to the social field, especially social problems such as deviations from adolescent behavior today. Social problems are problems that arise in society, are social in nature and closely related to social values and community institutions. So basically social problems concern social and moral values. Therefore social problems will not be discussed without considering the community's measurements of what is considered good and what is considered bad. Various kinds of social problems in Indonesia development, among others, are as follows:

1. Educational Problems
2. Problems of Poverty
3. Environmental Problems
4. SARA Conflict Problems
5. Problems of Crime
6. Problems of Protest, Regional Upheaval, and Human Rights Violations
7. Problems with Adolescent Behavior and Juvenile Delinquency

The definition of adolescent deviant behavior is all forms of adolescent behavior that are not in accordance with the values and social norms that apply in society. Among the forms or kinds of deviant behavior of adolescents include:

a. Fighting between students/brawl;
b. Sexual deviations include freex, homosexuality, lesbianism, and premarital sexual relations;
c. Alcoholism;
d. Abuse of drugs or narcotics;
e. Speeding on the highway/illegal street racing;
f. Theft or fraud, and other forms of action
g. other crimes.

Social problems related to adolescent deviant behavior fall into the category of cultural factors, including divorce, juvenile delinquency, etc [2]. So here we explore data sourced from social media twitter using keywords #kenakalanremaja, #tawuran, #seksbebas, #minumankeras, #narkoba, #balapanliar, dan #pencurian. We will analyze the data in the social field using statistical approach by SVM (Support Vector Machine) and semantic approach by LSA (Latent Semantic Analysis) [3]. SVM (Support Vector Machine) algorithm is an algorithm that is often used in classification for twitter data sentiment analysis because it can provide good accuracy results compared to other methods such as Naïve Bayes, Decision Tree, and so on. The SVM (Support Vector Machine) algorithm has the advantage of being able to identify a separate hyperplane which in turn can maximize the margin between 3 different classes or labels. However SVM also has disadvantages, namely the selection of appropriate parameters or features [4].

There are several techniques in modelling topics including LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation). LSA (Latent Semantic Analysis) is a statistical method in extracting contextual features or terms and also representing the meaning of a text through terms consisting of these unique strings [5]. LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation) have the same input which is a collection of words in matrix format. But LSA (Latent Semantic Analysis) focuses on the decomposition of the dimensional matrix while LDA (Latent Dirichlet Allocation) focuses on solving problem modeling topics so that it is suitable for research related to information retrieval. Therefore, in this research related to sentiment analysis we use LSA (Latent Semantic Analysis) with the aim of reducing the value of complexity from the word dimension matrix. LSA (Latent Semantic Analysis) utilizes a mathematical operation SVD (Singular Value Decomposition) to get a smaller collection of features. In LSA (Latent Semantic Analysis), we can get term-by-factor and doc-by-factor matrix. From the matrix can interpret the results based on the values for each document. LSA (Latent Semantic Analysis) can calculate proximity distance or vector values similarity to semantics by calculate cosine similarity between sentence vectors with vector topics to produce a score for each
sentence by measuring how closely the semantic meaning between the sentence and all the words on the topic. The sentence covered by the topic in the document is a sentence that produces the highest cosine similarity. This is done by comparing sentence vectors with topic vectors so it is expected to produce results more effective than statistical approaches.

After that, the precision and recall are calculated. Precision and recall are calculation methods that have often been used to measure performance a system. Precision is the level of accuracy can also be called sensitivity value which is a measurement of how much quality the usefulness of a system from the relevant results obtained by the system. Precision is related to the system's ability not to call documents irrelevant. Whereas recall is the success rate the system can be called a specificity value that displays the quality of how relevant the results are obtained by a system. Recall relates to the system's ability to call relevant documents. Using precision or accuracy alone in measuring the performance of a system/method can cause a very fatal bias a system may still make a mistake or display results that are less than expected even though only 1% however this shows the quality of the work carried out by the system is not yet maximal. if it is associated with a twitter document and the similarity value of one word vector to another then recall describes the relevance of the results of the collection of words obtained from twitter documents with see the value of the proximity or similarity of one word vector to another [6].

2. Related study and scope of work

In the previous study an analysis of twitter data was conducted using the LSA (Latent Semantic Analysis) approach to improve search in the text. Then to evaluate accuracy using precision and recall calculations to get the accuracy of up to 84% [4]. Then there is also research on clustering tweets using LSA (Latent Semantic Analysis) which is twitter data analysis was carried out on the MH370 tragedy. Then tweets are grouped based on the frequency of terms obtained from the document classification process using LSA (Latent Semantic Analysis) so that there are 2 types of tweets that are emotional and non-emotional [7].

