12-2022

Opinion Mining of Bird Preference in Wildlife Parks

Isiwat Adenopo
East Tennessee State University

Opinion Mining of Bird Preference in Wildlife Parks

A thesis

presented to

the faculty of the Department of Computing

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Masters of Science in Computer Science, AI and Machine Learning

---

by

Isiwat A. Adenopo

December 2022

Dr. Ghaith Husari, Chair

Dr. Brian Bennett

Mr. Matthew Harrison

Keywords: Opinion Mining, Text Analytics, Polarity Score, Tweet Analysis, Wildlife Parks

ABSTRACT

Opinion Mining of Bird Preference in Wildlife Parks

by

Isiwat A. Adenopo

Opinion Mining is becoming the fastest growing area to extract useful and insightful information to support decision making. In the age of social media, user's opinions and discussions have become a highly valuable source to look for users preferences, likes, and dislikes.

The industry of wildlife parks (or zoos) is a competitive domain that requires careful analysis of visitor's opinions to understand and cater for their preferences when it comes to wildlife. In this thesis, an opinion mining approach was proposed and applied on textual posts on the social media platform, Twitter, to extract the popularity, polarity (sentiment), and emotions toward birds and bird types such as owls, sparrows, etc. Then, the thesis provides recommendations based on popularity of birds and bird types and a ranked list of the most desired birds based on consumer emotions toward them. The findings of this thesis can help wildlife parks in the decision-making process on the types of birds to acquire.

DEDICATION

This thesis is dedicated to God, my family, and friends.

ACKNOWLEDGEMENTS

Thank you to my committee chair, Dr. Husari for the support and guide provided through my thesis process. I couldn't have done this without your help. I am grateful to my committee members, Dr. Bennett, and Mr. Harrison, and my graduate coordinator, Dr. Roach for supporting me and guiding me through my thesis process. I am also grateful to the department of Computing for the support I received through my degree.

I am extremely grateful to my family for their support through my degree. I couldn't have done it without them. In addition, I would like to express my sincere gratitude to my family and friends for being my personal alarm clock despite being thousands of miles away from them.

I want to specially acknowledge my amazing friends, both home and abroad for their love and support through my degree and thesis process, I am immensely grateful for everything.

I would also like to specially recognize the immense love and support I received from my Nigerian WhatsApp family throughout my degree. I will always treasure and cherish the beautiful and healthy relationship I have with them. I will never trade them for anything, and they will always have a special place in my heart.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction

1.1. Motivation

A growing amount of data is being captured and maintained by organizations across various sectors, changing how they conduct business. Several businesses today utilize opinion mining to understand how their customers feel about their products, enabling them to understand their level of satisfaction with their products better. Opinion mining is a more practical approach when conducting market research that allows for improved strategy and planning. Today, there are a variety of microblogging websites like Twitter that provide access to a wide range of information.

Microblogging typically involves posting real-time messages about different topics, discussing current events, complaining, and expressing positive sentiments about products one uses regularly. As a result, the product manufacturers began polling these microblogging websites to determine people's opinions regarding these products. Companies frequently use microblogs to evaluate their users' reactions and interact with them. This paper uses Twitter data on different birds to mine users' opinions on these birds and focuses on using machine learning techniques to analyze the data.

1.2 Problem Statement

Analyzing human opinions and feedback is crucial in identifying and catering to consumer needs. With the growth of social media platforms such as Facebook, Twitter, and other social networking sites, user-generated content to express opinions about products and services has become the primary information source for businesses to learn about how customers feel about services, their strengths, and weaknesses. Thus, automated solutions that can extract

business and marketing insights would enhance the decision-making process based on consumer opinions.

Accordingly, this thesis proposes a solution that aims to tackle three main challenges:

1. The fast-growing consumer-generated feedback about the desired types of birds has become a time-consuming and infeasible process for non-automated human analysis.
2. Textual posts may contain bird hashtags for irrelevant posts containing hyperlinks or extremely short tweets (e.g., a hashtag and an emoji). This type of post requires cleaning or preprocessing to obtain meaningful and informative data.
3. The lack of a star-based ranking system for each bird makes it hard (if altogether feasible) to aggregate thousands of posts in a meaningful manner to provide a ranked list of the most desirable bird and bird types.

1.3. Objectives

To tackle the challenges in this domain, this thesis proposes a solution with three main objectives:

1. Mine consumer-based insights and information from social media posts about the kinds of birds consumers desire to see.
2. Conduct sentiment and polarity analysis of consumers' textual posts about the desired types of birds to classify and aggregate them into positive and negative.
3. Develop a system to support wildlife parks by recommending:
    a. Birds and types of birds mentioned most frequently in social media, and
    b. The most desired birds and types of birds based on a sentiment (polarity) score extracted and aggregated for each bird and bird type.

