TOPIC DISCOVERY AND CLASSIFICATION COMPARISON ON THE COMMENTS OF INDONESIAN ENTERTAINMENT YOUTUBE CHANNEL VIDEOS USING SMOTE, N-GRAM, AND LDA APPROACHES

TUGAS AKHIR

Annisa Rizki Liliandari
41517010006

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2021**

https://lib.mercubuana.ac.id

TOPIC DISCOVERY AND CLASSIFICATION COMPARISON ON THE COMMENTS OF INDONESIAN ENTERTAINMENT YOUTUBE CHANNEL VIDEOS USING SMOTE, N-GRAM, AND LDA APPROACHES

Tugas Akhir

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh:
Annisa Rizki Liliandari
41517010006

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2021

https://lib.mercubuana.ac.id

LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM	: 41517010006
Nama	: Annisa Rizki Liliandari
Judul Tugas Akhir	: Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan didalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 11 Januari 2021

METERAI TEMPEL
6000
ENAM RIBU RUPIAH

Annisa Rizki Liliandari

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Annisa Rizki Liliandari
NIM	:	41517010006
Judul Tugas Akhir	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul diatas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Mercu Buana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 11 Januari 2021

METERAI TEMPEL
6000
ENAM RIBU RUPIAH

Annisa Rizki Liliandari

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Annisa Rizki Liliandari
NIM	:	41517010006
Judul Tugas Akhir	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis		Status	
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi		Diajukan	√
		Jurnal Nasional Terakreditasi			
		Jurnal International Tidak Bereputasi		Diterima	
		Jurnal International Bereputasi	√		
	Disubmit/dipublikasikan di :	Nama Jurnal	: The International Arab Journal of Information Technology (IAJIT)		
		ISSN	: 1683-3198 (print) 2309-4524 (Online)		
		Link Jurnal	: https://iajit.org/		
		Link File Jurnal Jika Sudah di Publish	:		

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

UNIVERSITAS
MERCU BUANA

Demikian pernyataan ini saya buat dengan sebenarnya.

Mengetahui
Dosen Pembimbing TA

Jakarta, 11 Februari 2021

METERAI TEMPEL
6000
ENAM RIBU RUPIAH

Dr. Ida Nurhaida, MT

Annisa Rizki Liliandari

https://lib.mercubuana.ac.id

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41517010006
Nama	:	Annisa Rizki Liliandari
Judul Tugas Akhir	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 09 Februari 2021

(Hery Derajad Wijaya, S.Kom, MM)
Penguji 1

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

PERSETUJUAN PENGUJI

NIM	:	41517010006
Nama	:	Annisa Rizki Liliandari
Judul Tugas Akhir	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 09 Februari 2021

(Dr. Harwikarya,MT)
Penguji 2

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

PERSETUJUAN PENGUJI

NIM	:	41517010006
Nama	:	Annisa Rizki Liliandari
Judul Tugas Akhir	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 09 Februari 2021

(Sri Dianing Asri, ST, M.Kom)
Penguji 3

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

LEMBAR PENGESAHAN

NIM : 41517010006

Nama : Annisa Rizki Liliandari

Judul Tugas Akhir : Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 09 Februari 2021

Menyetujui,

(Dr. Ida Nurhaida, MT)
Dosen Pembimbing

Mengetahui,

(Diky Firdaus, S.Kom, MM)
Koord. Tugas Akhir Teknik Informatika

(Desi Ramayanti, S.Kom, MT)
Ka. Prodi Teknik Informatika

UNIVERSITAS
MERCU BUANA

ABSTRAK

Nama	:	Annisa Rizki Liliandari
NIM	:	41517010006
Pembimbing TA	:	Dr. Ida Nurhaida, MT
Judul	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

YouTube saat ini menjadi media sosial paling populer dengan 88% pengguna aktif mengaksesnya. Komentar yang berisi opini dan saran semakin meningkat dan menantang untuk ditafsirkan satu per satu. Penelitian ini difokuskan kepada klasifikasi teks dan pemodelan topik terhadap analisis data komentar YouTube terkait konten video channel YouTube hiburan di Indonesia yang dilakukan dengan mengaplikasikan metode klasifikasi data mining untuk membandingkan kinerja metode Multinomial Naïve Bayes, K-Nearest Neighbor, dan Support Vector Machine serta mengetahui pengaruh dari berbagai eksperimen untuk menemukan metode yang memiliki akurasi tertinggi dengan prediksi tepat dalam mengklasifikasikan teks sebagai komentar positif, negatif, atau netral. Sedangkan untuk proses pemodelan topik menggunakan Latent Dirichlet Allocation. Kesimpulannya adalah preprocessing yang lengkap, penerapan teknik SMOTE, penyetelan parameter dan fitur-fitur canggih N-gram berkontribusi pada peningkatan akurasi. Hasil penelitian menunjukkan bahwa tingkat akurasi terbaik diperoleh dari model yang menerapkan teknik SMOTE dengan proporsi 80% data latih dan 20% data uji. Model SVM+SMOTE lebih unggul dibandingkan model MNB+SMOTE dan K-NN+SMOTE yaitu dengan akurasi 97,2% untuk dataset 1, 96,1% untuk dataset 2 dan 96,3% untuk dataset 3. Pemodelan topik menunjukkan bahwa dua dari ketiga dataset memiliki persamaan topik dalam penyajiannya.

UNIVERSITAS
MERCU BUANA

Kata kunci:
YouTube komentar, klasifikasi teks, topik modeling, machine learning, smote, n-gram.

https://lib.mercubuana.ac.id

ABSTRACT

Name	:	Annisa Rizki Liliandari
Student Number	:	41517010006
Counsellor	:	Dr. Ida Nurhaida, MT
Title	:	Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

YouTube is currently the most popular social media platform, with 88% of active users having easy access to it. Comments containing opinions and suggestions are increasing, and have become challenging to be interpreted individually. This research specifies on the data analysis of text classification and topic modeling of YouTube comments, related to entertainment video contents in Indonesia. This was carried out by applying the data mining classification method, to compare the performance of the Multinomial Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine techniques, and also ascertaining the effect of various experiments, in locating the accurate model for classifying text as positive, negative, or neutral comments. However, the topic modeling process uses Latent Dirichlet Allocation. In conclusion, the complete preprocessing, SMOTE technique application, parameter setting, and N-gram advanced features, contribute to improving accuracy. The results showed that the best level of accuracy, was obtained from a model that applied the SMOTE technique, with a proportion of 80% training data, and 20% testing data. Therefore, the SVM + SMOTE model is superior to the MNB + SMOTE and K-NN + SMOTE techniques, with an accuracy of 97.2% (dataset 1), 96.1% (dataset 2), and 96.3% (dataset 3). The topic modeling shows that two of the three datasets, have the same topic in the content presentation.

Key words:
YouTube commentary, text classification, topic modeling, machine learning, smote, n-gram.

UNIVERSITAS
MERCU BUANA

KATA PENGANTAR

Assalamu'alaikum Wr.Wb

Puji syukur saya panjatkan kehadirat Allah SWT, yang senantiasa telah melimpahkan rahmat, hidayah dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan Tugas Akhir yang berjudul "Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches" ini dengan baik dan tepat pada waktunya. Tujuan dari penulisan laporan Tugas Akhir ini dibuat sebagai salah satu syarat untuk dinyatakan LULUS sebagai sarjana Ilmu Komputer dari Universitas Mercu Buana.

