SENTIMENT ANALYSIS ON CLOSURE OF ILLEGAL MOVIE STREAMING SITES USING NAÏVE BAYES ALGORITHM

Dinda Ayu Muthia

Information System
Universitas Bina Sarana Informatika, Jakarta, Indonesia
www.bsi.ac.id
dinda.dam@bsi.ac.id

Abstract— The closure of illegal movie streaming sites IndoXXI has been a trending topic on Twitter at the end of 2019. The reaction of netizens on Twitter shows positive and negative sentiments. Until now, there have been many studies in the field of Sentiment Analysis using data in the form of Tweets from Twitter users. In sentiment analysis research, there are so many method used, and Naïve Bayes is one of it, because it is very simple and efficient. The method has advantages and disadvantages. Naïve Bayes is so sensitive in feature selection. Too many features not only increase calculation time but also reduce classification accuracy. In order to solve the disadvantages and increase the performance of the Naïve Bayes classifier, this method often being combined with many kind of feature selection methods. This research aims to classify tweets into positive and negative using the Naïve Bayes classifier combined with the Genetic Algorithm. The accuracy of Naïve Bayes before using the combination of feature selection methods reaches 79.55%. While after using feature selection methods, which is the Genetic Algorithm, accuracy increased up to 88.64%. The accuracy improved by up to 9.09%.

Keywords: Naïve Bayes, Sentiment Analysis, Twitter.

Abstrak— *Penutupan situs streaming film ilegal IndoXXI telah menjadi trending topic di Twitter pada akhir 2019. Reaksi netizen di Twitter menunjukkan sentimen positif dan negatif. Sampai sekarang, ada banyak penelitian di bidang Analisis Sentimen menggunakan data dalam bentuk Tweet dari pengguna Twitter. Dalam penelitian analisis sentimen, ada begitu banyak metode yang digunakan, dan Naïve Bayes adalah salah satunya, karena sangat sederhana dan efisien. Metode ini memiliki kelebihan dan kekurangan. Naïve Bayes sangat sensitif dalam pemilihan fitur. Terlalu banyak fitur tidak hanya menambah waktu perhitungan tetapi juga mengurangi akurasi klasifikasi. Untuk mengatasi kekurangan dan meningkatkan kinerja pengklasifikasi Naïve Bayes, metode ini sering dikombinasikan dengan berbagai jenis metode pemilihan fitur. Penelitian ini bertujuan untuk mengklasifikasikan tweet-tweet menjadi positif dan negatif menggunakan pengklasifikasi Naïve Bayes yang dikombinasikan dengan Genetic Algorithm. Keakuratan Naïve Bayes sebelum menggunakan kombinasi metode pemilihan fitur mencapai 79,55%. Sementara setelah menggunakan metode pemilihan fitur, yaitu Genetic Algorithm, akurasi meningkat hingga 88,64%. Peningkatan akurasi hingga 9,09%.*

Kata Kunci*: Naïve Bayes, Sentiment Analysis, Twitter.*

INTRODUCTION

The administrator of movie streaming sites through an unofficial distribution channel IndoXXI announced it will close the site starting January 1, 2020 (Ramadhani, 2019). Twitter's timeline on January 24, 2019 was bustling with the hashtag #IndoXXI because netizens talked about closing the illegal movie streaming site. The reaction of netizens on Twitter shows positive and negative sentiments. Until now, there have been many studies in the field of Sentiment Analysis using data in the form of Tweets from Twitter users, including research to classify sentiments from film reviews whose data are taken from Twitter (Samad, Basari, Hussin, Pramudya, & Zeniarja, 2013), research for analyzing unstructured text content on Facebook and Twitter, case studies in the Pizza industry (He, Zha, & Li, 2013), research that addresses sentiment analysis based on Tweets about airlines (Mostafa, 2013), research about scoring sentiment Tweets (Kumar & Sebastian, 2012), and research with three class classification including positive, negative, and neutral (Passonneau, 2011).