4. Provide wildlife park stakeholders with insightful data and opinion visualizations of the most desired (ranked) birds and types of birds in wildlife parks.

1.4 Background

1.4.1 Machine Learning (ML)

Artificial intelligence (AI) solutions perform complex tasks. The concept of machine learning (ML) is broadly defined as the capability of a machine to imitate intelligent human behaviors [1]. Like natural language processing (NLP), ML is a type of AI. NLP and machine learning solutions can work together to make even more complex decisions. The machine learning model approach is classified into supervised and unsupervised learning.

Supervised Machine Learning

Data classification and prediction are usually accomplished using supervised machine learning. It learns the relationship between inputs and outputs to predict outcomes. Supervised learning requires human interaction to label the data accurately or automatic labeling after collecting historical data to develop a model.

Unsupervised Machine Learning

In machine learning, unsupervised methods are commonly used to identify relationships within datasets. In this method, models are trained on raw, unlabeled training data. Unsupervised learning is used when similar data needs to be clustered into a specific number of groups. Unlike supervised learning, unsupervised learning does not require pre-determined labels. Instead, it requires setting hyperparameters like the number of cluster points.

1.4.2 Natural Language Processing (NLP)

Natural Language Processing, called NLP, translates text into a character-based language in real-time or non-real-time. It could be audio, video, or text represented here by the information [2]. Converting language to speech is the primary function of this process. NLP primarily aims to reveal hidden information over unstructured/uncertain data and make it machine intelligible.

1.4.3 Sentiment Analysis

One frequently used text classification tool is sentiment analysis, which analyzes incoming messages and determines whether the sentiment is positive, negative, or neutral [3]. Sentiment analysis, also known as opinion mining [4] , uses NLP and text analysis to identify emotions by extracting subjective information from source texts using NLP and text analysis. Opinion mining aims to determine whether customers are satisfied with a product based on their surveys and reviews. The process involves mining text for sentiment and personal information using data mining, ML, and AI. Organizations use sentiment analysis to analyze unstructured, unorganized text on social media platforms, web chats, emails, blog posts, and comments. The analysis then determines and categorizes opinions about their products, services, and ideas.

1.4.4 Text Analytics

Text analytics is extracting meaning from text [5]. Most of the text on the web is unstructured and scattered. Text analytics aims to identify patterns or trends in unstructured data. This data can be collected and analyzed after cleaning to provide valuable business information that could be used for business decisions. As part of this project, text analytics was used to identify and understand the popularity of different bird types. Data visualization techniques can facilitate understanding of text analytics results and prompt decision-making.

1.4.5 Twitter Application Programming Interface (API)

Finding data to analyze is the first step in analyzing data. Among the best places to find data is Twitter. Twitter's APIs provide developers with access to most of its functionality including information about tweets, users, retweets, media, likes, direct messages, favorites, and trends on Twitter [6]. Twitter API requests are authenticated using OAuth, a widely used open authorization protocol [6].

1.4.6 Tweepy

Tweepy is a Python package that allows access to Twitter's API using Python. Several of Tweepy's classes and methods represent Twitter's models and API endpoints [6]. Tweepy has a method that returns the most recent statuses posted by the specified user. Access is provided to all RESTful API methods in the Twitter API class. The methods handle various implementation details transparently, including encoding, decoding, HTTP requests, pagination of results, and OAuth authentication [6]. Invoking an API method returns a Tweepy model class instance, and this contains the data returned from Twitter which can then be used within our application [7].

1.5 Challenges of Sentiment Analysis

1. The length of the text. Tweets are characterized by their short length, up to 140 characters, in contrast to other social media platforms. These texts are shorter than other social media platforms or movie reviews and might affect the performance of sentiment analysis. The length of the text of tweets can be a limitation when performing sentiment analysis because a user mostly tries to fit in the content of a tweet in a short string.
2. Stop words. Stop words are commonly used words found in any language, not just English. Examples of English stop words include "a", "an", "the" and words like this are

not suitable for the performance of sentiment analysis, and they are often filtered out during preprocessing steps.