Then the research related to twitter data classification using SVM (Support Vector Machine) method entitled "Experimentation System Classification of Twitter Sentiment Analysis on Surabaya City Official Accounts Based on Machine Learning" classifies using the Naive Bayes algorithm and SVM (Support Vector Machine). Data sources were taken from twitter @e100ss and @SapawargaSby during the period 1 September 2015 to 13 October 2015 which were grouped into 3 classes namely positive, negative, and neutral. The best classification model is obtained using the SVM (Support Vector Machine) algorithm with an accuracy of 79.81% [8]. There is also research related to the application of sentiment analysis on Indonesian-language twitter as a rating provider. In this study classification was carried out with SVM (Support Vector Machine) algorithm and stemming from 175,000 tweet data divided into 2 positive and negative classes. The results obtained a fairly good accuracy of 73.43% [9].

The limitation in this study is that this study emphasizes the feature extraction process and not feature selection. Then data retrieval is done through twitter related to the social field, especially the problem of juvenile deviation behavior which consists of juvenile delinquency, street children, brawls, liquor, free sex, wild racing, drugs, theft, and fraud. And also labeling tweets in this study is specifically used for Indonesian sentiment analysis. And the comparison between sentiment analysis using the semantic approach with a statistical approach.

3. Proposed method

The stages or methodology in sentiment analysis using LSA (Latent Semantic Analysis) can be seen in the following figure 3.1. Research Methodology:
The first stage in first model, we collecting corpus data from social media, namely twitter. Then in the twitter crawling process, there were tweets of data about 1800 data in the form of tweets including retweets from twitter users who discussed social issues which included juvenile delinquency, street children, brawls, liquor, free sex, wild racing, drugs, theft, and fraud. Out of 1800 data tweets were filtered again to become 524 data tweets. Data from Twitter consists of posting dates, user names, posting sentences, posting hours, and much more. But here, we will only take tweet sentences as sentiment analysis material. The results of the tweets obtained can be seen in table 3.1. Tweet Collection Results

| No. | Tweet Collection Result |
|-----|-------------------------|
| 1   | Miris dengan kenakalan remaja jaman sekarang |
| 2   | Sakit nih orang2 yg ngasi rokok atau minuman keras ke hewan di kebun binatang..\nDiciduk pihak berwajib pake nangis-nangis bi\u2026 |
| 3   | Keturunan baik-baik itu selalu jauh dari larangan Terhadap tempat hiburan malam |
| 4   | Wahai pemuda pemudi. Janganlah kita meminum minuman keras. karen joins itu adalah minuman para setan. Kalo kita meminum\u2026 https://t.co/EY3oxNASGt |
| 5   | Masa Muda Ingin Bebas.. Terbang Di Angkasa Luas\nJangan Sampai Iman Terlepas.. Terjerumus Minuman Keras\nDunia Tak Kan A\u2026 |
| 6   | Ngasih minuman keras ke binatang kena pasal penganiayaan hewan |

The second stage is the labeling process of tweet data that is carried out through a survey of three people who conduct research in the same field of text mining. In the labeling process, the tweet data is divided into three classes: positive, negative, and neutral. From the survey results, the most subjectivity was taken from the three people to be used as target of training data and can be seen in table 3.2. Manual Labeling Results
Table 3.2. Manual Labeling Results

| No. | Manual Labeling Results                          |        |
|-----|--------------------------------------------------|--------|
| 1   | Miris dengan kenakalan remaja jaman sekarang    | negative |
| 2   | RT @uutrahmat: Sakit nih orang2 yg ngasi rokok atau minuman keras ke hewan di kebun binatang..\nDiciduk pihak berwajib pake nangis-nangis bi\'u2026 | negative |
| 3   | Keturunan baik-baik itu selalu jauh dari larangan Terhadap tempat hiburan malam | netral |
| 4   | Wahai pemuda pemudi. Janganlah kita meminum minuman keras. Karena itu adalah minuman para setan. Kalo kita meminum\'u2026 https://t.co/EY3oxNASGt | netral |
| 5   | RT @PrcsSyahriniNew: Masa Muda Ingin Bebas.. Terbang Di Angkasa Luas\'nJangan Sampai Iman Terlepas.. Terjerumus Minuman Keras\'nDunia Tak Kan \u2026  | netral |
| 6   | Minuman alkohol memaksa ginjal bekerja keras    | netral |

