Harnessing Social Media Sentiment Analysis for Wildlife Conservation

Abdur Rashid Sangi *, Ma Zhen, Zhang Chi, Ye Nan, and Baha Ihnaini *

International Association for Neuro-Linguistic Programming, Department of Computer Science, College of Science, Mathematics and Technology, Wenzhou-Kean University, Wenzhou, Zhejiang, China
Email: rsangi@wku.edu.cn (A.R.S.); 1162217@wku.edu.cn (M.Z.); 1162683@wku.edu.cn (Z.C.); 1162287@wku.edu.cn (Y.N.); bihnaini@wku.edu.cn (B.I.)
*Corresponding author

***Abstract*—In today's digital age, social media platforms have become powerful tools for collecting public sentiment on various issues, including environmental conservation. This research employs data from Twitter, YouTube, TikTok, and Instagram to enhance the conservation efforts for endangered species through sentiment analysis. We collected and preprocessed a high-quality dataset from these platforms and applied multiple models to perform sentiment analysis. Among the models tested, Logistic Regression (LR) and Valence Aware Dictionary and sEntiment Reasoner (VADER) showed the highest accuracy rates. Key preprocessing steps included cleaning emojis, slang, and non-English text to standardize the input data. Our results suggest that social media can be a strategic asset for conservationists by providing insights into public sentiment and engagement. Future work will focus on improving data processing techniques and exploring hybrid models to further boost the effectiveness of sentiment analysis in conservation efforts.**

***Keywords*—sentiment analysis, endangered species, environmental protection, logistic regression, Valence Aware Dictionary and sEntiment Reasoner (VADER)**

I. Introduction

The loss of biodiversity and endangered species is closely tied to priorities, economies, ecosystems, and human well-being around the world [1]. Field research is generally the way to protect these species through direct observation and conservation programs. In the digital age, social networking sites have also pioneered powerful tools to not only increase public awareness and engagement, but also to collect valuable data on a wide range of issues, including environmental conservation [2].

Social media is one of these powerful tools, which includes platforms such as Twitter, YouTube, TikTok, and Instagram, which are rich sources of user-generated content that reflect public sentiment, awareness, and engagement on environmental issues [3]. This data may contain deep insights into the views and attitudes of a wide audience towards endangered species. Analysis of this data will inform conservation strategies and, in fact, drive educational campaigns, leading to more effective conservation. The challenge remains in managing and analyzing the large amounts of unstructured data generated on these platforms. Handling the diversity and complexity of content on social media is very challenging for traditional data analysis methods.

This study collects a large amount of data on endangered species, all of which are sourced from many social media to obtain advanced data sets. Conservation of endangered species now becomes a pressing issue requiring creative ways of creating awareness and driving action. The dynamic and widespread nature of public sentiment and engagement on most traditional methods of data collection and analysis seems always to lack in regards to this. Social media channels, used extensively by vast masses of people and functioning in real-time, could provide immense opportunities to collect data on a large scale about public perception and awareness regarding endangered species.

However, managing this data comes with many difficulties. The first difficulty may be the huge amount of data posted on social media, which is likely to require extensive data cleaning and processing to ensure relevance and accuracy. Second, social media content is highly unstructured and varies greatly in tone, context, and sentiment, which complicates the tagging and analysis process. Third, traditional analytical models may not be sufficient to thoroughly mine the nuances of social media interactions. Therefore, this will limit the effectiveness of proactive action recommendation models.

To address these challenges, this study proposes to collect and analyze data from multiple social media sites, specifically Twitter, YouTube, TikTok, and Instagram. The intended approach is to create an extensive database on endangered species, process it for accuracy, and apply advanced machine learning models. Our focus on positive and negative sentiment aims to gain meaningful insights that can be used to support conservation efforts for endangered species.