In sentiment analysis research, there are so many method used, and Naïve Bayes is one of it, because it is very simple and efficient (Muthia,

2014b). The method has advantages and disadvantages. According to Chen et al. on (Muthia, 2014a), Naïve Bayes is so sensitive in feature selection. Too many features not only increase calculation time but also reduce classification accuracy (Uysal & Gunal, 2012).

In order to solve the disadvantages and increase the performance of Naïve Bayes classifier, this method often being combined with many kind of feature selection methods. This research aims to classify tweets into positive and negative using Naïve Bayes classifier combined with Genetic Algorithm.

MATERIAL AND METHOD

This research uses Naïve Bayes as classification method. But it has disadvantage, so it is combined with feature selection method, namely Genetic Algorithm to increase the accuracy of Naïve Bayes. The dataset used in this research are collection of Tweets from Twitter, consist of 110 pro tweets and 110 contra tweets about closure of illegal movie streaming site IndoXXI. Preprocessing is performed, consist of tokenization and generate N-grams (2-grams). Feature selection method used is genetic algorithm, whereas the classifier which is used is Naive Bayes. 10 fold cross validation testing will be performed, the accuracy of the algorithm will be measured using the confusion matrix and the processed data in the form of ROC curves. RapidMiner Studio 7.2 is used as a tool to measure the accuracy of experimental data.

RESULTS AND DISCUSSION

A. Result

1. Preprocessing

After collecting data from Twitter and split it into Positive (for pro tweets) and Negative (for contra tweets), next step is preprocessing which consist of tokenization and generate N-grams (2-grams). Tokenization is used for removing punctuation mark. Generate 2-grams is used for combining the previous and the next word, because the combination of two words might be create a sentiment word. Initial data processing result can be seen in Table 1.

Table 1. Initial Data Processing Result

Tweet	Tokenization	Generate 2-grams
harusnya kita dukung loh..buat film modalnya g dikit..kita dengan gampangnya streaminga n dan donlod gratis..sama aja kek pencuri tau..dosa..	harusnya kita dukung loh buat film modalnya g dikit kita dengan gampangnya streaminga n dan donlod gratis sama aja kek pencuri tau dosa	harusnya harusnya_kita kita kita_dukung dukung dukung_loh loh loh_buat buat buat_film film film_modalnya modalnya modalnya_g g g_dikit dikit dikit_kita kita kita_dengan dengan dengan_gampangnya gampangnya gampangnya_streami ngan streamingan streamingan_dan dan dan_donlod donlod donlod_gratis gratis gratis_sama sama sama_aja aja aja_kek kek kek_pencuri pencuri pencuri_tau tau tau_dosa dosa
rasakan kalian sobat tak modal	rasakan kalian sobat tak modal	rasakan rasakan_kalian kalian kalian_sobat sobat sobat_tak tak tak_modal modal
Habus aja pak lagian daerah rumahku ga ada sinyal buat nonton film di indoxxi	Habus aja pak lagian daerah rumahku ga ada sinyal buat nonton film di indoxxi	Habus Habus_aja aja aja_pak pak pak_lagian lagian lagian_daerah daerah daerah_rumahku rumahku rumahku_ga ga ga_ada ada ada_sinyal sinyal sinyal_buat buat buat_nonton nonton nonton_film film film_di di di_indoxxi indoxxi
Ya baguslah. Jangan ngebajak mulu. Kasian yg susah2 bikin film.	Ya baguslah Jangan ngebajak mulu Kasian yg susah bikin film	Ya Ya_baguslah baguslah baguslah_Jangan Jangan Jangan_ngebajak ngebajak ngebajak_mulu mulu mulu_Kasian Kasian Kasian_yg yg yg_susah susah susah_bikin bikin bikin_film film
alhamdulillah, ada titik terang untuk menghargai suatu karya	alhamdulillah ada titik terang untuk menghargai suatu karya	alhamdulillah alhamdulillah_ada ada ada_titik titik titik_terang terang terang_untuk untuk untuk_menghargai menghargai

