International Journal of Computer, Network Security and Information System (IJCONSIST)
Vol. 5 No. 2, Maret 2024

Sentiment Analysis on Generation Z News Article using Support Vectore Machine (SVM) with Synthetic Minority Over-sampling Technique (SMOTE)

Kartini
Informatics
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Surabaya, Indonesia
kartini.if@upnjatim.ac.id

Kartika Maulida Hindrayani
Data Science
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Surabaya, Indonesia
kartika.maulida.ds@upnjatim.ac.id

Betty Dewi Puspasari
Institut Teknologi Nasional Malang, National Dong Hwa University
Indonesia, Taiwan

***Abstract*— The development of digital media has increased the volume of news articles discussing various issues, including those involving Generation Z. Understanding public perception of these news items can be achieved by applying a crucial approach, namely sentiment analysis. This study aims to classify sentiment in news articles about Generation Z using the Support Vector Machine (SVM) algorithm. The main challenge in sentiment analysis is data class imbalance, where the amount of positive and negative sentiment data is often unbalanced. Therefore, the Synthetic Minority Over-sampling Technique (SMOTE) is used to address this problem by balancing the class distribution before model training. The datasets used were collected from various online news portals and analyzed through text preprocessing, feature extraction using Bag of Word, and SVM model training. The evaluation results show that the application of SMOTE significantly improves the model's performance in classifying sentiment, with improvements in accuracy, precision, recall, and F1-score compared to the model without data imbalance handling. This study demonstrates that the combination of SVM and SMOTE is effective in conducting sentiment analysis on Generation Z news articles. The accuracy shows 84% with 83% precision and 76% recall.**

Keywords*— Sentiment analysis, Generation Z, Support Vector Machine, SMOTE, text classification.

I. INTRODUCTION (*HEADING 1*)

Generation Z has now entered the world of offices and higher education and will subsequently become the dominant population [1]. The characteristics of Gen Z are quite different from previous generations, namely millennials and boomers, also shaped by the technology that is available today. In their development, Gen Z are more accustomed to coexist with technology [2][3]. Gen Z can receive a lot of information anywhere and anytime with the support of the internet, social networks, and long-distance communication. The use of technology by Gen Z is already very advanced. They are also used to being bombarded with a large amount of information and instant access to something. This phenomenon can provide both advantages and disadvantages for Gen Z. Proficiency in technology can provide useful competencies in carrying out activities. Meanwhile, instant requests by Gen Z should be minimized to better understand the process of carrying out an activity.

The explosion of information and technological skills received by Gen Z does not mean that Gen Z is the smartest generation compared to previous generations [4]. With characters formed by technological support, Gen Z should be educated according to the times. Educators are required to find new ways of teaching and transferring knowledge to meet the needs of Gen Z. Educational strategies must be created to accommodate Gen Z's education so that it can reach its maximum potential. The education needed by Gen Z is tailored to the characteristics and expectations of the recipients, including knowledge, experiential learning, technical skills, communication, and problem-solving [5].

Engaging learning for Gen Z involves the use of interactive technology and social media [6]. A hybrid learning approach supported by media such as images, audio, and video is more appealing to them than text. Some learning tools besides social media that can be used include Quizziz [7]. Popular social media platforms include YouTube, WhatsApp, Twitter, Instagram, and TikTok [8], [9].

News articles have discussed Gen Z behavior. Some believe that the ease of use of technology at a young age also contributes to Gen Z's expectations of the world. However, this perception is not only negative, but also positive. Thanks to their creativity in the digital world, many have become entrepreneurs and become successful at a young age. News about Gen Z can be analyzed through sentiment analysis. Sentiment analysis is a method for determining the positive or negative perspective of text data. Some popular sentiment analysis methods include Support Vector Machine (SVM) and Naive Bayes.

