JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 8, No. 2, December 2025, pp. 788~797
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v8i2.4707

Classification and Mapping of Online Gambling Based on News Articles Using NER and SVM

Wisnu Mukti Darwansah[1], Amalia Anjani Arifiyanti[2], Rizka Hadiwiyanti[3]
[1,2,3] Department of Information Systems, Faculty of Computer Science, UPN "Veteran" Jawa Timur, Indonesia

Article Info

Article history:

Received 07 22, 2025
Revised 10 13, 2025
Accepted 11 30, 2025

Keywords:

Online Gambling
Support Vector Machine (SVM)
Named Entity Recognition (NER)
Risk Classification
Data Visualization

ABSTRACT

The phenomenon of online gambling in Indonesia has developed rapidly, posing serious social and economic threats. This thesis aims to classify and map online gambling activities based on digital news using the Support Vector Machine (SVM) algorithm and Named Entity Recognition (NER). Data were collected from the news portals Detik.com, Kompas.com, and Tribunnews from 2017 to 2024 through a web scraping approach. The research process included setup and library import, data upload, data exploration, data labeling according to Law No. 1 of 2023, data preprocessing, data filtering, location normalization and extraction, and location data cleaning. Subsequently, the SVM model was trained for risk classification and followed by prediction. Evaluation was conducted using accuracy and F1-score metrics to assess overall model performance and classification balance. Based on the evaluation results, the Normal SVM model demonstrated the best performance with an accuracy of 96.94% and an F1-score of 0.97. The findings indicate that the combination of NER and SVM effectively identifies the location and risk level of online gambling activities. This research is expected to contribute to law enforcement authorities and policymakers in their efforts to prevent and address online gambling activities in Indonesia.

Corresponding Author:

Wisnu Mukti Darwansah
Department of Information Systems
Universitas Pembangunan Nasional “Veteran” Jawa Timur
Surabaya, Indonesia
Email: wisnumuktid@gmail.com

1. Introduction

Digital news portals such as *Detik.com*, *Kompas.com*, and *Tribunnews* have become major sources of information for the public. According to the 2024 Reuters Institute report, 50% of Indonesian respondents accessed *Detik.com* in a given week, making it the most-visited online media, followed by *Kompas.com* (39%) and *Tribunnews* (28%)[1]. The high level of public trust in these platforms positions them as key players in delivering current issues, including the online gambling phenomenon, which frequently garners media attention. However, the ever-growing volume of articles from multiple news sources often makes it difficult for readers to locate specific information. In this context, the implementation of text mining to integrate data from various news portals offers a promising solution to extract information, identify location entities, map affected regions, and classify the associated risks within online gambling news. This approach not only produces more structured information[2], but also supports data-driven analysis to provide meaningful insights and assist decision-making.

Online gambling is a serious issue in Indonesia, as it violates both social and religious norms and poses economic risks due to large amounts of money flowing out anonymously[3][4]. According to the 2023 annual report by the Financial Transaction Reports and Analysis Center (PPATK), there were 168 million online gambling transactions with a total value of IDR 327 trillion significantly increasing from just 250,000 transactions worth IDR 2 trillion in 2017. Furthermore, around 3.29 million Indonesians were reportedly involved in online gambling, often using accounts obtained through identity abuse or account trading to conceal illegal activity. This is further exacerbated by the transfer of funds abroad, creating challenges for law enforcement[5]. In this situation, text mining technology offers a potential solution to analyze news content, classify risk levels, and map affected regions[6], thereby supporting enforcement and offering strategic insights for policy-making.

Location mapping in the context of online gambling plays a vital role in understanding affected regions. Text mining-based mapping enables the visualization of specific areas identified as risk centers for online gambling[7]. By utilizing Named Entity Recognition (NER) to extract location entities from text and integrating this data with mapping techniques, the analysis becomes more precise and targeted[8]. For law enforcement, this location mapping can serve as a reference for prioritizing operations in high-risk areas. Additionally, it helps the government in formulating data-driven prevention and enforcement policies.

Text mining is a data analysis technique used to extract meaningful information from unstructured text, such as digital news or social media posts [9]. In this study, text mining is applied to analyze news articles related to online gambling. The Support Vector Machine (SVM) algorithm is used to classify risk levels. Compared to manual methods, text mining allows a more efficient analysis process and yields accurate risk mapping, which can serve as the foundation for more targeted decision-making [8].

