A study of Natural Languange Processing for Information Classification at XYZ University

Wiwin Sry Adinda Banjarnahor[1]
[1]Department of Informatics and Computer Engineering, Politeknik Negeri Medan, Indonesia

ABSTRACT

Undoubtly that social media has been used to exchange idea easier and faster. That increasingly widesprea exchange of data on social media boosts the data growth and generates various structures of data. University ca benefit this large of data by gaining knowledge related to topics discussed by social media users. By gaining th knowledge, university can define which information must be disseminated by analyzing the topic that most users discussed on social media. However, the diversity of data structures spread on social media complicates th process of obtaining this knowledge. This can be overcome by analyzing social media data using the Natur Language Processing method. This study suggests a framework for text classification using topics in the form Indonesia language that is spread on social media. The data analyzed in this research are social media data o Facebook and Twitter at the XYZ University. This analysis was carried out by conducting experiments usin applications from previous research

Keyword : Social Media, Natural Language Processing, , Text Classification, Algoritma

Corresponding Author:
Wiwin Sry Adinda Banjarnahor
Department: Informatics and Computer Engineering
Affiliation: Politeknik Negeri Medan
Address: Jl. Almamater No.1 Kampus USU Padang Bulan Medan, Indonesia
Email : wiwinbanjarnahor@polmed.ac.id

Article history:
Received Jan, 2021
Revised Jan, 2021
Accepted Jan, 2021

1. INTRODUCTION (10 PT)

Nowadays university benefits social media as a media of communication with prospective students and the wider community in general. This should make it easier for university to disseminate information. However the information disseminated by the university was classified as one-sided. The interaction of social media users and university only occurs when there is something that is asked by the user directly to the university's social media account. To gain knowledge of the topics discussed by social media users, university needs to analyse data in social media. The results of this social media data analysis can be used by university to consider the information that will be disseminated on social media. So that university does not only provide information based on one-sided thinking, it also considers the results of the analysis of social media data. It also increase the university's interaction with social media users [1, 2].

However, the data on social media which is the source of extracting information has a large volume. The content that is not structured in terms of form and format such as text, text with numbers and various non-uniform writing will make the analysis process more difficult. The difficulty of analyzing data from social media can be helped by applying Natural Language Processing (NLP) [3]. NLP was chosen because it is able to make computers understand the natural language given and respond to the desired processing results as humans do [4]. NLP tries to make the computer able to understand a command written in everyday language and it is expected that the computer will also respond in a language similar to natural language [5]. So that the analysis of social media data will be easier to do by utilizing the process in NLP.

This research focuses on university case studies in overcoming the problems previously described. Data that has been obtained from university social media can be processed into knowledge by classifying the information discussed by university social media users into relevant topics. The results of this study can

be used to compare the results of information classification with the application of the NLP process algorithm, namely stemming, spelling correction and text classification [6]. The comparison evaluation of this algorithm is seen from the classification of information generated in the current research. The results of this information classification can be a description of various topics that are being discussed or asked by social media users. So that university has a consideration of topics that will be disseminated further at a certain time in a more efficient manner and can help answer the needs of social media users [4].

2. RESEARCH METHOD/ LETERATURE REVIEW

Natural Language Processing is a research and application field that explores how computers understand language as humans do naturally [4]. NLP or Natural Language Processing can be used to process automatically the search for information from human language that previously could only be done by humans. The goal in the field of NLP is to carry out the process of making a computational model of language, so that human-computer interaction occurs with natural language intermediaries [7]. This computational model can be useful for scientific purposes such as examining the properties of a natural form of language as well as for everyday purposes, in this case facilitating communication between humans and computers.

NLP poses a scientific challenge to develop robust methods and algorithms that extract relevant information from large volumes of data originating from multiple sources and languages that have a free or even unstructured structure [8]. So that in this study three processes were applied to NLP, namely the process of stemming, spelling correction and text classification which are expected to support the final objectives of the study (Figure 1. Research Method). Social media data processing is carried out by applying the NLP task to classify documents into certain categories, namely text classification. Before the data is grouped, spelling correction is performed to overcome spelling errors found in the social media data. To perform spelling correction, the word needs to be returned to the original root word by removing the affixes in the word (stemming). However in this article, we only explained the text classification work that had been done in this study. Thus stemming and spelling correction work in this study are out of discussed.