The third stage is the preprocessing stage where twitter data is processed starting from eliminating punctuation, links, hashtags, retweets, and spelling correction. Then go into tokenisasi namely to separate words/sentences, Indonesian stopwords to discard non-essential words or not contain sentiments by using StopWordRemoverFactory from Indonesian Literature. The results can be seen in table 3.3. Preprocessing Results

Table 3.3. Preprocessing Results

| No. | Preprocessing Results                          |
|-----|------------------------------------------------|
| 1   | miris kenakalan remaja jaman sekarang         |
| 2   | sakit orang kasih rokok minuman keras hewan kebun binatang ciduk berwajib nangis |
| 3   | keturunan baik larangan hiburan malam         |
| 4   | pemuda pemudi jangan minuman keras minuman setan |
| 5   | masa muda ingin bebas terbang angkasa luas jangan sampai iman terlepas terjerumus minuman keras dunia |
| 6   | kasih minuman keras binatang pasal penganiayaan hewan |

The fourth stage is feature extraction using TF-IDF. From the preprocessing results, we get a vector that contains features or terms that can be calculated by the weight of occurrence per document. Each different term will be given weight using TF-IDF with equation 3.1:

\[
W_{t,d} = (1 + \log 10 \frac{tf_{t,d}}{n})((\log 10 n) / df_{t})
\]

Where :
- \( n \) = total number of documents
- \( tf_{t,d} \) = total number of term in the document \( d \)
- \( df_{t} \) = total number of document which contains terms in the entire document
TF-IDF will help in converting the unique integer id generated into the matrix by calculating the frequency of occurrence of words in the document where each row represents a unique word, while each column represents the context from which the words taken. After that, twitter data classified using SVM (Support Vector Machine) method and divided into 3 sentiments classes, namely positive, negative, and neutral sentiments.

In the second model, we reduced the TF-IDF matrix using SVD (Singular Value Decomposition) to minimize the value of complexity in processing the term document matrix and decomposed into 3 matrices namely matrix $U$, $D$, transpose matrix $V$ with the equation 3.2 below:

$$A_{mn} = U_{mn} \times S_{mn} \times V_{nn}^T$$

(3.2)

Where:
- $A_{mn}$: beginning matrices
- $U_{mn}$: orthogonal matrix $U$
- $S_{mn}$: diagonal matrix $S$
- $V_{nn}^T$: orthogonal matrix transpose $V$

Then put it into LSA model, in LSA (Latent Semantic Analysis) model we calculate proximity distance or vector values similarity to semantics by calculate cosine similarity between sentence vectors with vector topics to produce a score for each sentence by measuring how closely the semantic meaning between the sentence and all the words on the topic. After that, compared the sentence vectors with vector topics so that the final value will be obtained by taking the highest value of cosine similarity. Cosine similarity calculation can be seen in the equation 3.3 below:

$$\cos \alpha = \frac{A \times B}{|A| \times |B|} = \frac{\sum_{t=1}^{n} A_t \times B_t}{\sqrt{\sum_{t=1}^{n}(A_t^2 \times B_t^2)}}$$

(3.3)

Where:
- $\alpha$: the angle formed between vector $A$ and vector $B$
- $A$: document vector
- $B$: query vector
- $|A|$: length of the vektor $A$
- $|B|$: length of the vektor $B$

In the LSA (Latent Semantic Analysis) process, the data containing the document is reduced by its complexity value to produce a collection of features or words which are the arrangement of documents. This collection of features or words can be grouped into 2 or more based on concepts. The concept here is a grouping of words based on the value of vector proximity or similarity between a feature and other features. So the results obtained are in the form of words and the value of the proximity of the vector. For more details about phases on the LSA process can be seen in figure 3.2. LSA (Latent Semantic Analysis) modeling:
4. Result and Discussions

In the first scenario, we use SVM (Support Vector Machine) algorithm with social field datasets that have been obtained which are 524 tweets. The process is adjusted to the basic algorithm of normal classification using linear kernels and the TF-IDF weighting method with C and gamma values in the default model of the system and random state 0. In labeling 524 data, 1655 features or terms are obtained. The labels that are mostly obtained from data in the social field related to juvenile behavior deviation contain negative sentiments, as well as the terms in them tend to express negative sentiments such as invective words. In the table 4.1 Twitter Data Sentiment we can see the amount of twitter data that has been separated by survey according to its sentiments:
Then the labeled data is processed through the preprocessing stage. Data is divided into 2, namely training data 60 and testing data 40. After that, classification is done with SVM (Support Vector Machine) algorithm and the results of precision average value is 0.56 and recall 0.54. Most recalls are generated on negative sentiment, which is 0.72.