Saya menyadari bahwa penyusunan Tugas Akhir ini tidak lepas tanpa bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Kedua orang tua Ayah dan Ibu, serta Kakak dan Adik yang tidak pernah lelah untuk senantiasa memberikan motivasi, semangat, doa, dukungan dan kepercayaan kepada saya, sehingga dapat menyelesaikan kuliah dengan baik.
2. Bapak Dr. Mujiono Sadikin, MT, selaku Dekan Fakultas Ilmu Komputer, Program Studi Teknik Informatika Universitas Mercu Buana.
3. Ibu Dr. Ida Nurhaida, MT, selaku Pembimbing Tugas Akhir yang selalu semangat memberikan waktunya untuk membimbing, memberi arahan dan memberi saran di setiap proses pembuatan Tugas Akhir ini.
4. Ibu Desi Ramayanti, S.Kom, MT, selaku Kaprodi Teknik Informatika Universitas Mercu Buana.
5. Bapak Drs. Ahmad Kodar, MT, selaku Pembimbing Akademik yang telah membimbing dan membantu dalam memberikan informasi dan penjelasan mengenai proses akademik selama perkuliahan berlangsung.

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

6. Bapak Diky Firdaus, S.Kom, MM, selaku koordinasi Tugas Akhir Fakultas Ilmu Komputer, Program Studi Teknik Informatika Universitas Mercu Buana.
7. Seluruh Dosen Program Studi Teknik Informatika yang telah memberikan ilmu, wawasan, dan pengalaman yang bermanfaat selama perkuliahan berlangsung. Memberi kesempatan untuk belajar, berkarya dan juga berkembang.
8. Seluruh Staff Administrasi dan Tata Usaha yang telah banyak membantu dan memberikan kemudahan atas semua pelayanan dan arahannya.
9. Kepada saudara Nabil Hizbullah yang telah membantu meringankan beban dalam menyelesaikan Tugas Akhir ini, serta selalu memberikan dukungan baik itu doa maupun semangat yang tidak pernah henti.
10. Sahabat-sahabat dan rekan-rekan mahasiswa/i program studi S1 Teknik Informatika khususnya teman-teman seperjuangan angkatan 2017 dan perempuan-perempuan Teknik Informatika atas kekompakan, kebersamaan, pertemananya, saling menyemangati, saling membantu, saling mengingatkan, dan saling mendoakan yang sudah terjalin hampir 4 tahun ini. Saya harap pertemanan ini tidak akan terputus untuk selamanya dan kami dapat menyelesaikan Tugas Akhir ini bersama dengan baik.
11. Keluarga besar Teknik Informatika yang telah memberikan semangat dan motivasi selama menjalani perkuliahan.
12. Semua pihak dan personal yang tidak dapat disebutkan satu per satu yang terlibat dalam pembuatan Tugas Akhir ini sehingga dapat selesai dengan baik.

UNIVERSITAS
MERCU BUANA

Hasil Tugas Akhir ini masih jauh dari sempurna. Masih terdapat kekurangan dalam penelitian, cara penjelasan, dan kekeliruan dalam penulisan. Untuk itu, kritik dan saran dari pembaca sangat dihargai dan diharapkan. Akhir kata, semoga Tugas Akhir ini dapat memberikan manfaat bagi para pembaca.

Jakarta, 11 Januari 2021
Annisa Rizki Liliandari

https://lib.mercubuana.ac.id

DAFTAR ISI

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches

Annisa Rizki Liliandari and Ida Nurhaida
Faculty of Computer Science, University of Mercu Buana, Indonesia

Abstract: *YouTube is currently the most popular social media platform, with 88% of active users having easy access to it. Comments containing opinions and suggestions are increasing, and have become challenging to be interpreted individually. This research specifies on the data analysis of text classification and topic modeling of YouTube comments, related to entertainment video contents in Indonesia. This was carried out by applying the data mining classification method, to compare the performance of the Multinomial Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine techniques, and also ascertaining the effect of various experiments, in locating the accurate model for classifying text as positive, negative, or neutral comments. However, the topic modeling process uses Latent Dirichlet Allocation. In conclusion, the complete preprocessing, SMOTE technique application, parameter setting, and N-gram advanced features, contribute to improving accuracy. The results showed that the best level of accuracy, was obtained from a model that applied the SMOTE technique, with a proportion of 80% training data, and 20% testing data. Therefore, the SVM + SMOTE model is superior to the MNB + SMOTE and K-NN + SMOTE techniques, with an accuracy of 97.2% (dataset 1), 96.1% (dataset 2), and 96.3% (dataset 3). The topic modeling shows that two of the three datasets, have the same topic in the content presentation.*

Keywords: *YouTube commentary, text classification, topic modeling, machine learning, smote, n-gram.*

UNIVERSITAS
MERCU BUANA

1. Introduction

Generally, researches related to sentiment analysis, are found to be more numerous on the social media platform, such as Twitter [1] - [5]. Also, several studies have started focusing on sentiment analysis on YouTube [6] - [15]. Based on the wearesocial.com site in 2020, YouTube is the first in rank of social media platforms often used, with 88% of users situated in Indonesia. It is a source of media information, where each user interacts through the sharing of videos, giving likes or dislikes, adding views, subscribing to a channel, and also providing comments.

The entertainment video contents of YouTube, are mostly favored and interesting to viewers than that of the educational sector. Furthermore, Indonesia's YouTube viewers are dominated by young people averaging 15-30 years, with children starting to enjoy the services. This age group is being categorized as the millennial generation. Explicitly, the millennial generation is more interested in video contents that are entertaining than educational. This phenomenon is an essential problem because, the video content in the entertainment category, tends to lack educational benefits and value. Therefore, it has both positive and negative impacts, that tends to be imitated by the millennial generation. The positive impact obtained, is in the form of acquiring the latest information about holiday destinations, new fashion trends, and many more. Moreover, the negative impact obtained is quite crucial, such as harsh word usage in videos, Western cultural and controversial lifestyle trends, etc. These problems are observed to be handled by sentiment analysis research. This [1] is the process of using text analytics, to obtain multiple data sources from the internet, and various social media platforms falling within the field of Natural Language Processing (NLP). Furthermore, several studies showed that, the level of interaction between content creators and viewers is a significant dimension of popularity on YouTube [15].

This research focuses on text classification and topic modeling on commentary data analysis related to entertainment video contents in Indonesia, by comparing the performances of the Multinomial Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine classification methods, and also measuring the effects of various experiments, to locate accurate techniques, in classifying text as positive, negative, or

https://lib.mercubuana.ac.id

neutral comments. Moreover, the topic modeling process is observed to make use of the Latent Dirichlet Allocation method. This study is expected to help YouTube content creators in Indonesia, to assess the advantages and disadvantages of the various videos uploaded. It is also expected to aid content creators in determining whether the information and presentation delivered, has a good impact on the audience or vice versa.

Specifically, content creators with lot of subscribers tend to receive more comments, causing valuable viewers feedback to be buried deep withiin those providing useless responses. Furthermore, it is used as reference in considering decision-making, as regards video contents being more creative, effective, informative, visible, increasing the number of views and subscriptions, avoiding plagiarism, and preventing millennial generation from showing unfavourable video content.

In literature, related research has previously been conducted on several datasets, such as [16, 17, 18, 19, 20, 21, 22, 23]. This research consists of six parts, namely, Section 1 (Introduction), 2 (Related Works and Theories), 3 (The Dataset), 4 (Methods), section 5 (Experiments and Results), and 6 (Conclusion).

2. Literature Study

2.1 Related Work

This section provides information about sentiment analysis studies, carried out on several datasets. Siti and Agus [16], attempted to remove stopwords, change slang vocabularies, eliminate subject/object type words, and increase the Extratree's classification accuracy quite significantly, between 3% to 3.5%. With the usage of the Unigram and CountVectorizer feature, extraction obtained an accuracy of 88.8%.