SVM can classify sentiment analysis on social media regarding the university social media. In this study, SVM produced an accuracy of 60.6% with Bag of Word feature selection [10]. Another study conducted on general elections using social media data. The method used was K-Nearest Neighbor. The final result was an accuracy of 76%. SVM can also be used to classify hate speech in online news data. The data collected consisted of 33 new[11]s items, resulting in an accuracy of 53.88% [12]. News data can also be classified using Naïve Bayes. Classification of COVID-19 news headlines with a total of 399 items achieved an accuracy of 72.5% . An analysis of online learning news using Naïve Bayes was conducted . The accuracy was 87% [13]. This study will discuss news sentiment analysis regarding Gen Z. The algorithms used are SVM and Naïve Bayes optimized with SMOTE.

II. RESEARCH METHODS

A. Research Framework

This research was conducted in several stages, starting with data collection, data preprocessing, modeling using SVM and Naïve Bayes, and then evaluating the results. An illustration can be seen in Figure 1.

The basic research framework in Figure 1 represents the systematic steps implemented to develop a news sentiment analysis prediction model for Gen Z. The first step of the research was a literature review on Generation Z, the SVM algorithm, and the Naïve Bayes algorithm. The outputs of the literature review were problem identification, research gaps, and objective formulation.

B. Data collection

Data collection was conducted by scraping data from publicly accessible online news. The keyword criteria used in data collection were Generation Z, Gen Z, Gen Z intelligent, Gen Z smart, Gen Z college, Gen Z education, and Gen Z learning interests. The data collected included media, titles, and news texts. A total of 186 news items were collected. Table 1 explains the metadata of the research variables.

TABLE I. METADATA

Variable	Description	Data Type
Media	Name of the media where the news text was taken	*String*
Title	Title of the news text taken	*String*
News Text	News text that will be used for sentiment analysis prediction	*String*

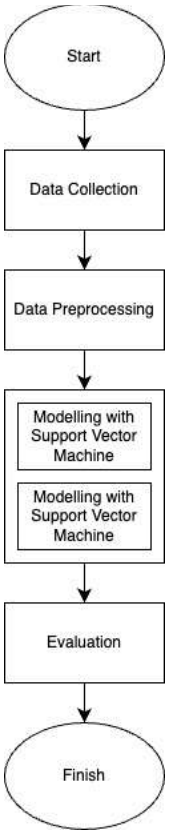

Fig. 1. Workflow

C. Data Preprocessing

Data preprocessing includes cleaning, case folding, tokenization, stopwords, and stemming. Cleaning is the process of data cleansing. Data cleaning performed at this stage includes removing numbers, formats, links, hashtags, punctuation, and excess spaces. Tokenization is the division of a text sentence into specific parts [12]. Stopwords are the removal of meaningless words [14]. The library used during the stopword process is NLTK (Natural Language Toolkit). Words in the library can be removed because they have no

function/meaning. Stemming is the process of removing initial and final affixes to form basic words [15].

D. Modeling using SVM

Support Vector Machine (SVM) is a supervised machine learning-based approach. The basic concept of SVM is to find a hyperplane that maximizes the distance between classes [16]. The hyperplane equation in SVM can be written as follows (Tripathi, 2021).

$$W^T - b = 0 \quad (1)$$

Where w is the normal vector of the hyperplane and $\frac{b}{\|W\|}$ shows the shift of the hyperplane from the origin along the normal vector w.

E. Modeling using Naïve Bayes

The Naïve Bayes algorithm uses Bayes' theorem with the following formula (Jackins et al., 2021).

$$P\left(\frac{a}{y}\right) = \frac{\left(P\left(\frac{y}{a}\right)P(a)\right)}{P(y)} \quad (2)$$

Where $P\left(\frac{a}{y}\right)$ indicates the posterior probability of the class. P(a) interprets the prior probability. $P\left(\frac{y}{a}\right)$ indicates the likelihood which is the probability of the predictor given the class. $P(y)$ indicates the initial probability of the predictor.