Previous studies have demonstrated the effectiveness of NER and SVM in analyzing unstructured text data, particularly for extracting and classifying geographic information [8]. In those studies, NER was employed to identify location entities from Twitter texts such as city names, provinces, or specific places which were then mapped geographically. Meanwhile, SVM was used to classify disaster types based on textual patterns, achieving high accuracy levels of up to 95%. This combination enables the transformation of unstructured text from news or social media into structured and actionable information. These findings provide a solid foundation for adopting similar methods in analyzing digital news related to online gambling, enabling more efficient mapping and risk classification of such illegal activities [8].

NER and SVM were selected in this study due to their suitability for text-based analysis. NER is used to extract important entities from unstructured text, such as names, locations, and organizations, which are crucial in identifying affected regions and relevant parties in the context of online gambling [10]. The extracted data is then classified using SVM, an algorithm known for its consistent performance in text classification tasks. SVM identifies the optimal hyperplane to separate classes, enabling the classification of risk levels into low, medium, or high categories with high accuracy. Previous studies reported average SVM accuracy of 90.72%, making it an efficient method, especially for large datasets [11]. The combination of NER and SVM ensures structured, accurate analysis results that can support affected-area mapping and informed decision-making.

Therefore, this study aims to analyze the online gambling phenomenon using a text mining approach, specifically by applying NER and SVM methods to generate structured and meaningful information. The research seeks to extract key entities such as location from digital news articles and classify the risk level of online gambling activity into low, medium, and high categories. Furthermore, this study aims to map affected regions to provide deeper insights into the risk distribution of online gambling across Indonesia. This approach is expected to support law enforcement and data-driven policy formulation, contributing to the mitigation of the social and economic impacts of such illegal activities. The data used in this study spans from 2017 to 2024, based on annual reports by PPATK, which highlight the growing prevalence and impact of online gambling in Indonesia.

2. Research Method

The research method consists of a structured series of steps designed to ensure that the thesis is conducted systematically, logically, and in a way that can be clearly understood by readers. Each stage in the process is arranged to guide the researcher from the initial identification of the problem through data collection, analysis, and interpretation of results. Figure 3.1 illustrates the complete research workflow, providing a visual representation of the sequence of activities carried out throughout the study. This diagram helps clarify how each step is interconnected and how the process progresses toward achieving the research objectives. By following these structured stages, the researcher is able to maintain consistency, ensure methodological rigor, and produce findings that are valid and reliable. The outlined procedure also serves as a reference for future researchers who may wish to replicate or extend the study.

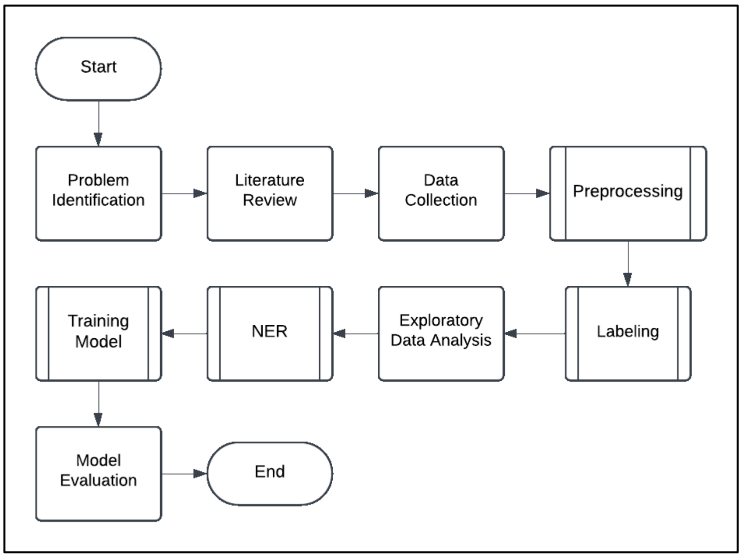

Figure 1. Research Methodology

2.1. Data Collection

Data collection was carried out using a web scraping method targeting three online news portals: *Kompas.com*, *Detik.com*, and *Tribunnews*. These sources were selected because they are mainstream media outlets that actively report on social and criminal issues, including the phenomenon of online gambling, which is the focus of this study. The keyword used for data retrieval was *"judi online"* (online gambling), with a data collection range spanning from 2017 to 2024. The data retrieved from each site includes the news headline and publication date, depending on the structure and availability of content on each website[12].