Widi and Retno [17] conducted a classification, using NB, SVM, and LR, while applying the SMOTE technique in handling class imbalance, and observed that the method used was quite effective in improving model performance, with an average increase of around 12%. The best model with a g-mean score, was obtained from the LR model (81.65%), followed by SVM (81.55%).

Further, Irma et al. [18] performed a rating prediction on beauty reviews, using MNB and N-gram. It was observed that the model, with the use of full preprocessing and N-gram combination, produced better accuracy, reaching 97% and 96% on tolerance and review sentiment testing, respectively. Therefore, the N-gram here is very influential on the accuracy results.

Yoga and Dhomas [19] conducted a topic modeling analysis on the title of a research in the health sector in Indonesia, and observed that, LDA modeling was confirmed to have the ability to model topics well, in producing the word distribution of Factors, Nutritional Status, Blood Pressure, Hospitals, Pregnant Women, Health Service, Public Health Centres, Extracts Ethanol, Public Health, Hiv-Aids, Diabetes, and Dengue.

Desi and Umniy [20] in their research, classified text as a complaint or not in the marine and fisheries domain, using Random Forest method, with the application of parameter settings. The best performance results for the model were 95%, while using parameters with min_samples_leaf (1), n_estimators (10), min_samples_split values (3), criterion entropy, max_features (3), and max_depth (0).

Mujiono and Fahri [21] predicted potential risks on customer credit datasets, using the C4.5 and NB algorithms. The results showed that the recommendation system, as a characteristic with the C4.5 model was more suiTabel, because it produces an accuracy of 83.33% than NB, which is based on the conditional probability of the input variable, with the accuracy achieved being 80.67%.

However, Anis et al. [22] compared the performance of K-NN with and without SMOTE, in the subjective/objective news classifications. The results showed that, the application of SMOTE, in the classification of news objectivity, resulted in less effective performance at k values 5, 7, and 9. This is shown when the performance of the K-NN algorithm, has an average decrease in the accuracy value of 6.67%. As for the value of k being 1 & 3, the performance of the K-NN algorithm has an average increase of 3.36% accuracy. Therefore, at k = 1 & 3, the accuracy obtained are 87.50% and 85.22%, respectively.

2.2 Learning Algorithm

During the learning phase, feature extraction uses the TF-IDF algorithm, by applying N-gram as selection. The text classification uses three algorithms, consisting of MNB, K-NN, and SVM, while the modeling topic uses the LDA algorithm.

TF-IDF is the weight calculation method, commonly used in information retrieval and text mining. It is used to calculate the weight of each word, and observe how often they appear in a document [1].

UNIVERSITAS MERCU BUANA

N-gram is a set of N-words, appearing in texts, which are very easy to obtain, and represented by means of vectors. In some models, the N for each successive word, is used as a feature. In the bigram model (N = 2), two consecutive words are used as features in the document's vector representation, same as in the trigram model (N = 3), which involves the use of three words. Furthermore, it is obvious that the N-gram feature provides better results, in the accuracy of the sentiment analysis model [16].

MNB is a widely used algorithm for text classification method, based on the assumption that, each document is extracted from a multinomial word distribution, and all conditional features being independent, with a given value of the class variable [24].

Furthermore, SVM acts by locating a hyperplane, which maximizes the distance between classes (margins). In this research, a linear kernel approach was used, with the 3 model classes tested. The research support authorizes the SVM to become multi-class, capable of classifying data into more than two classes. There are two choices of approach for implementing multiclass SVM. The first approach combines several binary SVMs, while the second bind all data from all classes into a form of an optimization problem. In this case, it uses the one-against-one method. This method builds a number of binary SVM models, that compare one class to another. To classify data into k classes, it has to build a number of, $\frac{K(K-1)}{2}$ binary SVM models [25].

K-NN is an algorithm for the classification method of supervised learning, based on distance. It only classifies the label (class) of an object, based on the class which is the majority of K-neighbors, in a set of training data. The number of K-neighbors, determines the performance of the K-NN algorithm. Distance measurement is then calculated using Manhattan method, which calculates the absolute difference between the coordinates of object pairs, with five heterogeneity in the number of neighbours, namely, 1, 3, 5, 7, and 9, in order to locate the best K-NN model [26].

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus, represented as a random mixture of latent topics, and used to ascertain patterns in a document that generate topics [19]. Modeling topics with LDA in this research, is to analyze for similarities related to the presentation of video content among YouTube channel datasets, based on public comments.

3. Dataset

This research collected data from YouTube, using an algorithm developed by the researchers, in the Python programming language. Only three samples of YouTube channel accounts based on the socialblade.com site, which are quite popular in Indonesia were collected, with the most subscribers and entertainment categories, consisting, Atta Halilintar, Baim Paula, and Ricis Official pages. The videos were selected, based on date of upload (January to August 2020), with the highest number of likes, and those with inappropriate titles, possessing large number of reactions. The total number of comments were 5400, with each account consisting 1800 feedbacks, as shown in Tabel 1. The data collected is stored in .txt format.

Tabel 1. Data Source.

Channel	YouTube Id	Like
Atta Halilintar 600/video (dataset 1)	https://www.YouTube.com/watch?v=j1d6rR2Jl7U	396K
	https://www.YouTube.com/watch?v=Ou8Hknbp2Nw	119K
	https://www.YouTube.com/watch?v=gxKN8JSQgT4	255K
Baim Paula 600/video (dataset 2)	https://www.YouTube.com/watch?v=mpHuvq-1sf4&feature=youtu.be	386K
	https://www.YouTube.com/watch?v=tQkNDPmR_8A&feature=youtu.be	435K
	https://www.YouTube.com/watch?v=jLZQ4T20REw	498K
Ricis Official 600/video (dataset 3)	https://www.YouTube.com/watch?v=RTzHbmTDsc0&t	262K
	https://www.YouTube.com/watch?v=ggMA9fuiv7k	267K
	https://www.YouTube.com/watch?v=h5KLC2G11Vk	253K

Data labeling was carried out manually using the sentiment scoring technique, by providing labels for each data, such as, positive code 2 (words such as praise and advice), negative symbol 0 (words of dislike or ridicule), and neutral code 1 (sentences without opinion elements).

3.1 Pre-Processing

The preprocessing of this study used Natural Language Processing (NLP). The technique consisted four steps, including, stopwords elimination, case folding, tokenizing, stemming, and several additional stages as follow;

- Cleansing, which includes deleting special texts, such as, (@, #, link), punctuation marks, blank characters, emoji representations, number symbols, and repetitive words. This was carried out because, vocabulary usually contains many repetitive characters. For example, the word *mantaaab* becomes *mantab* (fabulous), as *Kereeennn* changes

UNIVERSITAS MERCU BUANA

https://lib.mercubuana.ac.id

to *keren* (cool).

- Converting slang vocabulary into standard Indonesian Dictionary words. For example, *guee* converts to *saya* (me), *lu* becomes *kamu* (you), *wkwkwk* becomes *tertawa* (laugh) and many much more. In this conversion process, the researcher created a slang dictionary, containing 1344 words.
- Negation is a very important linguistic, because it affects the polarity of other words. It includes words, such as, no, less, and do not. Furthermore, its scope is limited only to the next word, or extended to other characters after the negation.