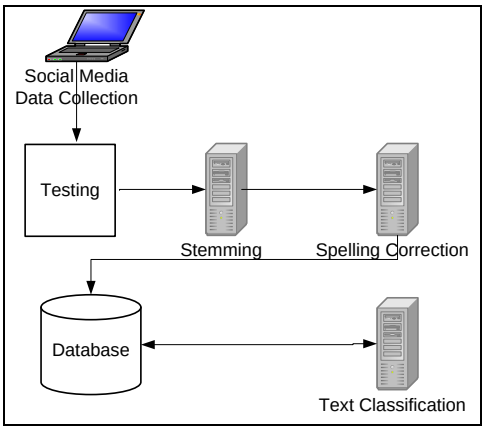

Fig 1. Research Method

Text classification is an activity carried out to find models or functions that explain or differentiate concepts / data classes with the aim of being able to estimate the class of an object [2]. This classification process is carried out automatically to determine or categorize a document into one or more groups based on the contents of the document. This classification process can also be done to categorize a word or sentence into a category. This categorization can be done by applying text classification algorithms such as Naïve Bayes, K-Nearest Neighbor, Support Vector Machines and ID3. The comparison between these algorithms was performed and this study used Naïve Bayes for the text classification process. This is because for data such as large and unstructured social media data, K-Nearest Neighbor and ID3 require high computation costs and long execution times and SVM can only categorize documents into two different categories. The Naïve Bayes algorithm is uncomplicated and effective in text classification.

The Naïve Bayes algorithm is a simple probability classifier that applies the Bayes Theorem. In this Naïve Bayes algorithm, sentences will be categorized into a category that corresponds to that category. The library used in this application was developed by Lukasz Krawczyk and the application can be obtained from https://github.com/LukaszKrawczyk/PHPNaiveBayesClassifier. In the application that has been made, a sentence categorization will be carried out to determine the sentence is included in the predetermined language category. In this study, the library will be developed so that it fits the needs of the study. A library that can handle only three categories before had been developed to be able to handle seven categories according to predefined categories. The dictionary in this library will also be changed so that it can handle sentences according to predetermined categories. The predefined categories based on the XYZ University needs are “Pendaftaran”, “Kegiatan”, “Beasiswa”, “Prestasi”, “Program Studi”, “Informasi Umum”, dan “Kategori Lain-lain”.

The steps for classifying the text are as follows:

1. Generate a dictionary for each category
   A dictionary was made of as many categories as the predetermined categories. In this case we classified the text into 7 categories. Therefore a file is created for each category such as registration.txt containing words related to registration and admission of new students such as word lists, registration, exams, tests, psychotest, usm, and so on, event.txt contains words related to activities such as the word event, activities, etc., scholarship.txt contains words related to scholarships, achievement.txt contains words related to championships or achievements achieved by PT, prodi.txt contains words related to PT study programs, general.txt contains words related to general information such as the location of PT, other. txt. contains words related to other things that are not included in the category such as the words solid, great, successful, and so on. Each file contains 60 words each according to the category. The category dictionary can be seen in Appendix A.
2. Perform word retrieval from the database
   Fetching words from the dictionary is handled in the example.php file. The word taken from the dictionary has passed the stemming application and the spelling correction application so that the word classified is free from errors and punctuation marks. The word classified is "when is the del entrance exam list".
3. Checking the dictionary.
   Before checking the dictionary that has been made beforehand, the classified word is broken down into individual syllables such as "when", "list", "exam", "enter", "del". Furthermore, the calculation of the number of words in the dictionary that has been made previously is carried out. After calculating the number of words in the dictionary, checking for each word is carried out to determine the probability for each word.
3. Determine the probability for each word
   The Naïve Bayes algorithm then calculates each word by calculating the probability of each word entered such as "when", "list", "test", "entry", "del" with each word in the dictionary. Each word will have its own probability value.
4. Determine the probability for each category
   After the text classification application using the Naïve Bayes algorithm determines the probability for each word, the probability value of each input word compared to the word in the dictionary will be used to determine the category probability value. Each category will have its own probability value. The more words in the dictionary that have similarities with the input word, the higher the probability value of that category.
5. Determine the classification results with the highest category probability
   The category with the highest probability value will be selected into the appropriate category according to the input. For example, the word "when is the del entrance exam list" will fall into the registration category. The process is complete.

4. RESULTS AND DISCUSSION

Text classification in this Final Project aims to classify each data on XYZ University social media into topics in accordance with the topic categories that have been determined at the analysis stage. The text classification process begins with data collection from the database which is the result of spelling correction. Each word in the data that has been retrieved is then compared into a dictionary to classify

the data into a category. The classification is selected based on the value of the posterior calculation which is the highest for each category. The more words in the sentence that match the words that appear in the category dictionary, the posterior value will be higher. The following is an example of the data to be classified and the expected classification results:

Table.1 An example of the data to be classified and the expected classification results

ID	Message	Harapan
D61	Min saya mau nanyak . verifikasi pembayaran saya sudh di konfirmasi. Dlm konfirmasi ituu ada tertuliss kode\, yg katanyaa diinformasikan ke panitia saat ujian. Tpi di website pdftaran stlah verifikasi pembayaran\, cetak kartu ujian. Tpi ga bsa min. Ga ad ktrangan utk mencetak. Gmna ya ? Tolong ya min. Terimakasih	pendaftaran
D100	kpn seleksi thp 2 bpk/ibu. syukur ada sma n 1 uluan msk.	pendaftaran
D107	Simak siaran ulang #MoraInteraktif brsama Bpk. Luhut Panjaitan dn di Radio Mora Nusantara. #TuneIn: Mora Sumut dan Mora Jabar	event

There are three combinations of processes used in this Final Project. The difference between this combination is on the algorithm used on spelling correction process. The text classification process begins with data collection from the database which is the result of spelling correction. Each word in the data that has been retrieved is then compared into a dictionary to classify the data into a category. The classification is selected based on the value of the posterior calculation which is the highest for each category. The more words in the sentence that match the words that appear in the category dictionary, the posterior value will be higher. The following are the results of the text data classification in Table 2.