II. Literature Review

Social media platforms have transformed how information is disseminated and discussed across global

Manuscript received June 22, 2024; revised July 14, 2024; accepted July 26, 2024; published November 8, 2024.

audiences, playing a pivotal role in environmental awareness and activism. Platforms such as Twitter, YouTube, and Instagram serve as vital tools for environmental advocacy, enabling rapid spread of information and mobilization of public support for conservation efforts [4]. Folke *et al.* [5] emphasize the utility of social media in capturing real-time, large-scale data that reflects public perceptions and behaviors towards environmental issues, providing a valuable resource for conservationists.

While social media offers a vast pool of data, the challenges of data volume, veracity, and variety cannot be understated. The need for robust data cleaning and preprocessing methods is crucial to ensure the relevance and accuracy of the data analyzed. Kietzmann *et al.* [6] discusses the complexities involved in social media analytics, emphasizing the importance of sophisticated tools and algorithms to manage and analyze this data effectively.

Our study illustrates the efficacy of multiple machine learning models in analyzing sentiment and generating actionable insights from social media content. This reflects a broader trend wherein AI and machine learning are increasingly applied to environmental sciences and conservation. The application of models such as Multinomial Naive Bayes, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM), as discussed by Zhou *et al.* [7], highlights the potential of these technologies to enhance conservation efforts by providing detailed and accurate analyses of public sentiment and engagement.

III. Methodology

A. Dataset

1) Collection and dataset build

Through a set of critical stages, such as data collection and data review, a profound dataset was made on the conservation of endangered species. Social media platforms were deemed as the most essential data source because many people generate various data on their posts. The solutions were found in social media—from Twitter to YouTube, TikTok, and Instagram.

Data generation through these sources received involvement of several necessary devices and techniques. They easily enabled you to extract Twitter pages directly and obtained much more high-quality data. This method was selected as the previous Application Programming Interface (API) test for Twitter indicated that there was a very small number of the tweets that were compatible with our requirements, so using the API was not commercially viable. Our YouTube approach used the tool YouTube Comment Downloader, an open-source tool, to gather video metadata and comments related to endangered species. On TikTok, we compiled the comments using the TikComments plugin. The data-collection method for Instagram used the Instagram Graph API to locate posts, captions, hashtags, and comments with reference to endangered species. These methods ensured high-quality data collection. Fig. 1 shows the proportion of data collected from each social media platform.

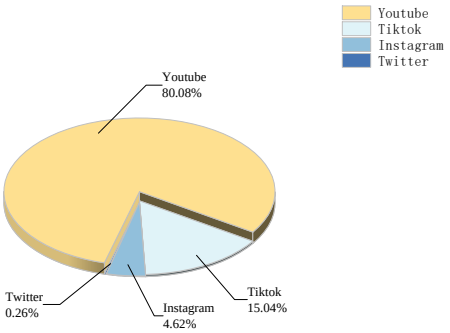

Fig. 1. Distribution of data sources for the dataset.

2) Preprocess

The processes are to make the dataset meet the requirement of sentiment analysis:

1. The different special emojis and garbled code are cleaned to ensure that the text data is standardized and treated uniformly when it is used to be in training [8]. The first step is applying the emojis library in Python to detect the emojis in text. Second, some special emojis that can't be detected by the emojis library, we translate these emojis to Unicode and match encoding range to identify emojis regular expression.
2. Numbers are cleaned, it often lacks semantic meaning in the context of natural language [8]. Unlike words or phrases, numbers typically do not carry rich contextual information. They provide numerical values but lack the nuanced semantics and associations present in language.
3. The text that is smaller than 7 characters are removed. Short text often prioritize brevity over clarity, leading to abbreviated or fragmented expressions. Understanding the intended meaning or sentiment behind such texts may require additional context or background knowledge, posing challenges for Natural Language Processing (NLP) models.
4. Empty intervals between a large amount of text, text that are not English. Our research is basing on English text. Other languages are not our target. First, powerful language detection library, like Langdetect, is applied to clean multilingual content accurately. Second, to assure every text, regular expression is used to detect by iterating Unicode of character in each text.
5. Slang is widely used in the comments of social media. The meanings of slang are very sentimental. Therefore, slang is translated in the paper. A basic dictionary concluded by ourselves about popular slang is made to translate slangs to specific text which is used to replace the slangs in text.