The scraping process was implemented using the Python programming language along with several supporting libraries, including Requests, BeautifulSoup, and pandas. The Requests library was used to send HTTP requests to the target websites, while BeautifulSoup was employed to extract relevant information from HTML elements containing news data. The pandas library was utilized to manage the scraped data by converting it into a DataFrame, which was then exported to an Excel file as the final storage format. To minimize the risk of being blocked by the server due to excessive requests in a short time, each scraping iteration included an automatic delay[13].

All data obtained through scraping was used as the main dataset for further analysis, including data cleaning, feature extraction, classification, and visualization. This study focuses on news headlines, as they typically represent the core content of the article and exhibit more consistent structure across different media platforms. The collected headlines served as the input for various text mining processes, such as keyword analysis, affected-area mapping, and risk-level classification. This approach was chosen to simplify the analytical process without compromising the main objective of the study, which is to identify patterns and map the risks associated with the spread of online gambling in Indonesia.

2.2. Preprocessing Data

The data preprocessing stage involves several essential steps to clean and prepare the text data for further analysis. This includes cleansing operations such as removing numbers, punctuation, and excessive spaces. These steps are crucial for improving the quality and consistency of the data. Numerical characters and punctuation are removed as they do not contribute meaningfully to text classification, and excess whitespace is eliminated to ensure cleaner structure. All of these processes are conducted using the re library in Python with the sub function[14].

Next, case folding is applied to standardize all characters into lowercase to avoid discrepancies between words with different letter cases, such as “Data”, “data”, and “DATA”, which should be treated as the same token. Tokenization then breaks down the text into smaller components or tokens (typically words), which serves as the foundation for further text analysis. In this study, tokenization is performed using the NLTK library, which efficiently splits sentences into individual words based on Indonesian language context.

The final preprocessing step is stopword removal, which eliminates common words that frequently appear in text but carry little meaningful information, such as "and", "to", "from", or "that". By removing these stopwords, the dimensionality of the data is reduced and the focus is placed on more relevant keywords. This step is vital for improving the performance of text mining processes like classification and keyword extraction, ultimately enabling more effective analysis of online gambling news content[15].

2.3. Labeling

At this stage, dataset labeling is carried out to assist the risk classification process according to the applicable laws and regulations. The labeling method used is direct-based learning, which involves aligning the data with a predefined list of keywords. To build a labeling system based on direct-based learning, several steps are required, such as manual labeling, calculating word weights for each class, and determining suitable keywords for direct-based learning. The process is explained as follows:

A. Manual Labeling

In the initial phase, manual labeling is performed by the author on the raw dataset as a foundational step to build a direct-based learning (DBL) system. This process involves assigning risk labels (low, medium, high) to news articles manually, based on the interpretation of the content in accordance with the legal context, as referenced in Appendix 1. The results of this manual labeling are then used as the initial reference data in the word weighting process using the TF-IDF method.

After the weighting is completed, words with the highest weights in each risk class are identified and extracted. These words are then used as the keyword list for the DBL system, allowing other news articles to be labeled automatically based on the appearance of these words. Thus, manual labeling serves not only as an initial step but also as a foundation for developing a more systematic rule-based automatic labeling system.

B. Word Weighting for Each Class

Following manual labeling, the next step is to calculate the weight of each word in each risk class using the Term Frequency–Inverse Document Frequency (TF-IDF) method. The purpose of this stage is to identify words that have high importance (high weight) in distinguishing between classes, namely low, medium, and high risk[16].

In this process, the manually labeled dataset is grouped based on its class. For each class, the term frequency (TF) of words is calculated within documents belonging to that class. Then, TF is multiplied by the Inverse Document Frequency (IDF), which is the logarithm of the ratio of the total number of documents to the number of documents containing the word. The result is a TF-IDF score for each word in each class.

Words with the highest TF-IDF scores are ranked in each class. However, not all high-scoring words are used directly. The final selection of keywords is filtered based on their relevance to applicable laws, particularly those regulating online gambling activities in Indonesia. This ensures that the keywords used in the DBL system are not only statistically relevant but also legally valid.

C. Automatic Labeling

In the automatic labeling phase, a direct-based learning (DBL) approach is used as the main method. The DBL system works based on a keyword list obtained from the previous steps: manual labeling and TF-IDF weighting for each risk class. These curated keywords are input into the system as references for recognizing word patterns that indicate specific risk levels in the news articles[17].

In implementation, the system reads and analyzes each news title and matches it with the available keyword list. If the title contains keywords corresponding to a specific risk class, the system will automatically assign the appropriate label. This method enables efficient and consistent labeling of large amounts of data without continuous manual intervention[18]. However, the system is designed to remain flexible and can be further developed by incorporating sentence context or integrating machine learning classification approaches.