3.2 Coping with Unbalanced Classes

To overcome the unbalanced dataset in this study, the SMOTE technique was applied. SMOTE is a balancing technique, used to equate sample data distributions, within the minority and majority class [22]. The SMOTE stage involves calculating the distance between minority and majority data, determining the model percentage value, ascertaining the nearest k number, and creating synthetic information, with the following equation [22]:

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \qquad (1)$$

Where x_{syn} is the synthetic data to be created, x_i is the information to be replicated, x_{knn} is the sample that has the closest distance, from that which is to be replicated, and δ is the random value between 0 and 1.

UNIVERSITAS MERCU BUANA

4. Method

4.1 Representing Text Classification

The text classification in this study was carried out through ten stages, which is seen in Figure 1

The first stage involves data collection, while that of the second is manual labeling. The third and fourth stage involves preprocessing, with the sharing of data (training and testing), respectively. The fifth stage involves feature extraction with TF-IDF, where all words in the data are converted into a list, and assigned to each document as a vector, through the application of N-grams, to obtain the best results.

The sixth stage is the SMOTE scenario, where the data is divided into two types to be compared, in order to observe which is best for classification, between those using the technique or not. The seventh stage is the cross-validation process, which divides the overall data into 10 parts, where a part is used for testing, with the remaining 9 portions used for training. This process was observed to be repeated for all other sections.

The eighth stage is the classification process, carried out by using training data therefore, forming the proposed model. Furthermore, the model is used to predict the test data. The ninth stage is the model validation after formation. The tenth stage is the model testing and evaluation process. Testing and evaluation of the predictive results, are used to measure the performance of the classification model that has been made, and then visualized towards the end.

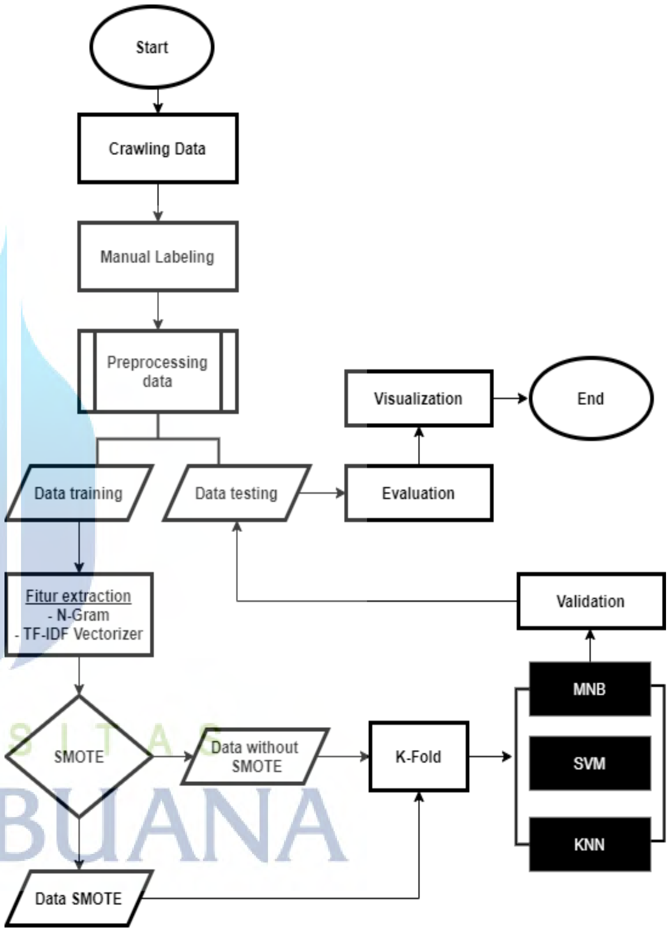


Figure 1. The flow of the text classification diagram.

Classification is carried out using various experimental scenarios, through the N-gram unigram feature, unigram-bigram, and unigram-trigram combination including,

1. The preprocessing stage does not remove stopwords, slang words, and does not handle negation, without selecting the N-gram feature and setting parameters. This scenario aims to determine the effect of the three preprocessing, N-gram features, and parameter settings on accuracy results.

2. The preprocessing stage uses the stopwords removal, slang word conversion, and negation handling processes, without selecting the N-gram feature and setting parameters. The aim is to determine the difference between the non-selective N-gram feature, without parameter setting, and those using both processes, with the effect of converting slang words and handling negation on the accuracy results.
3. All preprocessing steps are carried out, by selecting the N-gram feature, setting parameters and data proportion sharing. The proportion of data proposed in this study, is to compare 80% training and 20% testing data with 70% and 30% of both information, respectively. The aim is to determine the accuracy of all the processes to obtain the best accuracy, and test how effective the N-gram and parameter settings are, in improving the performance of the classification model.

4.2 Representing Topic Modeling

Modeling the topic of this research was carried out, to exploit the similarity in the presentation of content between YouTube channels, using LDA as observed in Figure 2.

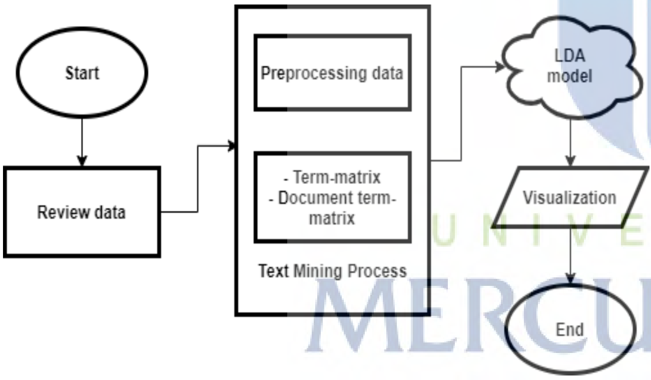


Figure 2. The flow of the topic modeling diagram.

In Figure 2, the first stage is reading the data collected. The second stage is the text mining process consisting of preprocessing, creating a term matrix and its document, using the BoW algorithm. After the data had been formed into a term-matrix document, it proceeds to the third stage, LDA modeling. Afterward, visualization is carried out at the last stage.

4.3 Evaluation Metrics

Classification performance measurement is carried out, to determine the performance results that have already been analyzed. The classification performance measurements used in this study, which are also preferred in similar studies are Accuracy, Precision, Recall, and F-measure [15, 17, 20, 21]. Meanwhile, to evaluate the modeling topic, the measurement of Perplexity and U_mass coherence is used.

5. Experiments and Results

Information relating to text classification, and topic modeling is presented in detail in this section.

5.1 Text Classification

To process the collected dataset, and learn the ML algorithm for text classification by using Python 3.6. The dataset used is 1800 comments. After each dataset had carried out data labeling and preprocessing (Tabel 2), the results in the form of the base words becomes the keywords for the feature extraction process.

Tabel 2. Preprocessing results.

Before Preprocessing	**Label**
Your content is not very useful, sorry for the children who watch it	0
Sister Afni looks like Sister Icis a few years ago, isn't it?	1
Just a suggestion, boss, make any content please, but if you meet difficult people (the poor), their faces are blurred because they also have self-respect	2
After Preprocessing	**Label**
the content is not useful, sir, sorry for the kids to watch	0
sis Afni really looks like icis, yes or no	1
just advise boss, in making content, when you meet a poor brother, please blur his face	2

The initial assumption of data is divided by the proportion of 80%:20% obtained 1440 training data and 360 testing data. Furthermore, after splitting the data, data is converted into feature vectors using TF-IDF. Problem comes, the number of comment labels on each dataset in the training data, experienced an imbalance class, which resulted in an unbalanced situation, in each section, e.g., leaning towards neutral or other sectors with details, is seen in Tabel 3. This needs to be addressed, in order not to affect the performance of the sentiment analysis carried out by applying the SMOTE technique.

Tabel 3. Number of training data labels.