Due to the significant difference between the positive and negative labels in the initial data, we supplemented the data with negative labels through synonymous rewriting, thus achieving a balance between the amount of data with positive and negative labels. Fig. 2 illustrates the distribution of the dataset before and after balancing the classes.

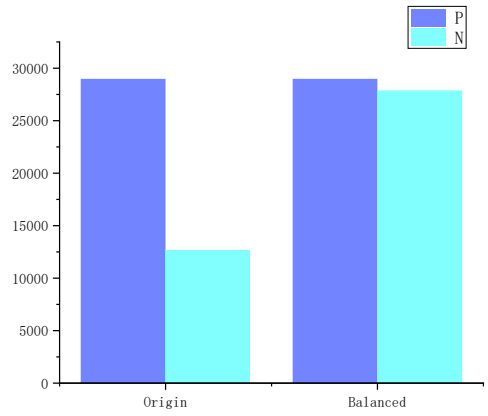

Fig. 2. Data distribution before and after balancing.

3) Kappa test

In order to examine the reliability of the labeling process for our dataset, we used the kappa coefficient, which measures the agreement level among multiple raters [9].

A random sample of 1,000 data points from the dataset, was taken to represent the total dataset heterogeneity. On each subtopic, the two annotators acted independently and followed such criteria as labeling the content as positive or negative.

The labels assigned by the two annotators were compared to calculate the Kappa coefficient (κ). The Kappa coefficient is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_0 is the observed agreement among raters: p_e is the expected agreement by chance; P_0 is the observed consistency; P_e is the expected consistency based on the distribution of labels assigned to each annotator.

The value of Kappa coefficient was obtained by means of collected data, which was κ = 0.793, showing a basic high-level agreement that the labeling process is reliable and that the guidelines provided to the annotators were clear and effective.

B. Splitting Data into Training Set and Test Set

It is usual to have a 4:1 ratio or more in the case of the training set and test set fashion that are used for training. First, 80% of the data makes the training fully enough, which helps the model to have high generalization ability. There are 20% of data used for the evaluation of the model, and that is to ensure that there are enough testing sets, which will help in seeing how well the model is performing on both unknown data and datasets [10]. It tests the generalization ability of the model by simulating actual application scenarios. Evaluating a model's performance on the test set can detect potential overfitting or under fitting issues in the model. As a team, we've also found that an 8:2 balance would provide a balance between expert training and case study review when data is limited.

C. Training Set Model Categories

In order to improve the reliability of the analysis and the availability of response data, the model classification in the training set includes the following categories:

1) Machine learning

a) Multinomial naïve bayes

A probabilistic classifier applying Bayes' Theorem with an assumption of independence between predictors [11].

b) Support Vector Machine (SVM)

A supervised learning model used for classification by finding the hyperplane that best divides the data points into classes.

c) Logistic regression

A statistical technique for binary recognition tasks, determining the probability of class membership.

2) Lexicon based model

a) Valence Aware Dictionary and sEntiment Reasoner (VADER)

VADER, or Valence Aware Dictionary and Sentiment Reasoner, is a sentiment analysis tool that makes use of dictionaries and rules. We utilized VADER to parse the text and applied a pre-built lexicon related to emotional words. In the dictionary, each word is assigned an emotional score and adjusted according to the rules. The VADER model will ultimately aggregate these scores to determine the overall sentiment of the text. This model can handle common informal and diverse languages in social media content, providing high-precision real-time emotional analysis. Therefore, it suits social media text sentiment evaluation. It gets by without the help of training data, and it is solely relying on dictionaries and sentiment-setting information. It has quick calculation speed and can be useful in real-time processing of data. VADER is mainly used for sentiment analysis, especially for NLP projects, and it is often used for short text sentiment analysis, which can quickly detect the emotions of the text [12].