Table 2. Using the data from the spelling correction results.

ID	Expected Category	1st Combination	2nd Combination	3rd Combination
61	Pendaftaran	Pendaftaran	Pendaftaran	Pendaftaran
100	Pendaftaran	Other	Other	Pendaftaran
107	Event	Other	Other	Prestasi

On data ID 61, the three process combinations yield the same classification, namely "Registration". And this result is in the expected category. This can happen because in sentences that have gone through the process of stemming and spelling correction, there are words that fit the "Registration" category such as the words "verification", "pay", "card", "information", "confirmation", "test" , "Write" and "committee". The suitability of the words in this sentence makes the posterior score of the "Registration" category higher than the other categories, namely 3.4516622514195E-140 in combination 1, 2.961526211718E-137 in combination 2 and 7.0265981546755E-140 in combination 3.

From the results of experiments conducted on 200 social media data obtained from Facebook and Twitter, the percentage of classification conformity with expectations in combination 1 is 72.5% with 153 data whose results are as expected and 7 data that are not suitable. with what is expected and in combination 2 is 71% with 152 data whose results are as expected and 48 data that are not as expected. Whereas in combination 3 it is 81% because only 162 data results are as expected and 38 data are not as expected. Based on these results, the combination of processes that produces the highest percentage for classification in accordance with the expected category is combination 3. These results support the conclusion of a study conducted by U.D, Sutisna with the title "Correction of Indonesian Language Query Spelling Using Damerau Levenshtein's Algorithm". In this study, it is stated that the

correction performance with the Damerau Levenshtein algorithm is better and optimal than the correction performance with the Levenshtein algorithm [9]. So that the results of using the Dmerau Levensthein algorithm will support the classification of sentences in a more precise category.

5. CONCLUSION

The experiment used in this study has carried out an analysis process based on the NLP stages on XYZ University social media data, namely preprocessing and stemming, spelling correction and text classification. The combination of processes with the best classification or most suitable for manual examination is the combination of the third process combination, namely the stemming + spelling correction algorithm of Damerau Leventshein Distance + text classification. This combination can be applied to build a recommendation system in future studies that apply a process like this study. Further research could adding university-related vocabulary or correct text according to Indonesian rules into the dictionary. So that the results of correcting the spelling of the wrong word can produce a better word.

REFERENCES

[1] Boyd & Ellison. 2008. Social Network Sites: Definition, History & Scholarship in Journal of Computer- Mediated Communication 13:210-230.
[2] Wahyudi, Adhie Tri and Anita Indrasari. 2009. Transformasi Media Sosial Pemasaran Online dalam Pembentukan Brand Perguruan Tinggi. Proceeding: 17 Oktober 2009 – ISBN 978 979 98125-2-0, Surakarta.
[3] Natural Language Processing 2004, 8 Lectures Ann Copestake (aac@cl.cam.ac.uk) http://www.cl.cam.ac.uk/users/aac/.
[4] Jurafsky, Dan. 2015. Stanford | Natural Language Processing - Course Introduction. (Online). https://class.coursera.org/nlp/lecture/124.
[5] Jurafsky, Daniel and James H. Martin. 2015. Speech and Language Processing, Classification: Naive Bayes, Logistic Regression, Sentiment Analysis. Draft Version April 20, 2014.
[6] Baran, Roger. 2010. Customer Relationship Management and the Asian Company Experience. AFBE Journal (December 2010), Vol 3, No. 2 ISSN: 2071-7873 p. 296.
[7] Vega, V. B., 2001. Information Retrieval for the Indonesian Language, Master's thesis, National University of Singapore.
[8] Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
[9] Sutisna, U. D. 2010. Koreksi Ejaan Query Bahasa Indonesia Menggunakan Algoritme Damerau Levenshtein. Jurnal Ilmiah Ilmu Komputer, 25-29. Fakultas MIPA, Institut Pertanian Bogor.

BIOGRAPHIES OF AUTHORS

	Wiwin Sry Adinda is a dedicated and knowledgeable professional having eight years progressive experience as a lecturer. She is also a skilled and experienced software engineer with over five years of Information Technology project in analysis, development, testing, and maintenance of large scale systems. She is currently a teachin member at Politeknik Negeri Medan. Her research area is focused on data mining, natural language processing, enterprise integration system, software engineering, business process management and knowledge management.