2.6 Addressing data imbalance with SMOTE

Synthetic Minority Oversampling Technique, is an oversampling technique, but it works differently from traditional oversampling. In classic oversampling, minority data is duplicated from a minority data population. This technique increases the amount of data but does not provide new information or variation to the machine learning model.

2.7 Evaluation of Results

The evaluation results used in this study were a confusion matrix, which includes accuracy, precision, and recall. Accuracy is the value of correct predictions, both positive (True Positive/TP) and negative (True Negative/TN). The accuracy formula can be described as follows:

$$Accuracy = \frac{TP + TN}{N} \quad (3)$$

Precision is the proportion between the predictions that are truly positive and the set of all true positive values. The formula for precision is

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall is a calculation that measures the proportion of the set of all positive values that are correctly predicted as positive by the model.

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

III. RESULTS AND DISCUSSION

A. Cleaning

The data cleaning process involves case folding. Capital letters are replaced with lowercase letters. Numbers and symbols are also cleaned in this step. The year and the hyphen (-) and comma (,) symbols discussed are removed. Examples of text and the results of the data cleaning can be seen in the table.

TABLE II. CLEANING TEXT PROCESS

Text	Cleaning Text
Generasi Z lahir ditengah perubahan teknologi digital yaitu: figital, fear of missing out (FOMO), hiperkustomisasi, terpacu, realistis.	lahir ditengah perubahan teknologi digital yaitu fear of missing out fomo hiperkustomisasi terpacu realistis
"Generasi Milenial terdiri dari mereka yang lahir antara tahun 1982 dan awal 90-an, sedangkan Generasi Z terdiri dari mereka yang lahir tahun 1995 sampai 2012	milenial terdiri dari mereka yang lahir antara tahun dan awal an sedangkan terdiri dari mereka yang lahir tahun sampai

B. Tokenization

An example of tokenization is dividing a sentence into words. The table shows that the sentence has been divided into individual words.

TABLE III. TOKENIZATION PROCESS

Cleaning Text	Tokenization
lahir ditengah perubahan teknologi digital yaitu fear of missing out fomo hiperkustomisasi terpacu realistis	['lahir', 'ditengah', 'perubahan', 'teknologi', 'digital', … , 'yaitu', 'fear', 'of', 'missing', 'out', 'fomo', 'hiperkustomisasi', 'terpacu', 'realistis']
milenial terdiri dari mereka yang lahir antara tahun dan awal an sedangkan terdiri dari mereka yang lahir tahun sampai	['milenial', 'terdiri', 'dari', 'mereka', 'yang', 'lahir', 'antara', 'tahun', 'dan', 'awal', 'an', 'sedangkan', 'terdiri', 'dari', 'mereka', 'yang', 'lahir', 'tahun', 'sampai']

C. Stopword Removal

An example of the application of stopword removal is that words that have no meaning such as 'dari', 'yang', 'dan' can be removed. As seen in Table III, the words from nineteen words have been reduced to two words. Most of the words did not have any meaning then they were removed.

TABLE IV. STOPWORD REMOVAL PROCESS

Tokenization	Stopword Removal
['lahir', 'ditengah', 'perubahan', 'teknologi', 'digital', … , 'yaitu', 'fear', 'of', 'missing', 'out', 'fomo', 'hiperkustomisasi', 'terpacu', 'realistis']	['lahir', 'ditengah', 'perubahan', 'teknologi', 'digital', …, 'fear', 'of', 'missing', 'out', 'fomo', 'hiperkustomisasi', 'terpacu', 'realistis']
['milenial', 'terdiri', 'dari', 'mereka', 'yang', 'lahir', 'antara', 'tahun', 'dan', 'awal', 'an', 'sedangkan', 'terdiri', 'dari', 'mereka', 'yang', 'lahir', 'tahun', 'sampai']	['milenial', 'lahir']

D. Stemming

Stemming is the process of turning a word into a root word. For example, stemming results in the word "terpacu" becoming "pacu." Words removed to form the root word are "ter-," "pen-," and "di-."