3. Multilingual content. Some tweets are not written in English. One challenges to detecting languages when applying sentiment analysis to texts with multiple languages.
4. Incorrect use of English. Sometimes, people write tweets that do not follow the proper English grammar, or they might use slang. For example, “me likey,” which means “I like it” or “I like this.” The analysis may not understand if it really means “I like this”. This is a social media way of texting, which becomes a challenge when encountered in a dataset.
5. Relevance of the topic. The topic of any tweet greatly influences sentiment analysis because its sentiment orientation is based on the types of topics discussed. An example of such tweet is “It is a commonly advertised monument #BirdWatching”. There is no specific bird reference in this tweet, but it does contain a hashtag related to birds.

1.6 Remaining Organization of Thesis

The remainder of this thesis is organized as follows. Chapter 2 includes a discussion of the literature that inspired this thesis. Chapter 3 presents an overview of the design and approach of this architecture. Chapter 4 discusses and evaluates the results of the polarity scores. “Chapter 5 presents the drawn conclusion and avenues for potential future work.

Chapter 2. Literature Review

2.1 Sentiment Analysis on News Data

The work in [8] conducted a sentiment analysis using live news data scraped from the web and machine learning algorithms to analyze the sentiment. Through web scraping, large amounts of information can be extracted from websites more quickly and efficiently than manually.

This paper applies sentiment analysis to live news data from the Redditt political news of India. Using the Multinomial Naive Bayes and Logistic Regression supervised classification learning algorithms, [8] evaluated, compared, and analyzed sentiment analysis on live news data. Next, this paper used Sklearn libraries to identify the negative language in news articles to determine the sentiment of the news archives. The authors used a lexicon-based approach with a supervised machine-learning algorithm to perform sentiment analysis on news headings.

The work in [8] built a sentiment classifier that can detect sentiment expressed through the news on a three-class scale. These classifiers are trained and tested on logistic regression and naïve bayes) classification algorithms. The authors in [8] reported that logistic regression algorithms outperform naïve bayes algorithms.

2.2 Sentiment Analysis on Newspapers

The approach proposed in [9] conducted a study that focused on identifying a target, collecting quotations (reported speech) about the target from three different newspapers, separating positive and negative news from positive and negative emotions expressed by the target, and comparing the results to manual data collection. Previous opinion mining studies have typically focused on highly subjective text types, such as reviews of movies or products, where

targets are clearly defined. Thus, applying Opinion Mining to the news domain requires a more precise definition of the essence of the problem than previous studies.

This study examines newspaper headlines to compare the results of the different studies on a single target because news usually deals with major national or international events filled with many emotions. Headlines are ideal for analyzing emotions at a sentence level because news articles are intended to attract the reader's attention.

Data extracted from different sources can be used in different contexts, depending on where they come from and the intended users. An example would be a public relations company monitoring public figures' images, analyzing social media for information about potentially dangerous situations, and tracking consumer opinions over time to conduct market and economic research. However, a reliable source of information is essential for acquiring relevant information.

According to the study, opinion mining/sentiment analysis is extracting an opinion regarding a given topic from a set of documents. New types of text on the social web, such as e-commerce reviews provide a snapshot of public opinions on various subjects, including economics, politics, and the environment. This study utilized a large amount of subjective data to compare consumer opinions about different products, and to detect the general mood of the people.

Furthermore, [9] used the SentiWordNet system to identify subjective, objective, and neutral news titles from three electronic newspapers: The Hindu, Times of India, and Deccan Chronicle, and to classify the identified subjective titles according to their emotional state (positive or negative). The authors conclude that topic-specific sentiment classification and other

subassemblies of opinion mining are less challenging to manage than open-domain opinion analysis.

The work in [9] concluded that annotations concerning a specific topic were not helpful for general opinion mining on the news and recommended additional research. These include analyzing sarcasm and negations in the text and exploiting a bag of words to enhance performance, employing methods for identifying targets and topics, and considering a topic's context when categorizing opinions.

2.3 Sentiment Analysis on Twitter Data

The work in [10] proposed a method for automatically detecting sentiments in Twitter messages (tweets) that examines the features of tweets and the semantic information contained within them. The researchers used noisy label sources from a few sentiment detection websites as training data.

In various applications, sentiment analysis of Twitter messages is essential to the analysis process. Several systems have attempted to automatically detect sentiment in text, news articles, blog posts, and web reviews. While these systems can detect sentiment from the raw word representation, these systems do not consider short Twitter messages of 140 characters.

This study developed a more effective solution by using a more abstract representation of tweets than raw tweet representations commonly used in previous studies. The study revealed that the proposed approach is more effective and robust than previous ones since the features developed can capture abstract representations of tweets.