Dataset 1	Dataset 2	Dataset 3	Label
147	103	101	0
369	698	710	1
924	639	629	2

The ratio of the original training data before oversampling, for example, in dataset 1 of positive, negative, and neutral classes is 924:147:369, while after

UNIVERSITAS MERCU BUANA

the process, possesses a ratio of 924:924:924 for each class.

Most importantly, before testing the model, it is necessary to determine the hyperparameter in training the model. The technique of selecting hyperparameters uses the grid search method, combined with cross-validation. Gridsearch is a scanning data method, to locate optimal parameters for a given model.

Some parameters considered for use are the MNB model, using alpha parameters with values 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, and 1. The linear kernel SVM model used parameter C, which consists of values 10^{-1}, 1, 10^{1}, 10^{2}, and 10^{3}. However, the K-NN model used the metric, n-neighbours, and leaf size parameters. Furthermore, the metric parameters used consisted of, Manhattan, Minkowski, and Euclidean. The N-neighbour parameters used, consisted of, 1, 3, 5, 7, and 9, while the leaf size parameters used, consisted of, 1, 2, 3, and 5.

The results further showed that, the best parameter of the MNB model was in the alpha data, with optimum value 0.1, and an accuracy of 91.94%. The SVM linear kernel model was then observed to be in parameter C, with the optimum value 10, and an accuracy of 93.89%, while the K-NN technique was in the leaf size 1 data, N-neighbour at 3, and Manhattan metric with a precise sum of 91.67%. The best parameter assessment of the GridsearchCV result, was observed from the base that had high average cross-validation accuracy, perfection, precision, recall, and f-measure. After obtaining the best parameters based on the results from the grid search, these data are to be used in the classification model testing phase. The overall grid search results are detailed in Tabel 4.

Tabel 4. Gridsearchcv results

Param	Value	Accuracy (%)	
		Test Score	Train Score
MNB (Alpha)	1	91,67	96.39
	0.1	**91.94**	97.99
	0.01	91.31	98.44
	0.001	90.90	98.66
	0.0001	90.90	98.61
SVM (C)	0.1	91.18	94.66
	1	93.68	99.37
	10	**93.89**	1.0
	100	93.89	1.0
	1000	93.89	1.0
K-NN (n_neighbor) leaf size 1	1	91.31	1.0
	3	**91.67**	1.0
	5	91.45	1.0
	7	90.20	1.0
	9	90.56	1.0

By conducting various experiments with two different data and 3 algorithm types, the model was put into operation. The test results are being observed in Table 5-9. The UNI column is for unigram features, the UNI + BI column is for the uni-bigram combination, while the UNI + TRI column is for the uni-trigram representation.

Tabel 5. Experiment 1 (sampel dataset 1).

Classification Models	Parameter Accuracy (%)	
	CV Score	Accuracy
MNB	77.55	77.33
MNB+SMOTE	90.22	90.22
K-NN	85.22	85.44
K-NN+SMOTE	67.66	66.11
SVM	88.44	90.22
SVM+SMOTE	89.88	91.11

The first experimental results by used proportion 80%:20% in Tabel 5 showed that, the SVM+SMOTE and MNB+SMOTE models, have better performances, even though they do not remove stopwords, do not change slang vocabulary, do not handle negation, parameter adjustment, and select N-gram features, than the MNB and SVM methods. This discusses that, the SMOTE technique application really helped improve model performance. Conversely, the K-NN model had better performance than the K-NN+SMOTE model, with very low accuracy because in the K-NN+SMOTE technique, the use of stopwords greatly affected the classification process.

Tabel 6. Experiment 2 (sampel dataset 1).

Classification Models	Parameter Accuracy (%)	
	CV Score	Accuracy
MNB	83.81	86.68
MNB+SMOTE	92.08	93.61
K-NN	87.91	88.61
K-NN+SMOTE	76.73	80.00
SVM	93.68	95.00
SVM+SMOTE	93.95	94.44

Based on the results of the second experiment by used proportion 80%:20% in Tabel 6 with complete preprocessing, all three models had increased from the initial experiment. The MNB+SMOTE model is still better than the MNB method, with an increase of 3.3%. Furthermore, the SVM and SVM+SMOTE models have competitive accuracy, with an increase of 4-5%. However, the K-NN+SMOTE model experienced an increase in accuracy of 14%. This confirmed that standard preprocessing, stopwords, slang vocabulary conversion, and effective negation handling improves model performance, even though the accuracy results of K-NN+SMOTE are still less than K-NN.

Tabel 7. Experiment 3 Dataset 1 (Atta Halilintar channel).

UNIVERSITAS MERCU BUANA

Classification Models	Proportion	TF-IDF Vectorizer Acc Score (%)		
		UNI	UNI+BI	UNI+TRI
MNB	80:20	93.88	93.61	**94.16**
	70:30	92.22	93.88	93.14
MNB SMOTE	80:20	93.33	95.00	**95.00**
	70:30	91.66	94.07	94.62
K-NN	80:20	**95.27**	92.50	93.05
	70:30	91.48	90.00	89.44
K-NN SMOTE	80:20	**94.44**	92.22	92.50
	70:30	93.14	91.11	89.81
SVM	80:20	95.56	**97.22**	96.94
	70:30	94.25	93.33	93.70
SVM SMOTE	80:20	95.56	**97.22**	96.94
	70:30	94.25	93.33	93.70

Tabel 8. Experiment 3 Dataset 2 (Baim Paula channel).

Classification Models	Proportion	TF-IDF Vectorizer Acc Score (%)		
		UNI	UNI+BI	UNI+TRI
MNB	80:20	91.94	92.22	**92.50**
	70:30	88.70	91.48	91.66
MNB SMOTE	80:20	91.66	93.05	**93.33**
	70:30	90.00	92.40	92.59
K-NN	80:20	91.94	**93.89**	93.33
	70:30	92.40	90.37	89.25
K-NN SMOTE	80:20	92.22	**94.16**	93.89
	70:30	91.48	90.00	89.44
SVM	80:20	**96.11**	95.83	95.27
	70:30	**96.48**	95.92	96.11
SVM SMOTE	80:20	**96.11**	95.83	95.27
	70:30	96.11	95.92	96.11

Tabel 9. Experiment 3 Dataset 3 (Ricis Official Channel).

Classification Models	Proportion	TF-IDF Vectorizer Acc Score (%)		
		UNI	UNI+BI	UNI+TRI
MNB	80:20	92.78	92.78	**92.78**
	70:30	89.44	89.81	89.81
MNB SMOTE	80:20	91.38	93.05	**93.89**
	70:30	90.37	89.81	90.37
K-NN	80:20	**90.27**	85.83	85.27
	70:30	90.18	89.25	88.14
K-NN SMOTE	80:20	**90.83**	88.33	88.05
	70:30	89.62	89.07	87.78
SVM	80:20	94.44	94.55	**94.72**
	70:30	94.25	93.33	93.70
SVM SMOTE	80:20	95.83	**96.38**	94.72
	70:30	94.25	93.33	93.70

The results of the third experiment in Tabels 7-9 showed that, the average share of the data proportions with 80% and 20% of both training and test dataset respectively, produced better accuracy in the three models, compared to the distribution of other information domains. This experiment showed that, the larger the number of training datasets, the better the accuracy of the model obtained.