2.4. Named Entity Recognition (NER)

The NER stage involves three key steps before and after location extraction. Before extraction, entity normalization is performed to convert general terms like place names or generic labels into specific and appropriate location names. After extraction, a data cleaning step is conducted to ensure that only valid and meaningful locations are retained for further processing.

A. Location Normalization

Location normalization is a crucial step carried out before automatic entity extraction. This step aims to standardize the spelling of region names in the news data to match proper forms. Often, place names appear in abbreviated, incomplete, or varied spellings that hinder accurate identification. Therefore, these names are converted into complete forms to ensure proper recognition and processing.

This normalization is applied to all labeled news titles, resulting in more consistent and accurate location extraction during the NER phase. Normalization also supports geographic mapping, as standardized names can be converted directly into coordinates using geolocation systems. It minimizes interpretation errors and enhances the quality of visual and spatial analysis[19].

B. Location Extraction

NER is used to extract relevant information for classifying text from a document, such as person names, locations, organizations, dates, times, and more. In this study, NER is used to extract location names related to activities with low, medium, and high-risk levels. The data used for NER has undergone preprocessing, such as removing numbers followed by periods, special characters, and excessive spaces, as well as location normalization to improve extraction accuracy[20].

This study uses the spaCy library with an Indonesian NER model, IndoBERT, adapted from the BERT architecture (Bidirectional Encoder Representations from Transformers) and specifically trained on Indonesian corpora. Its strength lies in its ability to understand bidirectional sentence context, making it more accurate for recognizing entities in Indonesian than standard models. IndoBERT is combined with spaCy to extract location entities, focusing solely on locations related to risk incidents[21]. Table 3.1 outlines the recognized entity types.

Table 1 Entity Types in spaCy

No	Entity Type	Description
1	PERSON	Person names
2	NORP	Nationalities, political groups, etc.
3	FAC	Buildings, airports, bridges, etc.
4	ORG	Companies, agencies, institutions
5	GPE	Countries, cities, provinces
6	LOC	Non-GPE locations
7	LANGUAGE	Language names

C. Location Data Cleaning

After the location normalization stage—where abbreviations and variations of regional names are standardized into their proper forms (for example, converting “jatim” to “Jawa Timur”)—the next step is to perform location entity extraction using Named Entity Recognition (NER). This process identifies and extracts location-related terms from the text and stores them in the extracted_locations column. However, the raw output from NER may still contain noise or inconsistencies, such as repeated locations, unnecessary whitespace, misspellings, or entities that do not correspond to valid geographic locations.

To address these issues, a comprehensive location data cleaning process is then applied. This step involves removing duplicate entries, eliminating null or empty values, correcting formatting inconsistencies, and ensuring that each extracted location matches recognized geographic names. In some cases, additional validation may be performed using external references or rule-based filtering to ensure accuracy.

The final output of this cleaning stage is a refined list of unique, standardized, and reliable location entities that can be used effectively in the subsequent stages of the research pipeline. This step is essential because clean and consistent location data supports more accurate geographic analysis, improves classification performance, and ensures that the results can be properly visualized and interpreted in later stages of the study..

2.5. Model Training

In this stage, the model is trained by dividing the dataset into training and testing subsets. The training portion is used to fit the Support Vector Machine (SVM) algorithm, enabling the model to learn patterns and relationships within the data, while the testing portion is used to evaluate its performance on previously unseen samples [22]. To ensure reliable and comparative results, two different data-splitting schemes are applied, namely 80:20 and 70:30. Each scheme is further tested under four different scenarios, resulting in a total of eight experimental tests. These scenarios are designed to analyze how varying preprocessing conditions and feature configurations affect the model’s accuracy and overall classification performance.