Twitter sentiment analysis uses a two-step classification method to classify messages as subjective and objective and to distinguish between positive and negative tweets. As opposed to using manually annotated data to compose the training data as is done with supervised learning

approaches, sources of noisy labels were used as training data to reduce the labeling effort in creating these classifiers. In this study, the classification algorithms could generalize better when there was a lack of information on tweets. To achieve the best outcome, [10] analyzed the noisy and biased labels provided by the sources and explored different strategies for combining them.

As a result of the proposed approach, there are cases where sentences contain antagonistic sentiments, which pose the major limitation of this study. Therefore, the study emphasized that a more advanced sentence analysis should be employed to identify the primary focus of sentiments for proper classification.

2.4 Sentiment Analysis to Identification Student's Behavior in Higher Education

A study by [11] noted the importance of popular sites such as Twitter, Facebook, and YouTube to higher education students in fields such as engineering, pharmacy, and medicine, etc. These sites contain data that can be used to identify the behaviors and opinions of students. The websites provide students with a platform to share their emotions, views, stress, and feelings related to learning [11].

An analysis of uncontrolled social web spaces was conducted in this study to understand students' learning behavior. The study focuses on higher education students' tweets about problems in their educational behavior, using both large-scale data mining and qualitative analysis.

The authors evaluated and discussed classification strategies for analyzing each student's behavior, such as the naïve bayes multi-label classifier and CRISP-DM process. A new opinion-based memetic classification algorithm proposed by the authors enabled a new system to be developed that overcomes the shortcomings of existing problems and systems. The memetic algorithms proposed by the authors are derived from genetic algorithms exploiting local search

or population-based methods. Compared to genetic algorithms, this classifier is primarily used for optimization.

An opinion-based memetic classifier was developed after a memetic algorithm was optimized with a classification technique, resulting in a fully optimized classification result. The authors propose that future studies will use videos, images, notations, or smiles so that the opinion-based memetic classifier can produce more accurate results.

2.5 Opinion Mining on Correlation between Events and Sentiment

According to [12], an approach was proposed to mine Twitter users' opinions and the correlation between events and sentiment. This paper applies Bayesian Logistic Regression (BLR) classification, a machine-learning technique commonly used for text categorization. This study used BLR to determine the correlation between this sentiment and significant events during the FIFA World Cup 2014.

First, the authors used Twitter's streaming API to extract data using the official FIFA World cup hashtags for filtering and processing the tweets. Using this dataset, they aim to analyze public sentiment toward unexpected and future events.

Using manually labeled positive and negative tweets, the authors built a trained method to search for correlations between Twitter sentiment and previous events. They used external lexicons, Unigram features and Bigram features to classify tweets as subjective or objective. Then, the authors filtered them using the Term Frequency-inverse Document Frequency) measure [12].

In addition, [12] introduced average sentiment polarity measures for various entities and events to know how people react and discuss them positively and negatively. Tweet sentiment analysis revealed interesting insights into users' opinions. The study displayed people's positive

and negative reactions to such events and how they can change based on incidents during the event.

2.6 Polarity Classification of Twitter Data using Sentiment Analysis.

The work in [13] conducted a sentiment analysis on Twitter data to determine the polarity of tweets. According to [13], sentiment analysis involves identifying and categorizing opinions expressed in texts to determine the users' positive, negative, or neutral attitudes toward a subject. By evaluating the success of an ad campaign in a new product launch, marketers can use this analysis to determine which version of a product or service is the most popular. In addition, this study also identified which demographics prefer or dislike certain product features.

For sentiment polarity categorization, this study evaluated two classifiers: linear and probabilistic classifiers. The authors used a Python machine learning software package called Scikit-learn to analyze tweets. They used three different algorithms to compare their results: support vector machine (SVM), naïve bayes, and logistic regression [13]. This study used SVM to represent a linear model, while the probabilistic models used a naïve Bayesian classifier and logistic regression. The linear and probabilistic approaches were compared at the end of this study to conclude.

According to the study, there are advantages and disadvantages to each of these three different algorithms. However, the SVM was the most effective because it could deal with linear and non-linear data. A significant finding in this study is that people increasingly rely on the information on the internet for everyday activities. Therefore, sentiment polarity categorization, a recurring problem of sentiment analysis, was carefully examined.

In computer science, sentiment analysis has become one of the fastest-growing research fields, making tracking all activities challenging. The sentiment level analysis aims to classify

every sentence according to its senti [14]ment. Sentiment analysis at the aspect or entity level aims to classify sentiments about specific aspects. Sentiment analysis, however, might be hindered by shortcomings in the online data available. Because people are free to post their content, it is impossible to validate or guarantee the quality of the opinions they provide.

