SENTILYZE: A BILINGUAL SENTIMENT ANALYSIS API USING LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINE ALGORITHMS

Juan Gabriel Baterina[1*], Jared Castañeda[2], Stephen Grant Sumadsad[3]

School of Computer Studies and Technology Colegio de San Juan de Letran Calamba , Bucal Bypass Rd, Calamba, 4027 Laguna, Philippines

Article Information

Received: 21-11-2024
Revised: 28-11-2024
Published: 05-12-2024

Keywords
Bilingual Sentiment Analysis; Logistic Regression; Support Vector Machine (SVM); Natural Language Processing (NLP)

****Correspondence Email:***
juan@gmail.com

Abstract

This study introduces a Bilingual Sentiment Analysis API employing Logistic Regression and Support Vector Machine (SVM) algorithms. Using descriptive research methods, the study focuses on the administrative personnel of the External Relations Department (ERD) at Colegio de San Juan de Letran Calamba, with the Director as the primary respondent. Data were collected through interviews and social media content, particularly Facebook posts and comments. The API development followed the Agile Model within the Software Development Life Cycle (SDLC). Results highlight the proposed model's precision and reliability, with Logistic Regression delivering accuracy and SVM effectively capturing sentiment patterns. Combined, these algorithms achieved a 90.19% accuracy rate across five recorded tests, showcasing the efficiency of automated sentiment analysis in reducing processing time and providing real-time insights. This tool aids the ERD in online reputation management and strategic decision-making by interpreting sentiment data. Furthermore, the research advances natural language processing by addressing sentiment analysis challenges in social media contexts and offering a practical solution for sentiment interpretation. The Bilingual Sentiment Analysis API provides a powerful resource for improving the ERD's capabilities in utilizing sentiment data to enhance communication strategies and operational efficiency.

1. Introduction

The Department of Education highlights the importance of technology in education, particularly during the pandemic, with platforms like Facebook playing a crucial role in communication and learning. At Colegio de San Juan de Letran Calamba, the External Relations Department (ERD) faces challenges in manually monitoring and analyzing the growing volume of social media comments, limiting their ability to manage online reputation effectively. To address these inefficiencies, this study proposes a Bilingual Sentiment Analysis API leveraging Logistic Regression and Support Vector Machine (SVM) algorithms, which are recognized for their accuracy and robustness in sentiment classification. Combining these algorithms enhances sentiment analysis performance, especially in bilingual contexts, as demonstrated in prior research. The API aims to streamline data collection, provide real-time insights, and improve sentiment analysis workflows, empowering the ERD to make informed decisions and strengthen engagement strategies on social media.

The External Relations Department (ERD) of Colegio de San Juan de Letran Calamba faces challenges in analyzing sentiment data from Facebook, crucial for managing the institution's reputation and engagement strategies. The current manual process is inefficient and delays decision-making. To address this, the study proposes developing an automated Bilingual Sentiment Analysis API using Logistic Regression and Support Vector Machine algorithms, aiming to streamline data analysis, provide real-time insights, and enhance online reputation management. Key research questions focus on identifying challenges, evaluating the API's impact, and addressing inefficiencies in current practices.

This study aimed to develop and implement a Bilingual Sentiment Analysis API utilizing Logistic Regression and Support Vector Machine algorithms to automate sentiment analysis of Facebook data related to Colegio de San Juan de Letran Calamba. It sought to enhance the External Relations Department's ability to analyze sentiment trends, identify issues, and improve the Colegio's online reputation management. Specific objectives included optimizing the model for reliability, comparing its efficiency and accuracy with manual analysis, and evaluating its effectiveness in automating sentiment analysis for social media data.

The significant value for various beneficiaries by developing a Bilingual Sentiment Analysis API that automates the analysis of social media data, particularly Facebook posts related to Colegio de San Juan de Letran Calamba. For the institution, it offers actionable insights to enhance online reputation management and stakeholder engagement. The External Relations Department (ERD) benefits from streamlined data gathering, resource savings, and real-time sentiment insights, improving decision-making processes. Students, faculty, and staff may experience a morale boost from improved institutional reputation. Additionally, future researchers and practitioners in sentiment analysis and machine learning can leverage the methodologies and findings to advance their work, while other educational institutions can adapt the study's approach to address similar challenges in managing online reputations and engagement.