A. Training and Testing Data Split

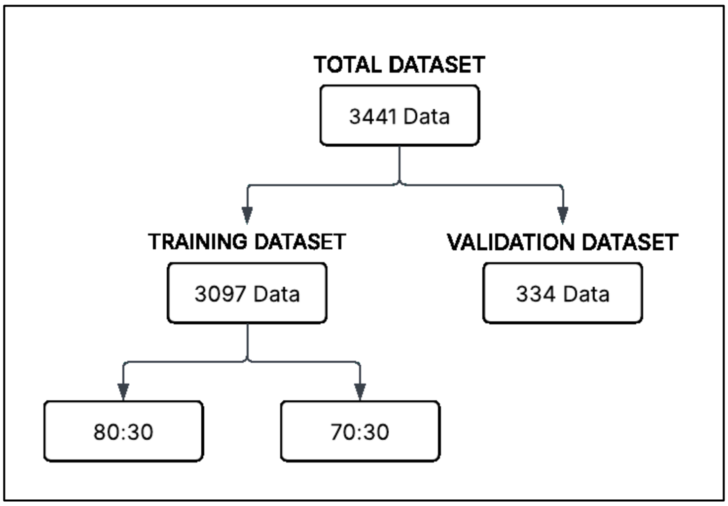

Figure 2. Split Data

Figure 2 illustrates the dataset splitting process. Out of 3,441 news items, 3,097 are used for model training and validation, while 334 are reserved as a final validation set to evaluate the model's generalization performance.

The 3,097 training items are split into two scenarios: one with an 80:20 ratio and the other with 70:30. Each scenario is used to train models with variations such as TF-IDF and oversampling (SMOTE), allowing performance comparisons across combinations.

B. Classification Model Development

The four classification scenarios used to train the SVM model are summarized in Table 2:

Table 2 SVM Classification Scenarios

Scenario	Description
SVM CountVectorizer	Uses bag-of-words representation (word frequency) via CountVectorizer, fed into SVM. Effective for texts with repeated word patterns.
SVM + TF-IDF	Applies CountVectorizer followed by TF-IDF weighting, then trains SVM. This accounts for word importance across the corpus.
SVM Oversampling	Applies oversampling (e.g., SMOTE) to balance class distribution before training SVM. Helps avoid bias toward majority class.
SVM Oversampling + TF-IDF	Combines oversampling and TF-IDF transformation before training SVM. Ensures both balanced data and focus on informative words.

The comparison aims to determine which combination yields the best model for classifying news related to online gambling risk. Table 2 outlines four different scenarios used for classifying online gambling risk in news articles using Support Vector Machine (SVM). The first scenario, SVM CountVectorizer, employs a simple bag-of-words approach, where word frequencies are captured using CountVectorizer and then passed into the SVM model. This method is particularly effective for recognizing repeated word patterns across the text. The second scenario, SVM + TF-IDF, enhances this by applying TF-IDF weighting, which considers the importance of each word within the overall corpus. This helps the model focus more on informative words rather than just frequent ones[23].

The third scenario, SVM Oversampling, addresses the issue of class imbalance by applying an oversampling technique such as SMOTE before training the SVM. Oversampling increases the number of minority class samples, reducing model bias toward the majority class and improving classification fairness.

The fourth and most comprehensive scenario, SVM Oversampling + TF-IDF, combines both approaches: it balances the dataset using oversampling and applies TF-IDF transformation to emphasize meaningful words. This scenario aims to ensure both balanced training data and rich textual representation, potentially leading to the best overall performance[24].

2.6. Model Evalution

The performance of the classification model is evaluated using a confusion matrix, which compares the predicted labels against the actual ground truth labels. This matrix provides the foundation for calculating various performance metrics that are essential in assessing how well the model performs in classifying news related to online gambling risk. Among these metrics are recall, precision, and accuracy, which each offer different perspectives on the model's effectiveness[25].

Recall measures the proportion of relevant instances that are correctly identified by the model, highlighting its ability to retrieve all relevant documents an important aspect when minimizing false negatives is critical. Precision, on the other hand, evaluates the proportion of retrieved instances that are actually relevant, thereby indicating the reliability of positive predictions. Accuracy provides an overall measure of correctness by calculating the percentage of total predictions that are correct. Together, these metrics offer a comprehensive evaluation of the model’s performance.

3. Result and Discussion

This section presents the results of the study, covering data collection, preprocessing, labeling, and classification using machine learning models. All processes are supported by Python programming with relevant libraries such as BeautifulSoup, Pandas, Scikit-learn, and Transformers.

3.1. Data Collection through Web Scraping

News articles were collected using a custom Python-based web scraping tool targeting three major Indonesian news portals: Detik.com, Kompas.com, and Tribunnews.com. The scraping algorithm employed the requests and BeautifulSoup libraries to access and parse HTML structures. Articles were retrieved based on the keyword *"judi online"*(online gambling) across the years 2017–2024.

Each entry includes the article title, summary, publication date, and channel. To avoid being blocked by the server, the program introduced randomized time delays between requests. A total of 3,096 valid articles were collected and saved into structured Excel/CSV files for further processing.