Also, the addition of the parameter settings from Tabel 4, and the selection of the N-gram feature, contributed to the increased accuracy. It was observed that the three datasets of the MNB+SMOTE model, had better performances than the MNB method, with a combination of unigram and trigram features, where MNB+SMOTE in dataset 1,2,3 obtained an accuracy of 95%, 93,33%, and 93,89% respectively. While MNB in dataset 1,2,3 is only 94,16%, 92,50%, and 92,78%. Furthermore, the SVM and SVM+SMOTE models had the same accuracy of 97,22 % and 96,11% in dataset 1 and 2, even though they only experienced an increase of 1-2% from the second experiment. However, the K-NN model has increased by 14-15% from the second experiment. In dataset 1, the K-NN and K-NN+SMOTE models have competitive accuracy results. However, in dataset 2 and 3, the K-NN+SMOTE model was better than the K-NN model. In the three datasets, the highest average accuracy was observed in the N-gram, which is a combination of unigram-bigram and unigram-trigram. This showed that the combination of N-grams had more information than others, and is able to capture negation words in order to minimize errors.

Tabel 10. Best model comparison with SMOTE.

	Models	Parameter Accuracy (%)					Time (s)
		CV	Acc	P	R	F-M	
1	MNB SMOTE	92.98	95.00	92.67	95.67	94.00	0.78
	K-NN SMOTE	91.18	94.44	92.00	93.67	93.00	0.75
	SVM SMOTE	93.95	**97.22**	97.33	96.33	96.67	3.85
2	MNB SMOTE	89.93	93.33	90.67	94.00	91.67	0.63
	K-NN SMOTE	89.72	94.16	95.67	92.67	94.00	0.77
	SVM SMOTE	93.26	**96.11**	94.00	94.00	93.67	2.00
3	MNB SMOTE	89.72	93.89	90.67	91.00	91.00	0.63
	K-NN SMOTE	89.45	90.83	92.33	86.67	89.00	0.65
	SVM SMOTE	94.55	**96.38**	96.33	91.33	93.33	2.78

After the classification process is complete, the model with the SMOTE technique was better than that without the SMOTE technique. Therefore, the next step was to evaluate the performance of the model, with accuracy and confusion matrix parameters. Confusion matrix is a measure to analyze how well the classification model, recognizes the actual samples from different classes. The results with 5 different metrics are presented in Tabel 10. In terms of validation, it showed that the average result of 10-cross validation, was appropriate to be used as validation in this study, because the percentage obtained was above 80%.

UNIVERSITAS MERCU BUANA

In terms of the dataset, the three models have the best accuracy on data 1, compared to others, where MNB+SMOTE produced accuracy, precision, recall, and f-measure of 95%, 92.67%, 95.67%, and 94%, respectively. Furthermore, K-NN+SMOTE produced accuracy, precision, recall, and f-measure of 94.44%, 92%, 93.67%, and 93%, respectively. However, SVM+SMOTE produced accuracy, precision, recall, and f-measure of 97.22%, 97.33%, 96.33%, and 96.67%, respectively. In terms of accuracy, the SVM+SMOTE model was better than the other models in datasets 1, 2, 3 at 97.22%, 96.11%, and 96.38%, respectively. Meanwhile, in terms of processing time, although the SVM+SMOTE accuracy is superior, the resulting computation period is longer than the other models.

Tabel 11.Confusion matrix of the final best model.

Datasets	Actual	Prediction Class SVM+SMOTE		
		Negative	Neutral	Positive
1	Negative	38	0	1
	Neutral	0	81	6
	Positive	1	2	231
2	Negative	21	3	0
	Neutral	1	150	6
	Positive	2	2	175
3	Negative	25	3	4
	Neutral	1	147	2
	Positive	0	3	175

Furthermore, the final best model results were obtained from the SVM+SMOTE technique, in the entire dataset. Based on Tabel 11, the result of confusion matrix from the negative class was 38 times, with only one predicted error as positive. From the neutral class side, only 81 times are correctly classified, with a predictive error of 6 as positive. Meanwhile, from the positive class side, only 231 times are classified as true positive, 1 is predicted as a negative, and the other 3 are forecasted as neutral sections.

The overall visualization of the best model classification results, was displayed through the website application made by the researcher in Figures 3-5.

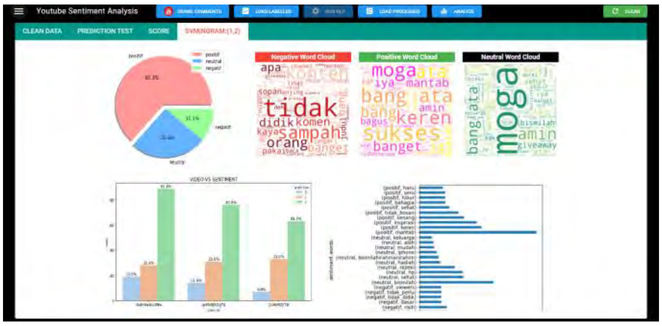


Figure 3. Visualization results of the Atta Halilintar channel.

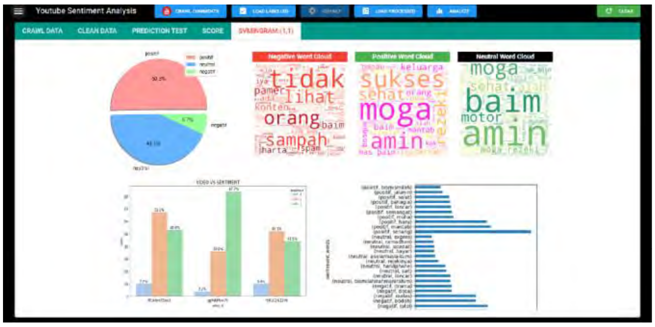


Figure 4. Visualization results of the Baim Paula channel.

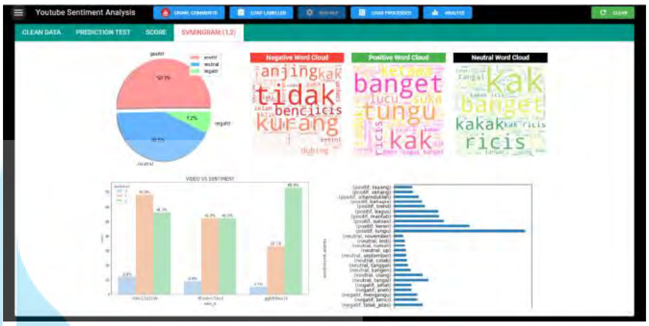


Figure 5. Visualization results of the Ricis Official channel.

Based on the results of the chart, word cloud visualization and TF-IDF weighting in Figures 3-5, the video content on dataset 1 of Atta Halilintar channel still had shortcomings, based on the negative comments received, namely, ‘Not educated’, ‘UncomforTabel’, ‘Lazy’, ‘Disrespectful’, and ‘Unclear’. Meanwhile, the advantages of Atta Halilintar’s content, based on positive comments are ‘Fun’, ‘Entertaining’, ‘Not bored’, ‘Inspiring’, ‘Cool’, and ‘Marvelous’. The percentage of sentiment from the predicted results obtained on dataset 1 is positive by 63.3%, negative by 11,1%, and neutral by 25.6%.

In dataset 2, the video content of Baim Paula channel also had disadvantages, based on negative comments received, namely, ‘Spam’, ‘Drama’, ‘Lazy’, ‘Weird’, ‘Showing off’, and ‘Stupid’. However, the advantages of Baim Paula content, based on positive comments are ‘Salute’, ‘Happy’, ‘Noble’, ‘Pleased’, ‘Marvelous’, and ‘Emotional’. The percentage of sentiment from the predicted results obtained on dataset 2 is positive by 50.3%, negative by 6.7%, and neutral by 43.1%.