The study focused on analyzing Facebook posts and comments related to Colegio de San Juan de Letran Calamba, aiming to provide insights into sentiments expressed by the online community. Data collection targeted a substantial dataset by 2023; however, challenges such as data availability and access restrictions may have affected the quantity and representativeness of the data. The study was limited to Facebook as the primary platform, excluding sentiments from other social media channels and private or restricted posts. Sentiment analysis relied on Logistic Regression and Support Vector Machine algorithms, which, while effective, could encounter misclassifications due to complexities like sarcasm or irony. Additionally, the scope was confined to sentiment classification without exploring underlying motivations or factors driving the sentiments. Despite these constraints, the research utilized rigorous methodologies to enhance sentiment analysis accuracy and contribute to understanding online perceptions of the institution.

1.1 Literature Review

This chapter contains the related readings, related literature, and related studies categorized in local and foreign areas. This serves as the basis and foundation and will contain a detailed background of Bilingual Sentiment Analysis using Logistic Regression Algorithm, through the social media platform 'Facebook', on Colegio de San Juan de Letran Calamba. Definitions and uses of Sentiment Analysis, through 'Facebook', will be tackled. This will give a better understanding to the readers as this aims to provide clarifications and clear any vague ideas regarding the study.

1) Facebook
   Facebook is a widely accessible social media platform where users can create profiles, connect with others, and share various content like pictures and posts. It supports multilingual interactions and had 80.3 million users in the Philippines by early 2023, with a high percentage of the population exposed to advertisements on the platform.
2) Sentiment Analysis
   Sentiment analysis uses natural language processing (NLP) to determine the emotional tone of a text, classifying it as positive, negative, or neutral. It is valuable in understanding public opinion, especially on social media platforms like Facebook, where unsolicited feedback provides insights into user perspectives.
3) Machine Learning
   Machine learning enables software to learn from data and make decisions without explicit programming. By analyzing patterns in data, machine learning models improve over time, offering insights that enhance decision-making. The success of these models depends on the quality of the data used for training.
4) Natural Language Processing (NLP)
   NLP combines linguistics and AI to enable machines to understand and interpret human language. In sentiment analysis, NLP helps in discerning emotions from text, such as through word sense disambiguation, which determines the intended meaning of words based on context.
5) Naive Bayes Algorithm
   Naive Bayes is a probabilistic machine learning algorithm commonly used for text classification tasks. It calculates the probability of a text belonging to a particular category, with strengths in speed and accuracy, though its assumption of feature independence can sometimes limit its performance.
6) Logistic Regression
   Logistic regression is a statistical method used for binary classification tasks, making it ideal for sentiment analysis where texts are classified into two categories (e.g., positive or negative). It is valued for its simplicity, interpretability, and effectiveness when the data is linearly separable.

2. Research Methods

This study focuses on the administrative personnel of the External Relations Department (ERD) at Colegio de San Juan de Letran Calamba, particularly the ERD Director, who had administrative access to the Bilingual Sentiment Analysis API for monitoring Facebook sentiments. Other respondents included current employees and occasional part-timers with user access to the API, contributing to sentiment analysis and evaluation of the institution's Facebook page. The research aimed to investigate the API's usability and effectiveness, using this specific group of respondents to assess its performance and potential for improving social media management.

During the model development stage, researchers implemented logistic regression and support vector machine algorithms using the Scikit-learn library for sentiment analysis. Textual data were converted into numerical feature vectors through CountVectorizer or TfidfVectorizer to enable model training. The models were trained on a designated dataset to learn patterns and make predictions, with their performance evaluated using metrics such as accuracy, precision, recall, and F1-score. This stage provided comprehensive documentation of the algorithms used, feature extraction techniques applied, and evaluation results, offering valuable insights into the models' performance and their potential for sentiment analysis tasks.

Building on this foundation, the model refinement and optimization phase focused on enhancing generalization and avoiding overfitting through techniques like cross-validation. These efforts led to a fine-tuned logistic regression model with improved accuracy and robustness. The optimization process was meticulously documented, detailing fine-tuned hyperparameters and cross-validation strategies, ensuring reproducibility. This refinement not only strengthened the sentiment analysis system but also established a solid framework for future research and advancements in the field.

The integration and testing phase focused on incorporating the trained sentiment analysis model into the desired system while ensuring its reliability in real-world scenarios. Rigorous testing was conducted using both positive and negative test cases to evaluate the model's behavior across diverse data types and sentiment classifications. Researchers identified and addressed potential issues or bugs during this phase, ensuring robust performance. Detailed documentation covered the integration process, testing procedures, executed test cases, and any modifications made, culminating in a fully integrated and thoroughly tested sentiment analysis system ready for deployment.