3.2. Text Preprocessing

To prepare the data for text analysis, a comprehensive preprocessing pipeline was implemented. The steps included:

- Duplicate removal
- Text cleaning (removal of numbers, punctuation, extra whitespace)
- Normalization (case folding, tokenization, stopword removal using Indonesian stopword list)
- Stemming using the Sastrawi library

The output was a clean, standardized corpus of article titles stored in CSV format, reduced to 2,991 records. Table 1 illustrates sample results after preprocessing.

Table 3. Sample Results of Text Preprocessing

Raw Title	Cleaned Title
Menteri HAM Wanti-wanti Jajarannya Tak Main Judi Online	menteri ham wantiwanti jajarannya main judi online
Empat Juta Pengguna Internet Main Judi Online	juta pengguna internet main judi online

3.3. Risk Labeling

3.3.1. Manual Annotation

An initial manual labeling was conducted on the cleaned dataset to categorize articles into High, Medium, and Low risk levels based on the content and context, referencing Indonesian Law No. 1 of 2023. For instance:

- High Risk: involves syndicates, financial crimes, or large-scale gambling (e.g., underground networks)
- Medium Risk: user involvement or minor law violations (e.g., casual users caught)

- Low Risk: promotional content or influencers (e.g., ads without active participation)

3.3.2. TF-IDF-Based Analysis

To identify key terms for each class, the Term Frequency–Inverse Document Frequency (TF-IDF) method was applied. Titles were grouped by label and tokenized into unigrams. The top keywords per risk level were extracted and ranked.

Table 4. Top TF-IDF Keywords per Risk Category

Risk Level	Top Keywords
High	bandar, situs, duit, buron, tangkap
Medium	pelaku, diperiksa, pria, usia
Low	endorse, promosi, selebgram, tiktok

3.3.3. Rule-Based Risk Classification (Direct-Based Learning)

Using the TF-IDF keywords and legal justifications, a rule-based system (Direct-Based Learning) was constructed for automatic labeling. Keywords were mapped to risk levels, and a Python function applied hierarchical matching logic. This automated approach labeled the dataset efficiently and consistently.

Tabel 5. Sample Auto-Labeled Data

Title	**Risk Label**
Artis Promosi Judi Online	Low
Polisi Tangkap Sindikat Judi Online	High

3.4. Named Entity Recognition (NER)

3.4.1. Location Normalization

After the risk labeling stage, the next step is location normalization to ensure consistency in writing region names. This involves converting abbreviated forms like "Jkt" into their full versions such as "Jakarta" or "Sby" into "Surabaya". The normalization process is crucial for accurate and consistent location extraction and mapping.

The normalization is performed using a dictionary (normalisasi_dict) that maps abbreviations to their full names. The dataset used in this process is the labeled data stored in labelled_preprocessed_result.csv. A function is applied to the ‘title’ column to create a new column, ‘title_normalized’, containing the normalized region names. This new column helps maintain both raw and cleaned versions of the data, which is then saved in normalized_location_result.csv. This process ensures uniformity, reduces errors, and facilitates more precise risk analysis.

3.4.2. Location Extraction

Named Entity Recognition (NER) is used to extract location entities from the normalized text using the IndoBERT pretrained model (cahya/bert-base-indonesian-NER) provided by the Transformers library. The system checks for GPU availability to optimize processing speed.

The dataset is processed in batches (100 records per batch) to manage memory usage. A dedicated function extracts location entities such as provinces, cities, or sub-districts using entity tags like B-LOC, I-LOC, and LOC. These extracted locations are added as a new column (extracted_locations) in the dataset and saved in location_extracted_result.csv.

At the end of this process, the system calculates the number and percentage of records with extracted location entities. Sample outputs are also displayed for verification.

3.4.3. Location Data Cleaning

The final stage involves cleaning the extracted location data to ensure validity and remove inconsistencies. The goal is to retain only meaningful and accurate location names.

A function is applied to the extracted_locations column to strip unnecessary spaces, remove duplicates, and retain unique entries in their original order. The cleaned results are stored in a new column called cleaned_locations.

The script also compiles a list of all unique locations from the cleaned data and saves them in two files:

- cleaned_location_result.csv (full cleaned dataset)
- unique_locations.txt (alphabetically sorted list of unique locations)

Statistics on the number of valid location entries and the total number of unique locations are generated to evaluate the success of the location extraction process.