		menghargai_suatu suatu suatu_karya karya	, untuk nonton legal ataupun non legal itu keputusan orang. Tapi lebih baik nya, nonton yang legal. Nah, masalahnya yg nonton ilegal dan bangga ini keliatan tolol dan bodoh nya. Kalau mau nonton ilegal, lebih baik nonton aja. Gk usah banyak tingkah, jadi keliatan bodoh	a untuk nonton legal ataupun non legal itu keputusan orang Tapi lebih baik nya nonton yang legal Nah masalahny a yg nonton ilegal dan bangga ini keliatan tolol dan bodoh nya Kalau mau nonton ilegal lebih baik nonton aja Gk usah banyak tingkah jadi keliatan bodoh	Sebenarnya_untuk untuk untuk_nonton nonton nonton_legal legal legal_ataupun ataupun ataupun_non non non_legal legal legal_itu itu itu_keputusan keputusan keputusan_orang orang orang_Tapi Tapi Tapi_lebih lebih lebih_baik baik baik_nya nya nya_nonton nonton nonton_yang yang yang_legal legal legal_Nah Nah Nah_masalahnya masalahnya masalahnya_yg yg yg_nonton nonton nonton_ilegal ilegal ilegal_dan dan dan_bangga bangga bangga_ini ini ini_keliatan keliatan keliatan_tolol tolol tolol_dan dan dan_bodoh bodoh bodoh_nya nya nya_Kalau Kalau Kalau_mau mau mau_nonton nonton nonton_ilegal ilegal ilegal_lebih lebih lebih_baik baik baik_nonton nonton nonton_aja aja aja_Gk Gk Gk_usah usah usah_banyak banyak banyak_tingkah tingkah tingkah_jadi jadi jadi_keliatan keliatan keliatan_bodoh bodoh
Kalo bisa objektif dong jngan hanya memikirkan tentang Kita, tapi bagaimana mereka yg berkarya	Kalo bisa objektif dong jngan hanya memikirkan tentang Kita tapi bagaimana mereka yg berkarya	Kalo Kalo_bisa bisa bisa_objektif objektif objektif_dong dong dong_jngan jngan jngan_hanya hanya hanya_memikirkan memikirkan memikirkan_tentang tentang tentang_Kita Kita Kita_tapi tapi tapi_bagaimana bagaimana bagaimana_mereka mereka mereka_yg yg yg_berkarya berkarya	Dihapus gak papa yg penting kasih situs berbayar buat nonton film film indo kita, gak papa bayar	Dihapus gak papa yg penting kasih situs berbayar buat nonton film film indo kita gak papa bayar	Dihapus Dihapus_gak gak gak_papa papa papa_yg yg yg_penting penting penting_kasih kasih kasih_situs situs situs_berbayar berbayar berbayar_buat buat buat_nonton nonton nonton_film film
Oh santai, gue juga masih download. Serial TV yang di Amerika, yang di sana free to air dan emang ga ada di streaming legal Tapi kalau film gue udah mengurangi, karena kebanyakan emang nonton di Bioskop.	Oh santai gue juga masih download Serial TV yang di Amerika yang di sana free to air dan emang ga ada di streaming legal Tapi kalau film gue udah mengurangi karena kebanyakan emang nonton di Bioskop	Oh Oh_santai santai santai_gue gue gue_juga juga juga_masih masih masih_download download download_Serial Serial Serial_TV TV TV_yang yang yang_di di di_Amerika Amerika Amerika_yang yang yang_di di di_sana sana sana_free free free_to to to_air air air_dan dan dan_emang emang emang_ga ga ga_ada ada ada_di di di_streaming streaming streaming_legal legal legal_Tapi Tapi Tapi_kalau kalau kalau_film film film_gue gue gue_udah udah udah_mengurangi mengurangi mengurangi_karena karena karena_kebanyakan kebanyakan kebanyakan_emang emang emang_nonton nonton nonton_di di di_Bioskop Bioskop			
Sebenarnya	Sebenarny	Sebenarnya			

yg penting bisa nonton broo	yg penting bisa nonton broo	film_film film film_indo indo indo_kita kita kita_gak gak gak_papa papa papa_bayar bayar bayar_yg yg yg_penting penting penting_bisa bisa bisa_nonton nonton nonton_broo broo

Source: (Muthia, 2020)

2. Classification

This process is to determine whether one word or combination of words is a member of a positive or negative class. After the classification process, it is obtained some words that most frequently appears and related to the pro on contra, such as **"setuju"**, "**setuju indoxxi"**, "**setuju aja"**, "**setuju kalo"**, "**gua setuju"**, **"dukung"**, "**aku setuju"**, "**gak setuju"**, "**setuju sih"**, **"sedih", "setuju banget", "setuju saja".** Total occurences of sentiment word related to the pro and contra about the closure of illegal movie streaming sites can be seen in Table 2.