Following successful integration, the model was deployed in the target environment, enabling its real-world application. To maintain its efficiency and reliability, monitoring mechanisms were established, including alert systems to promptly detect and address any performance issues or drifts. Regular evaluations were conducted to ensure consistent accuracy and effectiveness in various scenarios. The deployment process and monitoring setup were meticulously documented, providing insights into the deployed environment and the protocols established to uphold the model’s optimal performance during ongoing use. Here is a breakdown of the key materials and tools used in this research, along with their roles and significance in supporting the sentiment analysis process:

Python was chosen for its simplicity, versatility, and powerful libraries supporting data analysis, machine learning, and natural language processing, making it ideal for sentiment analysis tasks. The Facebook Graph API was used to extract relevant content and metadata, such as posts, timestamps, and user information, which served as the dataset for training and evaluating the sentiment analysis model. For data analysis, tools like NumPy were used for numerical computations, Pandas for organizing and manipulating data, Seaborn for creating visual plots, and Matplotlib for detailed visualizations. Scikit-learn was utilized to implement the logistic regression algorithm for sentiment analysis, providing a reliable framework for model training and evaluation. Visual Studio Code was the integrated development environment (IDE) used for code development, debugging, and version control, with plugin support and Git integration for collaboration.

The system requirements for running the API are minimal, focusing on compatibility with Google Chrome's software requirements. The necessary hardware includes Windows 10 64-bit OS, an Intel Pentium Core 4 CPU, 4GB of RAM for multitasking, and 256GB SSD storage for fast data processing.

Data was collected from Facebook posts using the Facebook Graph API, followed by integrity checks to ensure accuracy and relevance. Key steps included classifying data to extract meaningful words, implementing a Logistic Regression Algorithm for sentiment prediction, and training the model using labeled data to identify positive or negative sentiments. The final product, "Sentilyze," was successfully deployed within the department's system, accompanied by a tutorial on its operation.

3. Result and Discussion

The researchers present the findings of the research study, which delves into a bilingual sentiment analysis API using Logistic Regression and Support Vector Machine Algorithm. Through rigorous data analysis and interpretation, they explored the key patterns, trends, and insights that have emerged from the research, shedding light on the implications and significance of these results. This section aligns with the study's objectives and assesses the successful utilization and functionality of the developed Bilingual Sentiment Analysis API.

4. Conclusions

The study demonstrated the effectiveness of combining the Logistic Regression model and Support Vector Machine (SVM) in the development of *Sentilyze: Bilingual Sentiment Analysis API*. This combined approach achieved superior accuracy and reliability compared to standalone algorithms, particularly in a bilingual context. While addressing challenges faced by the External Relations Department (ERD) of Colegio de San Juan de Letran Calamba, *Sentilyze* also contributed valuable insights to the broader field of sentiment analysis. Despite limitations, such as the scarcity of non-lexicon Tagalog datasets, the research underscores the potential for future enhancements in bilingual sentiment analysis.

The evaluation revealed that the combined model outperformed manual analysis in efficiency and accuracy. The automated system reduced analysis time from 5 hours per month to seconds, achieving an overall accuracy of 90.19% compared to Logistic Regression's 72.33% and SVM's 74.2%. This efficiency enables real-time analysis, delivering timely and precise insights into social media sentiments. These findings highlight the value of integrating advanced algorithms to streamline sentiment analysis processes and optimize social media strategies for institutions like Letran Calamba.

5. References

Abro, A. A., Talpur, M. S. H., & Juman , A. . (2022). Natural Language Processing Challenges and Issues: A Literature Review. Natural Language Processing Challenges and Issues: A Literature Review. https://doi.org/10.35378/gujs.1032517 Access Date: July 20, 2023

Adi, A. C., Lestari, D. P., Elsa, Saputri, F. S., & Sabui, Y. (2022). Online School Sentiment Analysis in Indonesia on Twitter Using The Naïve Bayes Classifier and RapidMiner Tools. International Journal of Innovative Science and Research Technology. https://ijisrt.com/assets/upload/files/IJISRT22JAN782_(1).pdf

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. Procedia Computer Science, 152, 341–348. https://doi.org/10.1016/j.procs.2019.05.008

Alawadh, H. M., Alabrah, A., Meraj, T., & Rauf, H. T. (2023). Semantic FeaturesBased discourse analysis using deceptive and real text reviews. Information, 14(1), 34. https://doi.org/10.3390/info14010034

Aliman, G., Nivera, T. F., Olazo, J. C., Ramos, D. J., Sanchez, C. D., Amado, T., Arago, N., Jorda, R., Jr, Virrey, G., & Valenzuela, I. (2022). Sentiment Analysis using Logistic Regression. De La Salle University. https://www.dlsu.edu.ph/wpcontent/uploads/pdf/research/journals/jciea/vol-7-1/4aliman.pdf

Amrani, Y. A., Lazaar, M., & Kadiri, K. E. E. (2018). Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia 99 Computer Science, 127, 511–520. https://doi.org/10.1016/j.procs.2018.01.150