In the second scenario, we use the LSA (Latent Semantic Analysis) model, which is the data that has been converted into the term document matrix is decomposed back into the SVD matrix then search for the angles between sentence vectors and topic vectors between lines with the cosine similarity equation. In cosine similarity calculation n_component is divided into 3 with the assumption that each concept contains negative, positive and neutral sentiments. N_component here determines the concept produced by using LSA (Latent Semantic Analysis). The use of the LSA method produces Bag-of-Words or a set of random terms. For example here using n_component = 3 then the results obtained can be seen in table 4.2. Collection of Words:

### Table 4.1. Twitter Data Sentiment

|          | Positif | Negatif | Netral |
|----------|---------|---------|--------|
|          | 175 tweet | 203 tweet | 146 tweet |

Table 4.1. Twitter Data Sentiment

Then from each document analyzed again such as the first document "miris kenakalan remaja jaman sekarang", words that have vector closeness with the concept 0 are “remaja”, “kenakala”, “jaman”, “sekarang” with vector values 1,313. 1,313 is the sum of the following word [remaja+kenakalan+jaman+sekarang] vector values. Then in the concepts 1 and 2 the results of the vector value are zero in other words there are no words that have a vector value that close to the concept. The results can be seen in table 4.3. Collection of Words on each concept and table 4.4. Total of Vector Values on each concept below

### Table 4.2. Collection of Words

| Concept 0 Vector Value | Concept 1 Vector Value | Concept 2 Vector Value |
|------------------------|------------------------|------------------------|
| Juvenile 0.63          | Racing 0.61             | Drugs 0.37             |
| Delinquency 0.63       | Illegal 0.60            | Sexual 0.33            |
| Drugs 0.12             | Children 0.19           | Abuse 0.33             |
| Children 0.16          | Street 0.14             | Brawl 0.33             |

Table 4.2. Collection of Words

### Table 4.3. Collection of Words on each concept

| Tweets | Concept 0 Related Terms | Concept 1 Related Terms | Concept 2 Related Terms |
|--------|-------------------------|-------------------------|-------------------------|
| Miris kenakalan remaja jaman sekarang | Remaja, kenakalan, jaman, sekarang | - | - |
| Sakit orang kasih rokok minuman keras hewan kebun binatang ciduk berwajib nangis | Orang, keras, minuman, kasih | Keras, minuman | Keras, kasih |
| Keturunan baik larangan hiburan malam | - | Malam | - |
| Pemuda pemudi jangan minum minuman keras minuman setan | Jangan, keras | Keras, jangan, minuman | Keras, jangan, minuman |

Table 4.3. Collection of Words on each concept
After getting the number of vector values, we can determine the largest vector value in what concept and our analysis shows that the most negative and neutral sentiment documents enter concept 2 and the most positive sentiment documents enter concept 0. For positive-sentiment documents we calculate the accuracy value manual and get 48% results. However, for negative and neutral sentiment documents there is no specific range or distance of similarity vector values that can distinguish the two. And the results of precision value is 0.57 and recall 0.56. The recall value of tweets related to juvenile behavior deviation here can be seen from the value of proximity or similarity of vectors from one word to another to a document. Examples can be seen in the following table 4.5. Vector Proximity Value:

### Table 4.3. Collection of Words on each concept

| Tweets | Concept 0 Related Terms | Concept 1 Related Terms | Concept 2 Related Terms |
|--------|-------------------------|-------------------------|-------------------------|
| kasih minuman keras binatang pasal penganiayaan hewan | - | keras | keras, minuman, kasih |
| dekil kayak anak jalanan | anak, jalanan | anak, jalanan | anak, jalanan, kayak |