Table 2. Total Occurences of Sentiment Word

Words	Total Occurences	Number of Tweets
Setuju	53	49
Setuju IndoXXI	12	12
Setuju aja	6	6
Setuju kalo	6	6
Gua setuju	5	5
Dukung	4	4
Aku setuju	4	4
Gak setuju	4	4
Setuju sih	4	4
Sedih	4	4
Setuju banget	1	1
Setuju saja	1	1

Source: (Muthia, 2020)

The main output of classification process is the accuracy of Naïve Bayes. This classificaion performed before combining Naïve Bayes with feature selection method, Genetic Algorithm. Naïve Bayes accuracy reaches 79.55%. The result can be seen in Table 3.

Table 3. Confusion Matrix of Naïve Bayes Classification Before Using Feature Selection Method

Akurasi Naive Bayes: 79.55% +/- 8.92% (mikro: 79.55%)			
	True Positive	**True Negative**	**Class precision**
Pred. Positive	104	39	72.73%
Pred. Negative	6	71	92.21%
Class recall	94.55%	64.55%	

Source: (Muthia, 2020)

3. Model Optimization and Experiments on Model's Parameters

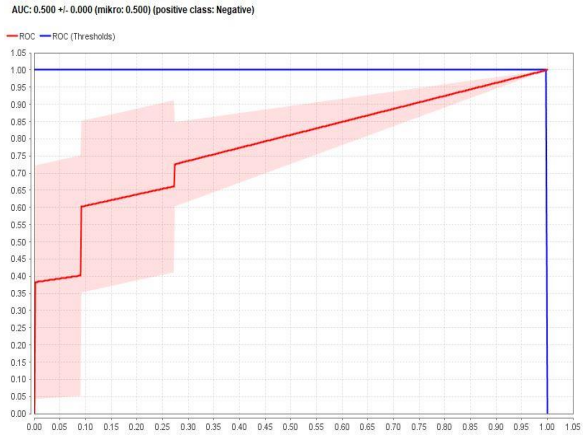
Wrapper feature selection method that is used in this study is Genetic Algorithm. In Genetic Algorithm there are some indicator, such as max number of new parameters, maximal fitness, population size, maximum number of generations, tournament size, start temperature, p initialize, p crossover, and p generate. To get a good model, value of some parameters had been adjusted to obtain high accuracy results. In adjustment indicator on Genetic Algorithm, the highest accuracy is obtained with the combination of population size = 10, p initialize = 0.5, p = 0.5 crossovers, and generate p = 0.1. The results achieve 88.64% accuracy. If other parameters also changed its value, may lead to the data processing becomes increasingly longer. The result can be seen in Table 4.

Table 4. Confusion Matrix of Naïve Bayes Classification After Using Feature Selection Method

Akurasi Naive Bayes: 88.64% +/- 6.18% (mikro: 88.64%)			
	True Positive	**True Negative**	**Class precision**
Pred. Positive	98	12	88.29%
Pred. Negative	12	97	88.99%
Class recall	89.09%	88.18%	

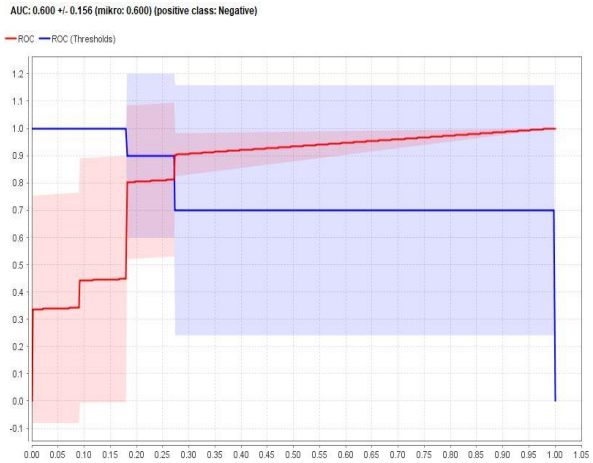
Source: (Muthia, 2020)

Here are some view of ROC curve test results data. Figure 1 is the ROC curve for Naïve Bayes models before using the feature selection methods and Figure 2 is the ROC curve for Naïve Bayes models after using the feature selection methods.