3.5. Model Training and Evaluation

To classify the risk levels of online gambling news, the Support Vector Machine (SVM) algorithm was employed. Preprocessed titles were vectorized using both CountVectorizer and TF-IDF methods. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples in the minority classes.

Model evaluation was conducted using two train-test splits: 80:20 and 70:30. Each configuration was tested under both normal and oversampled datasets. The performance metrics used include Accuracy, Precision, Recall, and F1-Score(macro average).

Table 6. Performance Comparison of SVM Models

Dataset	Configuration	Accuracy	Precision	Recall	F1-Score
Normal	**SVM + CountVectorizer (80:20)**	**0.98**	**0.98**	**0.96**	**0.97**
	SVM + TF-IDF (80:20)	0.96	0.96	0.94	0.95
	SVM + CountVectorizer (70:30)	0.98	0.98	0.95	0.97
	SVM + TF-IDF (70:30)	0.95	0.95	0.92	0.93
SMOTE	SVM + Oversampling (80:20)	0.88	0.85	0.90	0.87
	SVM + TF-IDF + SMOTE (80:20)	0.97	0.97	0.97	0.97
	SVM + Oversampling (70:30)	0.86	0.79	0.88	0.82
	SVM + TF-IDF + SMOTE (70:30)	0.97	0.97	0.95	0.96

The results show that SVM with CountVectorizer yielded the highest accuracy (0.98), while TF-IDF combined with SMOTE provided the best balance across all evaluation metrics, particularly in improving recall and F1-score for minority risk classes. Therefore, the combination of TF-IDF and SMOTE is recommended for achieving more robust and fair classification performance in imbalanced datasets. Furthermore, the selected best-performing model will be further evaluated using a separate validation dataset to assess how well it generalizes to unseen data.

4. Conclusion

This study successfully addressed two main objectives: (1) the classification of online gambling activity risk in Indonesia based on digital news, and (2) the mapping of such activities using location-based information extracted from news articles. The risk classification was conducted by extracting named entities using Named Entity Recognition (NER) and categorizing them into three risk levels: low, medium, and high. The classification model utilized the Support Vector Machine (SVM) algorithm, with risk categories defined based on legal criteria from Indonesian law.

Secondly, this study successfully visualized the geographic distribution of online gambling activities by processing news data from Detik.com, Kompas.com, and Tribunnews.com, covering the period from 2017 to 2024. Through a series of stages—web scraping, preprocessing, entity extraction, and risk classification—the final output was presented in the form of an interactive map-based visualization. This map provides a clear overview of affected regions and can be used as a strategic decision-support tool by policymakers and authorities.

By integrating text mining, NER, and SVM classification, this research demonstrated that data-driven approaches can effectively process unstructured news texts into meaningful insights and spatial visualizations. The findings confirm that digital news analysis can play a significant role in understanding and managing the spread of online gambling, supporting more targeted and informed social interventions.

References

[1] Erlina F. Santika, “10 Media Online yang Paling Banyak Digunakan Warga Indonesia 2024,” https://databoks.katadata.co.id/media/statistik/4b024acf115a988/10-media-online-yang-paling-banyak-digunakan-warga-indonesia-2024.

[2] N. Nurchim, N. Nurmalitasari, and Z. A. Long, “Indonesian news classification application with named entity recognition approach,” *JURNAL INFOTEL*, vol. 15, no. 2, pp. 130–134, May 2023, doi: 10.20895/infotel.v15i2.909.

[3] Fidyan Hamdi Lubis, Melisa Pane, and Irwansyah, “Fenomena Judi Online di Kalangan Remaja dan Faktor penyebab Maraknya Serta Pandangan Hukum Positif dan Hukum Islam (Maqashid Syariah),” *Jurnal Pendidikan dan Konseling*, vol. 5, pp. 2656–2657, 2023.

[4] I. Tasya Jadidah *et al.*, “Analisis maraknya judi online di Masyarakat,” 2023.

[5] Adi Ahdiat, “Judi Online Kian Marak, Transaksinya Tembus Ratusan Triliun,” https://databoks.katadata.co.id/ekonomi-makro/statistik/2bdcd34bb7533f5/judi-online-kian-marak-transaksinya-tembus-ratusan-triliun.

[6] S. Supriyatna and E. Fahrudin, “PEMANFAATAN ALGORITMA TEXT MINING DALAM MENEMUKAN POLA RISIKO BENCANA SEBAGAI PENGETAHUAN KEBENCANAAN DARI DOKUMEN KAJIAN RISIKO BENCANA (KRB) 1*,” *Jurnal Informatika Utama*, vol. 2, no. 1, 2024, doi: 10.55903/jitu.v2i1.xx.