### Table 4.4. Total of Vector Values on each concept

| Tweets | Vector Values Concept 0 | Vector Values Concept 1 | Vector Values Concept 2 | Max Vector Values Concept |
|--------|-------------------------|-------------------------|-------------------------|---------------------------|
| miris kenakalan remaja jaman sekarang | 1.313 | 0 | 0 | 0 |
| sakit orang kasih rokok minuman keras hewan kebun binatang ciduk bervajib nangis | 0.062 | 0.021 | 0.297 | 2 |
| keturunan baik larangan hiburan malam | 0 | 0.034 | 0 | 1 |
| pemuda pemudi jangan minum minuman keras minuman setan | 0.021 | 0.021 | 0.335 | 2 |
| kasih minuman keras binatang pasal penganiayaan hewan | 0 | 0.021 | 0.196 | 2 |
| dekil kayak anak jalanan | 0.166 | 0.271 | 0.216 | 1 |

### Table 4.5. Vector Proximity Value

| Tweets | Concept 0 Related Terms with Vector Value | Concept 1 Related Terms with Vector Value | Concept 2 Related Terms with Vector Value |
|--------|------------------------------------------|------------------------------------------|------------------------------------------|
| miris kenakalan remaja jaman sekarang | Remaja=0.63 Kenakalan=0.63 Jaman=0.27 Sekarang=0.17 | - | - |
| suka video via vallen cerita anak jalanan | Anak=0.16 Jalanan=0.05 Suka=0.05 Video=0.04 | Anak=0.19 Jalanan=0.14 Suka=0.10 Video=0.07 | Anak=0.22 Jalanan=0.18 Suka=0.09 Video=0.07 |
From the table 3.8. Vector Proximity Value, each document is analyzed each word contained in each concept and then the vector value is calculated. The greater vector value obtained, the more similar to the topic. And the results are a lot of negative documents contained in concept 2 so that concept 2 is considered as a negative sentiment and the concept 0 is considered as a positive sentiment. While for neutral sentiments it was not detected in the LSA modeling process so that in the end we only used 2 classes of sentiment, namely negative sentiment and positive sentiment.

5. Conclusions
From the comparison of the 2 scenarios above, it can be concluded that:
1. The twitter data classification results related to juvenile delinquency deviations using the SVM method which is run in a linear kernel with the default settings divided into 3 sentiment classes namely positive, negative and neutral. And gives precision result of 0.56 and recall of 0.54.
2. Analysis of data using LSA, divide the division of components or concepts into 3 with the assumption that each component contains a collection of words that contain positive, negative, and neutral sentiments. The results obtained by sorting each component or concept containing 100, 300, and 500 features are a collection of random words but when viewed on the value of proximity/similarity of the vector, the most negative and neutral documents are found in concept 2. Then for the most positive documents there are many in the concept 0. And gives precision result of 0.57 and recall of 0.56.
3. For future research, we will explore more about the weight of sentiments through the vector value of the closeness or similarity of words to a document or other words and it is expected that research related to the weight or score of sentiments will grow and develop in order to enrich the word with resource or sentiment dictionary Indonesian.

6. References
[1] Julia Kreutzer & Neele Witte 2014 Opinion Mining Using Sentiwordnet (Uppsala University)
[2] Soerjono Soekanto 2013 Sosiologi Suatu Pengantar
[3] B. Liu 2012 Sentiment Analysis and Opinion Mining Chapter Sentiment Leksikon
[4] Dinar Ajeng Kristiyanti 2015 Analisis Sentimen Review Produk Kosmetik Menggunakan Algoritma Support Vector Machine Dan Particle Swarm Optimization Sebagai Metode Seleksi Fitur
[5] S. G. Chodhary , Rajkumar S. Jagdale , Sachin N. Deshmukh 2016 Semantic Analysis of Tweets using LSA and SVD
[6] Masfulatul Lailiyah 2015 Sentimen Analisis Menggunakan Rule Based Method pada Data Pengaduan Publik Berbasis Lexical Resource (Institut Teknologi Sepuluh Nopember)
[7] Norsuhaili Mahamed Rasidi, Sakhinah Abu Bakar, Fatimah Abdul Razak 2017 Tweets clustering using latent semantic analysis
[8] Nuke Y. A. Faradhillah, Renny P. Kusumawardani, Irmasari Hafidz 2016 Eksperimen Sistem Klasifikasi Sentimen Twitter Pada Akun resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin
[9] Nuvierta Monarizqa, Lukito Edi Nugroho, Bimo Sunarifri Hantono 2014 Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating

Acknowledgement
Acknowledgments for Ministry of Communication and Informatics and Institute of Technology Sepuluh Nopember also Mojokerto City Planning Development Office.