b) TextBlob

TextBlob is an open-source text processing library in Python, and widely used in very core natural language processing. This API makes text processing very simple to perform in an operation. Some of the main functionalities that TextBlob deals with are word tokenization, tagging, named entity recognition, spelling correction, translation, n-gram generation, and sentiment analysis. The sentiment analysis module will try to estimate the polarity and subjectivity of the text. Polarity ranges between –1 (negative) and 1 (positive), while subjectivity ranges between 0 (objective) and 1 (subjective). TextBlob is an excellent library for basic sentiment analysis tasks and quick prototyping [13]. We used these scores to classify the sentiment of each post. Its ease of use and versatility made it suitable for quick prototyping and basic sentiment analysis.

c) SentiWordNet

SentiWordNet is a lexicon-based sentiment analysis tool that associates sentiment scores with WordNet synsets. Each synset has a positive, negative, and objective

score, which provides fine-grained sentiment analysis. SentiWordNet can be used with a Part-of-Speech (PoS) tagger in studies on complex tasks. SentiWordNet relies heavily on WordNet's systems. Hence, it may not prove effective if some new words or slang are used that are not part of WordNet. It is helpful for in-depth research and applications that call for in-depth sentiment analysis [14, 15]. SentiWordNet labels each word in a post with a corresponding sentiment score. After calculating the aggregated score, we can analyze the overall sentiment of social media post.

3) Machine learning and lexicon based model

In evaluating the performance of our machine learning, deep learning and lexicon based models on the dataset, we used three key metrics: accuracy, F1-Score, and Recall. The selection of these metrics is to comprehensively evaluate the effectiveness of the model in classifying content related to endangered species.

D. Evaluation Metrics

1) Accuracy

Accuracy is essential metric which determine the ratio of Correctly classified instances from total image. It gives a general perception of how good the model works in all of the points included.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. In our evaluation, we used an accuracy rate to comprehend the overall correctness of the models' predictions.

2) F1-Score

The F1-Score measures the balance between precision and recall, making it a useful and robust metric. It is particularly accurate for imbalanced datasets and accounts for false positives and false negatives. The F1-Score is calculated as follows:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Precision is the ratio of correctly predicted positive observations to the total predicted positives. It indicates the accuracy of the positive predictions.
- Recall (or Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. It indicates the ability to capture all relevant instances.

A high F1-Score ensures that our models not only capture relevant content (precision) but also capture all relevant instances (recall), confirming the input quality.

3) Recall

Recall, also known as sensitivity, ensures that the model locates all priority instances in the dataset. It is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall becomes one of the most imperative because it determines the ability of the model to detect endangered species content by the model. Low recall ensures that there is an important level of miss that could be mistaken to be irrelevant. This is critical for comprehensive analysis

IV. Result and Discussion

A. Machine Learning

Comparing the three models, Logistic Regression stands out as the best performer in terms of accuracy, F1-Score (0.8665), and recall (0.9049) (see Table I). The high recall and F1-Score of Logistic Regression suggest that it is highly effective at both detecting positive instances and minimizing false positives [9]. Support Vector Machine (SVM), while slightly less accurate than Logistic Regression, still offers a good balance and performs better than Naive Bayes in all metrics [16]. Multinomial Naive Bayes, though the least accurate, still provides a reasonable baseline performance [18].