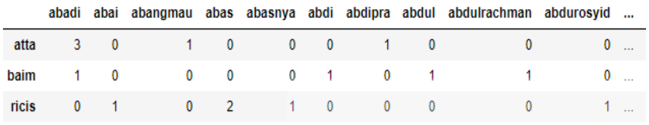
Furthermore, in dataset 3, video content of Ricis Official channel had disadvantages, based on negative comments received, namely, ‘Unclear’, ‘Dislike’, ‘Not long enough’, ‘Annoying’, and ‘Weird’. Moreover, the advantages of Ricis Official content, based on positive comments were ‘Pleased’, ‘Happy’, ‘Trendy’, ‘Good’, ‘Marvelous’, ‘Cool’, and ‘Successful’. The percentage of sentiment from the predicted results obtained on

UNIVERSITAS MERCU BUANA

dataset 3 is positive by 50.3%, negative by 7.2%, and neutral by 42.5%.

5.2 Topic Modeling

To implement topic modeling with LDA, it utilizes the library provided by python, namely, gensim. The process requires a document matrix and the number of topics, wanted by the algorithm. The first step is to combine the text list into one large part, in the form of pandas dataframe, for data cleaning. Furthermore, the corpus is converted into a term-matrix document using the BoW algorithm, which is shown in Figure 6.

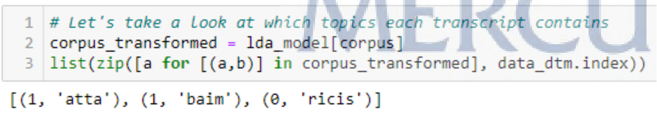


	abadi	abai	abangmau	abas	abasnya	abdi	abdipra	abdul	abdulrachman	abdurosyid	...
atta	3	0	1	0	0	0	1	0	0	0	...
baim	1	0	0	0	0	1	0	1	1	0	...
ricis	0	1	0	2	1	0	0	0	0	1	...

Figure 6. Result of term-matrix document.

After obtaining a term-matrix document, the next step is to place it into the new gensim format. Also, the gensim requires id2word, which is a dictionary of all terms, and their respective locations, in the term-matrix document. After obtaining the corpus and id2word, topic modeling is carried out, using LDA with the number of topics is 2, number of words is 10, and 60 passes value. The value of each parameter is adjusted to the dataset which results in a good, reasonable distribution of topics, and more human interpreTabel. To obtain similarities in topics, all corpus are processed in zipped function to map the same index with more than one itterable, which is shown in Figure 7.

UNIVERSITAS
MERCU BUANA



```
# Let's take a look at which topics each transcript contains
corpus_transformed = lda_model[corpus]
list(zip([a for [(a,b)] in corpus_transformed], data_dtm.index))
```

```
[(1, 'atta'), (1, 'baim'), (0, 'ricis')]
```

Figure 7. Result of code program.

The results of LDA modeling on each dataset, with coherence values is shown in Tabel 12 and Figures 8-9.

Tabel 12. Results of topic modeling.

	Topic	Channel
0	0.021*"laugh"+0.017*"wait "+ 0.016*"good"+0.012*"funny"+ 0.012*"like"+ 0.010*"cool"+ 0.010*"movie"+0.008*"video"+ 0.008*"content"+0.007*"broadcast"	Ricis Official (Dataset 3)
1	0.026*"success"+0.019*"sustenance"+ 0.018*"bismilah"+0.013*"family"+ 0.013*"giveaway"+0.013*"good"+ 0.011*"content"+0.011*"marvelous"+ 0.011*"cool"+0.010*"motorbike"	Atta Halilintar (Datasset 1) and Baim Paula (Dataset 2)

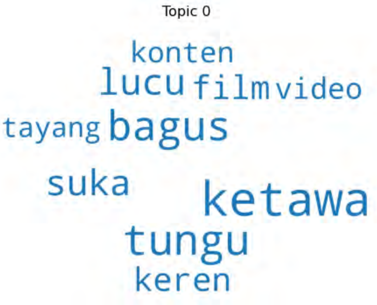


Figure 8. The visualization results of topic modeling for topic 0.

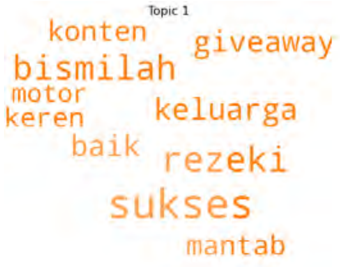


Figure 9. The visualization results of topic modeling for topic 1.

After the model is done, it is continued by evaluating the model using perplexity and coherence score. The perplexity result of the modeling topic obtained is -7.166900721475966 and the u_mass coherence score of -0.039479497367376465.

Based on the results of the topic modeling in Tabel 12, the Latent Dirichlet Allocation (LDA) method obtained similarities in topics related to content presentation, between the three YouTube channels, in this study. The Atta Halilintar channel (dataset 1) and the Baim Paula channel (dataset 2), have similar topics, namely, Topic 1. Results of the distribution topics on Topic 1 explains that, the Atta and Baim channels present content with a theme of family, sheer sustenance, and giveaways, in the form of motorbikes. Furthermore, results of the distribution topic on Topic 0 for the Ricis Official channel (dataset3) had different topics from the Atta and Baim accounts. The Ricis channel presents content with funny themes, triggers laughter, and episode movies.

6. Conclusion

Based on the analysis and testing results by various experiments, it is concluded that in the text classification with an imbalanced dataset, the use of SMOTE technique is quite effective in improving the performance of the MNB and SVM models, even without complete preprocessing. However, for the K-

NN model, the SMOTE technique does not work well, without complete preprocessing.

To obtain good classification results does not only depend on what classification algorithm is used, but also what steps has been applied. The use of the complete preprocessing, parameter setting, and selection of N-gram's advanced features, also contributed to improving the performance of the three models. The selection of features in the three models of the datasets, showed that the best classification performance was obtained at N-gram, which is a combination of unigram-bigram and unigram-trigram.

The classification model that produced the best performance, was obtained from a model with the SMOTE technique. Based on the accuracy value, SVM+SMOTE was superior to other models on datasets 1, 2, and 3, at 97.22%, 96.11%, and 96.38%, respectively. Furthermore, it was followed by MNB+SMOTE on dataset 1, 2, 3, at 95%, 93.33%, 93.89%, respectively, and K-NN+SMOTE on information 1, 2, 3, at 94.44%, 94.16%, 90.83%, sequentially.

However, for the topic modeling results, the Latent Dirichlet Allocation (LDA) method was confirmed to have the ability to obtain similarities in topics from the content presentation, between the three YouTube channels in this study. The results further showed that, the Atta Halilintar (dataset 1) and Baim Paula (dataset 2) channels have similar topics. This topic modeling prevents plagiarism towards content presentation.

Thus from this research can be used as an example of a reflective model to improve future learning in the field of sentiment analysis and topic modeling. For further development in sentiment analysis, it was suggested that, can test more data, a lexicon-based automatic data labeling process should be carried out and try various experiments by using other methods, such as deep learning for classification. Meanwhile for further development in topic modeling can be done with different data and methods.

References

[1] Sutoyo E. and Almaarif A., "Twitter sentiment analysis of the relocation of Indonesia's capital city," *Bulletin of Electrical Engineering and Informatics*, no. 1, pp. 474–478, 2017.

[2] Tricahyo V.A. and Isa S.M., "Classification of Indonesian Presidential Campaign on Twitter Using Word2Vec," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 9, no. 4, pp. 5501-5508, 2020.

[3] Kaur C. and Sharma Dr. A., "Sentiment Analysis of Tweets on Social Issues using Machine Learning Approach," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 9, no. 4, pp. 6303-6311, 2020.