TABLE V. STEMMING PROCESS

Stopword Removal	Stemming
['lahir', 'ditengah', 'perubahan', 'teknologi', 'digital', …, 'fear', 'of', 'missing', 'out', 'fomo', 'hiperkustomisasi', 'terpacu', 'realistis']	['lahir', 'tengah', 'ubah', 'teknologi', 'digital', …, 'fear', 'missing', 'out', 'fomo', 'hiperkustomisasi', 'pacu', 'realistis']
['milenial', 'lahir', ‘pendidik’]	['milenial', 'lahir', ‘didik’]

E. Wordcloud Visualization

After the sentences have been preprocessed, visualization can be performed to review frequently occurring words. The word cloud can be seen in Fig 2.

Fig. 2. Wordcloud

F. Labelling

Positive and negative labeling was performed to understand the sentiment analysis generated. Overall, positive sentiment accounted for 72.97% of the data, while negative sentiment accounted for 27.03% as seen on Fig 3. This led to an imbalance in the data. SMOTE was applied to improve the balance of the dataset.

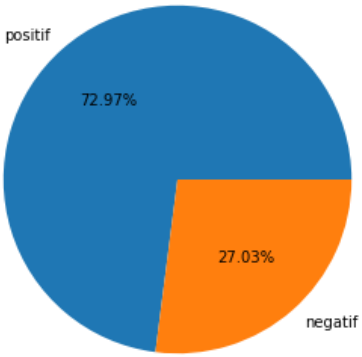

Fig. 3. Positive – Negative Sentiment

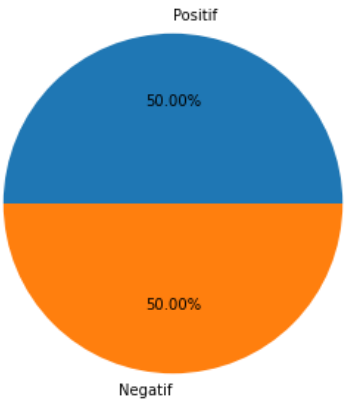

Fig. 4. After Balancing

Fig 4 shows balance dataset. Positive and negative sentiments were balance with 50% each.

TABLE VI. EVALUATION RESULTS

Algorithm	Balance Dataset Optimization	Accuracy	Precision	Recall
SVM	-	82%	80%	74%
SVM	SMOTE	84%	83%	76%
Naïve Bayes	-	64%	56%	56%
Naïve Bayes	SMOTE	75%	69%	64%

The best accuracy was achieved at 84% after applying the SVM algorithm with SMOTE data processing. Precision and recall also improved to 83% and 76%, respectively. Overall, Naïve Bayes performance was still lower than SVM in terms of accuracy, precision, and recall.

IV. CONCLUSIONS

In this study, the author has conducted a classification on Gen Z news article using Support Vector Machine and Naïve Bayes model. SMOTE was used to handle the imbalance dataset. In addition to obtain accuracy and performance results for each label class, it was found that the model used was good enough to conduct this gen z sentiment analysis. The experimental results obtained showed that gen z news article has a sentiment that is quite positive rather than negative. By obtaining an accuracy value of 84% and the performance for each label class is also quite good.

REFERENCES

[1] A. Sękala, O. Schultz, K. Foit, and T. Blaszczyk, "Technical Education for Industry 4.0: Generation Z's motivation to study technical subjects - a comparative study of selected universities in Denmark and Poland," in *2023 32nd Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE)*, IEEE, 2023.

[2] L. Nurin Nabila, F. Putra Utama, A. Ahya Habibi, I. Hidayah, and M. Aliyah Negeri, "Aksentuasi Literasi pada Gen-Z untuk Menyiapkan Generasi Progresif Era Revolusi Industri 4.0," 2023.

[3] C. Paulina and E. Ernawati, "How to Develop Learning Styles to Encourage Gen Zers in Their Academic Performance and Workforce," *Business Economic, Communication, and Social Sciences (BECOSS) Journal*, vol. 4, no. 2, pp. 121–132, Jun. 2022, doi: 10.21512/becossjournal.v4i2.8378.