TABLE I. Results of Machine Learning

Models	Performance		
	Accuracy	F1-Score	Recall
Multinomial Naïve Bayes	0.7628	0.7303	0.7628
Support Vector Machine (SVM)	0.7934	0.7733	0.7934
Logistic Regression (LR)	0.8017	0.8665	0.9049

B. Lexicon-Based Models

VADER did best in this test run because the lexicon was optimized on text from social media; it conducted the analysis very quickly and is particularly suited for processing informal text and real-time data streams. TextBlob performs generally well with sentiment analysis because of its simplicity and versatility, but it is somewhat less suitable for processing complex expressions. In the meantime, SentiWordNet gives the overall fine-grained sentiment analysis; performance is relatively poor since it is complex and very much dictionary coverage dependent. According to the comprehensive test results and the features of each one, it is recommended that VADER should be used first in processing text on social media, and TextBlob should be used for general sentiment analysis. If one requires complex sentiment analysis, wherein the complexity is under control, one can use SentiWordNet.

C. Comparison of All Models

The research presented the capacity that machine learning models and lexicon-based models can hold for the exploration of social media data for conservation purposes. The Logistic Regression model emerged as the nimblest, while the VADER model turned out to have the best performance of all lexicon-based models. These tools can generate precise public opinions on social movements, forestalling the failure of the implementation of conservation projects. The goal of the ongoing project should be to develop the preprocessing techniques of the data and investigate the cooperation between the machine learning process and the lexicon-based approach to get the best of both worlds (see Table II).

TABLE II. ALL THE RESULTS OF THE MODELS

Model Type	Model Name	Performance		
		Accuracy	F1-Score	Recall
Machine Learning	Multinomial Naïve Bayes	76.28%	73.03%	76.28%
	Support Vector Machine (SVM)	79.34%	77.33%	79.34%
	Logistic Regression (LR)	80.17%	86.65%	90.49%
Lexicon-Based Model	TextBlob	64.5%	64.56%	59.65%
	SentiWordNet	59.99%	60.92%	57.53%
	VADER	77.23%	84.49%	90.14%

D. Discussion

These findings align with other studies in the field. Zhou *et al.* [7] reported similar results, highlighting the effectiveness of machine learning models like SVM and Logistic Regression in environmental conservation contexts. Kietzmann *et al*. [6] emphasized the challenges of social media analytics, which this study addressed through robust preprocessing and the use of VADER, aligning with their emphasis on sophisticated tools. Additionally, a study by Chmiel *et al.* [19] found that negative emotions boost user activity, indicating the importance of accurately capturing sentiment for effective conservation messaging. Another survey of sentiment analysis methods by Giachanou and Crestani [20] discussed the utility of Twitter sentiment analysis in various contexts, including conservation efforts.

E. Limitation

Social media data is inherently more difficult to classify due to its unstructured nature, use of slang, emojis, and varying tones [21]. Ensuring the relevance and accuracy of the data collected is challenging, as social media platforms often contain noise, irrelevant information, and duplicates. Despite robust data cleaning and preprocessing techniques, some degree of irrelevant or misleading data may still influence the analysis.

V. CONCLUSION

In conclusion, the study uses social media data to help endangered species conservation efforts. We collected data from Twitter, YouTube, TikTok, and Instagram to obtain a comprehensive and all-encompassing dataset on public sentiment and interactions.

Logistic regression and VADER are the best performing models for sentiment analysis. The dataset was rigorously preprocessed and validated.

This study shows that social media can play a key role in the process of environmental protection and provides answers that can be used to plan and implement strategies. Further research should process more data and develop hybrid methods for better analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Ma Zhen was responsible for the implementation of machine learning models and conducted the overall data analysis. Zhang Chi focused on the development and application of the lexicon-based model. Ye Nan handled the data processing tasks. Abdur Rashid Sangi and Baha Ihnaini provided the initial concept for the proposed work, mentored the team members throughout the research, regularly reviewed the progress, supervised all activities, and arranged the funding for the project. All authors have reviewed and approved the final version of the manuscript.