Source: (Muthia, 2020)

Figure 1. ROC curve for Naïve Bayes models Before Using The Feature Selection Methods



Source: (Muthia, 2020)

Figure 2. ROC curve for Naïve Bayes models Before Using The Feature Selection Methods

B. Discussion

In this study, the using of Generate 2-grams also increase the accuracy of the classification. If Generate 3-grams applied, it might get different result and the process could take a longer time. The classification obtained 12 words related to the pro and contra of closure of illegal movie streaming sites, such as, "setuju", "setuju indoxxi", "setuju aja", "setuju kalo", "gua setuju", "dukung", "aku setuju", "gak setuju", "setuju sih", "sedih", "setuju banget", "setuju saja". The other Tweets are more like expression of sadness, goodbye, giving alternative sites to watch movies legally and opinion about piracy.

Without the use of feature selection methods, Naïve Bayes algorithm itself has resulted in an accuracy of 79.55%. Accuracy is still less accurate, so it needs to be improved using feature selection methods. After using the feature selection method of the filter and wrapper are combined, Naïve Bayes algorithm accuracy increases to 88.64%. If the accuracy of classification reaches 70-80%, it is still categorized as "Fair Classification". If the accuracy of classification reaches 80-90%, it is still categorized as "Good Classification". This study obtained classification accuracy up to 88.64%, so for this case Naïve Bayes has a good classification in classifying Tweets.

CONCLUSION

From data processing which has been done, the combination of Naïve Bayes algorithm and Genetic Algorithm as feature selection methods obtained a good result, it is proven it can increase classification accuracy of Naïve Bayes. Data Tweets from Twitter can be classified well into the form of positive and negative. The accuracy of Naïve Bayes before using the combination of feature selection methods reaches 79.55%. While after using feature selection methods, which is Genetic Algorithm, accuracy increased up to 88.64%. The accuracy improvement up to 9.09%.

The model that has been built could be implemented to all text, so that we could see the result directly in the form of positive and negative. In the future research, this model can be applied in other domain and add a new class to be classified a neutral expression, beside positive (pro) and negative (contra).

REFERENCES

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Kumar, A., & Sebastian, T. M. (2012). Sentiment Analysis on Twitter. *IJCSI International Journal of Computer Science*, *9*(4), 372–378

.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251. https://doi.org/10.1016/j.eswa.2013.01.019

Muthia, D. A. (2014a). Analisis Sentimen Pada

Review Buku Menggunakan Algoritma Naive Bayes. *Jurnal Paradigma*, *XVI*(1), 1–9.

Muthia, D. A. (2014b). Sentiment Analysis of Hotel Review Using Naive Bayes Algorithm and Integration of Information Gain And Genetic Algorithm as Feature Selection Methods, 25–30.

Muthia, D. A. (2020). *Laporan Akhir Penelitian: Sentiment Analysis Of Twitter On Closure Of Illegal Movie Streaming Sites*.

Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Proceedings Ofthe Workshop on Language in Social Media*, (June), 30–38.

Ramadhani, Y. (2019). IndoXXI akan Tutup 1 Januari 2020 & Penyebabnya dari Kominfo. Retrieved May 7, 2020, from https://tirto.id/indoxxi-akan-tutup-1-januari-2020-penyebabnya-dari-kominfo-eoGz

Samad, A., Basari, H., Hussin, B., Pramudya, I. G., & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, *53*, 453–462. https://doi.org/10.1016/j.proeng.2013.02.059

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, *36*, 226–235. https://doi.org/10.1016/j.knosys.2012.06.005