[4] A. Szymkowiak, B. Melović, M. Dabić, K. Jeganathan, and G. S. Kundi, "Information technology and Gen Z: The role of teachers, the internet, and technology in the education of young people," *Technol Soc*, vol. 65, May 2021, doi: 10.1016/j.techsoc.2021.101565.

[5] J. P. Landry *et al.*, "Reaching and Retaining the Next Generation: Adapting to the Expectations of Gen Z in the Classroom," *Information Systems Education Journal (ISEDJ)*, vol. 16, no. 3, pp. 45–54, Jun. 2018, [Online]. Available: http://www.edsigcon.org

[6] C. K. Y. Chan and K. K. W. Lee, "The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial generation teachers?," *Smart Learning Environments*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40561-023-00269-3.

[7] T. A. Atieka, "ANALISIS PENGGUNAAN QUIZIZZ DALAM PENINGKATKAN LITERASI DIGITAL GENERASI Z MENDUKUNG SDGs," vol. 16, pp. 177–188, 2024, [Online]. Available: www.quizizz.com

[8] C. Nismeta Rotua Mendrofa and N. Aprilia, "Pemanfaatan Media Sosial oleh Generasi Z sebagai Media Pembelajaran (Utilization of Social Media By Generation Z As A Learning Medium)," *Jurnal Tonggak Pendidikan Dasar : Jurnal Kajian Teori dan Hasil Pendidikan Dasar*, vol. 2, no. 1, pp. 23–35, Jul. 2023, doi: 10.22437/jtpd.v2i1.22874.

[9] D. D. Sholekah and S. Wahyuni, "Pemanfaatan Media Sosial dalam Proses Pembelajaran di SMPN 1 Mojo Kediri," *Indonesian Journal of Islamic Education Studies (IJIES)*, vol. 2, no. 1, pp. 50–60, Aug. 2019, doi: 10.33367/ijies.v2i1.850.

[10] F. Azka, W. #1, A. Romadhony, and H. #3, "Sentiment Analysis of University Social Media Using Support Vector Machine and Logistic Regression Methods", doi: 10.34818/indojc.2022.7.2.638.

[11] F. Delfariyadi, A. Helen, and S. Yuliawati, "Klasifikasi Sentimen Judul Berita Pemberitaan COVID-19 Tahun 2021 pada Media DetikHealth."

[12] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," vol. 7, no. 1, pp. 1–10, 2020, [Online]. Available: http://jurnal.mdp.ac.id

[13] U. Kusnia, F. Kurniawan, and S. Artikel, "Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes INFO ARTIKEL ABSTRAK", doi: 10.35891/explorit.

[14] B. A. Yuniarossy *et al.*, "ANALISIS SENTIMEN TERHADAP ISU FEMINISME DI TWITTER MENGGUNAKAN MODEL CONVOLUTIONAL NEURAL NETWORK (CNN)," vol. 5, no. 1, 2024, doi: 10.46306/lb.v5i1.

[15] M. Daffa, A. Fahreza, A. Luthfiarta, M. Rafid, M. Indrawan, and A. Nugraha, "Analisis Sentimen: Pengaruh Jam Kerja Terhadap Kesehatan Mental Generasi Z," *JOURNAL OF APPLIED COMPUTER*

SCIENCE AND TECHNOLOGY (JACOST), vol. 5, no. 1, pp. 2723–1453, 2024, doi: 10.52158/jacost.715.

[16] D. B. Vukovic, K. Romanyuk, S. Ivashchenko, and E. M. Grigorieva, "Are CDS spreads predictable during the Covid-19 pandemic? Forecasting based on SVM, GMDH, LSTM and Markov switching autoregression," *Expert Syst Appl*, vol. 194, May 2022, doi: 10.1016/j.eswa.2022.116553.