[7] W. Saefudin, A. Komarudin, and R. Ilyas, “Visualisasi Kumpulan Berita Dalam Bentuk Peta Digital Dengan Metode Term Frequency-Inverse Document Frequency dan Gazetteer,” *Seminar Nasional Sains dan Teknologi Informasi (SENSASI)*, Jul. 2019, [Online]. Available: http://prosiding.seminar-id.com/index.php/sensasi/issue/archivePage|665

[8] A. Suganda Girsang and B. Krisna Noveta, “Location Prediction using Named Entity Recognition for Indonesia Natural Disasters in Data Twitter,” *Elsevier*, Nov. 2022, [Online]. Available: https://ssrn.com/abstract=4276345

[9] A. Adhitama, S. Hidayatullah, and M. Rahman, “Klusterisasi Judul Berita Pada Website Detik Menggunakan Algoritma Kmeans,” *Indonesian Journal of Innovation Science and Knowledge*, vol. 1, Jul. 2024.

[10] S. Aliff, S. Ramadhani, A. Rahmanqa, D. Nur, and A. Rakhmawati, “Deteksi Lokasi Siswa SMP di Instagram dengan Metode Named Entity Recognition,” *Jurnal Sosial dan Teknologi (SOSTECH)*, vol. 1, no. 7, 2021, [Online]. Available: https://greenvest.co.id/

[11] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, “Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.

[12] C. C. Aggarwal and C. Zhai, “Mining Text Data,” *New York: Springer*, 2012.

[13] Y. A. Hafiz and E. Sudarmilah, “IMPLEMENTASI WEB SCRAPING PADA PORTAL BERITA ONLINE,” *Inisiasi*, vol. 12, 2023, doi: https://doi.org/10.59344/inisiasi.v12i1.120.

[14] H. Kaur and A. Saini, “A Review of Data Preprocessing Techniques for Text Mining,” *Int. J. Comput. Appl*, vol. 103, 2014.

[15] S. García, J. Luengo, and F. Herrera, “Intelligent Systems Reference Library 72 Data Preprocessing in Data Mining,” cham, 2015. doi: 10.1007/978-3-319-10247-4.

[16] I. S. Wibowo, A. Witanti, and I. Susilawati, “Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF”, [Online]. Available: http://jurnal.mdp.ac.id

[17] H. Zhao, X. Li, F. Wang, Q. Zeng, and X. Diao, “Incorporating keyword extraction and attention for multi-label text classification,” *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 2, pp. 2083–2093, Aug. 2023, doi: 10.3233/JIFS-230506.

[18] Y. Zhang, M. Jiang, Y. Meng, Y. Zhang, and J. Han, “PIEClass: Weakly-Supervised Text Classification with Prompting and Noise-Robust Iterative Ensemble Training,” Oct. 2023.

[19] Z. Qiang, K. Taylor, and W. Wang, “How Does A Text Preprocessing Pipeline Affect Ontology Syntactic Matching?,” Nov. 2024.

[20] B. Jehangir, S. Radhakrishnan, and R. Agarwal, “A survey on Named Entity Recognition — datasets, tools, and methodologies,” *Natural Language Processing Journal*, vol. 3, p. 100017, Jun. 2023, doi: 10.1016/j.nlp.2023.100017.

[21] J. Ortega and G. Brotosaputro, “Analisis Sentimen Tokoh Politik pada Situs Berita Menggunakan NER. Studi Kasus: IMMC,” *Prosiding SISFOTEK*, vol. 3, Oct. 2019.

[22] K. Lee *et al.*, “Deduplicating Training Data Makes Language Models Better,” Jul. 2021.

[23] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, “A survey on imbalanced learning: latest research, applications and future directions,” *Artif Intell Rev*, vol. 57, no. 6, p. 137, May 2024, doi: 10.1007/s10462-024-10759-6.

[24] M. Mujahid *et al.*, “Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering,” *J Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[25] Robert Antonius, A. R. Zulkarnain, and H. Irsyad, “Pendekatan TF-IDF, SMOTE, dan SVM dalam Klasifikasi Sentimen Masyarakat terhadap Pemblokiran Judi Online,” *Buletin Ilmiah Informatika Teknologi*, vol. 2, no. 3, pp. 115–122, Jun. 2024, doi: 10.58369/biit.v2i3.65.