[4] Metin, B., and Köktas, H., "Sentiment Analysis with Term Weighting and Word Vectors," *The International Arab Journal of Information Technology*, vol. 16, no. 5, 2019.

[5] Fitriana, D.N., and Sibaroni, Y., "Sentiment Analysis on KAI Twitter Post Using Multiclass Support Vector Machine (SVM)," *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi)*, vol 4, no. 5, pp. 846-853, 2020 .

[6] Bhuiyan, H., Ara, J., Bardhan, R., and Islam, M.R., "Retrieving YouTube Video by Sentiment Analysis on User Comment," *in IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Malaysia, pp. 474–478, 2017.

[7] Aribowo, A., Basiron, H., Herman, N.S., and Khomsah, S., "An evaluation of preprocessing steps and tree-based ensemble machine learning for analysing sentiment on Indonesian YouTube comments," *International Journal of Advanced Trends in Computer Science and Engineering,* vol.9, no. 5, pp. 7078-7086, 2020.

[8] Samsudin, N.M., Foozy, C.F., Alias, N., Shamala, P., Othman, N.F., and Din, W., "YouTube spam detection framework using naïve bayes and logistic regression," *Indonesian Journal of Electrical Engineering and Computer Science.* vol. 14, no. 3, pp. 1508-1517, 2019.

[9] Chaithra V.D., "Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 9, no. 5, pp. 4452-4459, 2019.

[10] Fernando, J., Budiraharjo, R., and Haganusa, E., "Spam Classification on 2019 Indonesian President Election YouTube Comments Using Multinomial Naïve Bayes," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM),* vol. 2, pp. 37–44, 2019.

[11] Tanesab F.I., "Sentiment Analysis Model Based On YouTube Comment Using Support Vector Machine," *International Journal of Computer Science and Software Engineering (IJCSSE),* vol. 6, pp. 180-185, 2017.

[12] Novendri, R., Callista, A.S., Pratama, D.N., and Puspita, C.E., "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes,"

https://lib.mercubuana.ac.id

Bulletin of Computer Science and Electrical Engineering, vol. 1, no. 1, pp. 26-32, 2020.

[13] Shaout A. and Crispin B., “Streaming Video Classification Using Machine Learning,” *The International Arab Journal of Information Technology,* vol. 17, no. 4A, 2020.

[14] Aufar, M., Andreswari, R., and Pramesti, D., “Sentiment Analysis on YouTube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study,” *in 2020 International Conference on Data Science and Its Applications (ICoDSA)*, Indonesia, pp. 1-7, 2020.

[15] Poché, E., Jha, N., Williams, G., Staten, J., Vesper, M., and Mahmoud, A., “Analyzing User Comments on YouTube Coding Tutorial Videos,” *in IEEE/ACM 25th International Conference on Program Comprehension (ICPC),* Argentina, pp. 196-206, 2017.

[16] Khomsah S. and Aribowo A.S., “Model Text-Preprocessing Komentar YouTube Dalam Bahasa Indonesia”. *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi),* vol. 4, pp. 648-654, 2020.

[17] Satriaji W. and Kusumaningrum R., “Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis,” *in 2nd International Conference on Informatics and Computational Sciences (ICICoS),* Indonesia, 2018.

[18] Pujadayanti I., Fauzi M., and Arum S.Y., “Prediksi Rating Otomatis pada Ulasan Produk Kecantikan dengan Metode Naïve Bayes dan N-gram,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 11, pp. 4421-4427, 2018.

[19] Sahira Y. and Fudholi D.H., “Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation),” *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no.2, pp. 336-334, 2020.

[20] Ramayanti D. and Salamah U., “Text Classification on Dataset of Marine and Fisheries Sciences Domain Using Random Forest Classifier,” *International Journal of Computer Techniques*, vol. 5, pp. 1-7, 2018.

[21] Sadikin M. and Alfiandi F., “Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no.6, pp. 4763-4771, 2018.

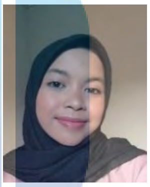
[22] Kasanah A.N., Muladi., and Pujianto U., “Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objectivitas Berita Online Menggunakan Algoritma K-NN,” *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no.2, pp. 196-201, 2019.

[23] Lubis, F.F., Rosmansyah, Y., and Supangkat, S.H., “Topic discovery of online course reviews using LDA with leveraging reviews helpfulness,” *International Journal of Electrical and Computer Engineering (IJECE),* vol. 9, no. 1, pp. 426-438, 2019.

[24] Ruan, S., Li, H., Li, C., and Song, K., “Class-Specific Deep Feature Weighting for Naïve Bayes Text Classifiers,” *IEEE Access,* vol.8, pp. 20151- 20159, 2020.

[25] Suyanto, *Data Mining Untuk Klasifikasi Dan Klasterisasi Data*, Edisi Revisi, Informatika Bandung, Bandung, 2019.

[26] Arhami M. and Nasir M., *Data Mining Algoritma dan Implementasi*, Edisi 1, ANDI, Yogyakarta.



Annisa Rizki Liliandari is a student in Faculty of Computer Science at Mercu Buana University in Indonesia. Born in Tangerang on October 30th, 1999, this author is interested in data mining, algorithm analysis and web programming. Current reseach scope is machine learning, natural language processing, text classification, and topic modeling.

Ida Nurhaida was born in Kuantan (Malaysia) in 1971. This author is a researcher, and member of Faculty of Computer Science Universitas Mercu Buana, Indonesia. Areas of interest and research are in the image processing pattern recognition and image retrieval system, which was formalised in 2010 (PhD), on this subject in the Faculty of Computer Science, University of Indonesia. This author has presented papers at conferences, both home and abroad, published articles and papers in various journals, and contributed a chapter to the book.

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id

KERTAS KERJA

Ringkasan

Kertas kerja ini merupakan material kelengkapan artikel jurnal yang telah di lampirkan sebelumnya dengan judul "Topic Discovery and Classification Comparison on the Comments of Indonesian Entertainment YouTube Channel Videos Using SMOTE, N-gram, and LDA Approaches". Kertas kerja ini berisi keseluruhan material hasil penelitan Tugas Akhir yang tidak dimuat atau disertakan di dalam artikel jurnal. Di dalam kertas kerja ini menjelaskan penelitian secara rinci dan jelas mulai dari literatur review, dataset yang digunakan, metodologi atau tahapan eksperimen, source code, dan hasil pengolahan serta eksperimen secara keseluruhan akan lebih jelas di lampirkan. Di dalam kertas kerja disajikan:

- Bagian 1: Literatur Review

  Membahas mengenai literatur review yang berisi artikel jurnal dari penelitian sebelumnya dan teori yang menjadi dasar atau landasan dalam penelitian ini.
- Bagian 2: Dataset

  Membahas mengenai dataset yang digunakan dalam penelitian ini, meliputi: cara perolehan data, sumber data, kriteria pelabelan data, sampel dari dataset, dan Npenyesuaian data akhir yang siap untuk diolah.
- Bagian 3: Tahapan Penelitian

  Menjelaskan mengenai tahapan penelitian yang disajikan dalam bentuk alur diagram dengan penjelasan dari setiap tahapan dan skenario eksperimen yang dilakukan.
- Bagian 4: Source Code
- Bagian 5: Hasil semua eksperimen

  Menjelaskan hasil keseluruhan dari eksperimen yang telah dilakukan dengan penjelasan yang lebih lengkap.
- Bagian 6: Kesimpulan dan Saran

UNIVERSITAS MERCU BUANA

https://lib.mercubuana.ac.id