REFERENCES

[1] E. van Huis. (May 2024). Biodiversity loss poses direct threat to economy. BNP Paribas. [Online]. Available: https://www.bnpparibas.nl/en/biodiversity-loss-poses-direct-threat-to-economy/#:~:text=A%20strong%20economy%20cannot%20exist,we%20often%20take%20for%20granted

[2] World Bank. (April 2024). Biodiversity, World Bank. [Online]. Available: https://www.worldbank.org/en/topic/biodiversity

[3] F. Isbell, "Causes and consequences of biodiversity declines," *Nature Education Knowledge*, vol. 3, no. 10, 2010.

[4] E. Goldsmith, "Social media's role in environmental conservation: a lens into public awareness and engagement," *Environmental Communication Journal*, vol. 12, no. 2, pp. 175–193, 2018.

[5] C. Folke, T. Hahn, P. Olsson, and J. Norberg, "Social media data for conservation science: A methodological overview," *Ecology and Society*, vol. 21, no. 3, 2016.

[6] J. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," *Business Horizons*, vol. 61, no. 3, pp. 259–265, 2018.

[7] X. Zhou, S. Chen, and X. Liu, "Harnessing social media for environmental conservation: A study on the effectiveness of multi-model engagement," *Journal of Environmental Management*, vol. 157, pp. 161–170, 2015.

[8] Data Preprocessing: Definition, Key Steps and Concepts. Data Management. [Online]. Available: https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

[9] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, pp. 23–34, Feb. 2012. doi: 10.20982/tqmp.08.1.p023

[10] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," Technical Report: UTEP-CS-18-09, University of Texas at El Paso, 2018.

[11] S. Wang, L. Jiang, and C. Li, "Adapting naive bayes tree for text classification," *Knowledge and Information Systems*, vol. 44, no. 1, pp. 1–17, July 2015. doi: 10.1007/s10115-014-0746-y

[12] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. the International AAAI Conference on Web and Social Media*, May 2014. doi: 10.1609/icwsm.v8i1.14550

[13] J. P. Gujjar and H. R. P. Kumar, "Opinion mining for the customer Feedback using TextBlob," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, no. 6, pp. 72–76, 2020. doi: 10.32628/cseit206418

[14] M. Shelke, D. D. Sawant, C. B. Kadam, K. Ambhure, and S. Deshmukh, "Marathi SentiWordNet: A lexical resource for sentiment analysis of Marathi," *Concurrency and Computation: Practice and Experience*, vol. 35, November 2022. doi: 10.1002/cpe.7497

[15] M. Fikri and R. Sarno, "A comparative study of sentiment analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 3, pp. 902–909, 2019. doi: 10.11591/IJEECS.V13.I3.PP902-909

[16] S. M. Walker II. F-Score: What are Accuracy, Precision, Recall, and F1 Score? KLU AI. [Online]. Available: https://klu.ai/glossary/accuracy-precision-recall-f1

[17] B. K. Shrivash, D. Verma, and P. Pandey, "Performance evaluation of machine learning and deep learning approaches for sentiment

analysis on COVID-19 sentiments," *Tuijin Jishu/Journal of Propulsion Technology*, vol. 44, no. 5, pp. 4295–4310, 2023.
[18] T. Islam, M. A. Sheakh, M. R. Sadik, M. S. Tahosin, M. M. R. Foysal, J. Ferdush, and M. Begum, "Lexicon and deep learning-based approaches in sentiment analysis on short texts," *Journal of Computer and Communications*, vol. 12, no. 1, pp. 11–34, January 2024. doi: 10.4236/jcc.2024.121002
[19] A. Chmiel, P. Sobkowicz, J. Sienkiewicz, G. Paltoglou, K. Buckley, M. Thelwall, and J. A. Hołyst, "Negative emotions boost user activity at BBC forum," *Physica A: Statistical Mechanics and Its Applications*, vol. 390, no. 16, pp. 2936–2944, 2011. https://doi.org/10.1016/j.physa.2011.03.040
[20] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–41, 2016. https://doi.org/10.1145/2938640
[21] S. Batbaatar, Y. Li, and S. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.