In conclusion, the authors suggested that machine learning and deep learning algorithms can be used to improve a classifier and better mining techniques for natural language processing.

2.7 Sentiment Analysis on COVID-19-Related Social Distancing

An analysis of Twitter data on Covid-19 related social distancing in Canada [14] is presented in the article. This study aimed to analyze and understand public sentiments towards social distancing through Twitter textual data. In this study, sentiment polarity was extracted from tweets using a tool called SentiStrength. This study used the SVM algorithm to analyze sentiment. The Twitter data used for this study was from a free, open-source website provided by IEEE [14].

The dataset included global tweets filtered using COVID-19 keywords for one month. [14] divided the tweets into positive, negative, and neutral sentiments according to the polarity of their expressed sentiments. Tweets were manually explored to better understand public concerns related to the pandemic, thereby increasing awareness and dissemination of the COVID-19 resources available. It provided a deeper understanding of social distancing from a Canadian perspective. Using this analysis, the authors claim governments can make decisions to improve the health of populations.

The study found that 40% of Canadians have neutral attitudes toward social distancing, 35% have negative attitudes toward the concept, and only a quarter are positive about it. The

SVM algorithm performed at 87% accuracy in terms of performance evaluation. The authors suggested that the algorithm's performance can be improved by increasing the training data.

2.8 Polarity Classification of Twitter Messages using Audio Processing

The work in [15] analyzed the polarity classification of Twitter messages using audio processing. They proposed a novel method called Sound Cosine Similarity Matching. The method is used as a polarity classification for Twitter messages, incorporating features based on audio rather than text properties.

The study's objective was to provide a novel approach to correcting misspelled and shortened words. They correct such words by extracting phonetic characteristics using their audio representation, expected to be similar to the correct form of the word.

The research used a Twitter dataset and incorporated audio based on recent observations on how human brains process text. The authors evaluated the model in two ways by capturing the rate of misspelled and shortened words with a classification of the feature set.

The work in [15] concluded that the proposed feature set provided high-accuracy results compared to the classical Bag of words and Text Cosine Similarity Matching models, especially for extensive data. It also showed that the beginning of the feature set was identical. However, they became different when the misspelled or shortened words using Sound Cosine Similarity matching were corrected.

Chapter 3. The Proposed Approach

This chapter describes the system proposed in this thesis. Figure 1 below shows the framework for the proposed approach. The framework consists of seven modules that start with bird name collection and ends with providing aggregated user-based knowledge about the most desirable types of birds. These modules are explained in detail in the following subsections.

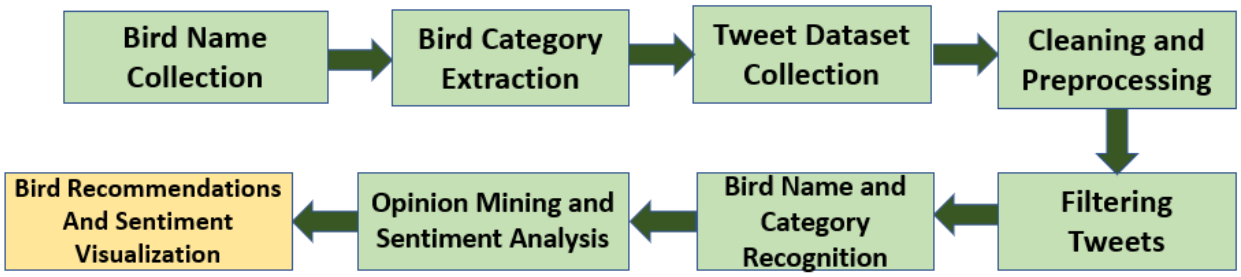

Figure 1 *The Framework of the Thesis Approach*

3.1. Data Collection

3.1.1. Collecting Bird Names

This thesis collected the bird names from Wikipedia [16]. This data included 11,001 initial bird names. However, we discovered that part of these were not bird names but contained only single characters from A-Z. To this end, we cleaned all single-character bird names from the list. The final list of 11,001 bird names was used as a search query with Tweepy API to collect tweets containing any of these bird names.

3.1.2 Collecting Bird Categories

Some tweets do not mention a specific bird name, e.g., “red-bellied woodpecker”. Those tweets only mention bird categories in general, such as “I love the kingfisher!”. To solve this problem, the bird categories (or classifications) from the bird names were collected by extracting the last word of the bird name. One example of a bird name is “Great Horned Owl” and by

extracting the bird name's last word, the bird's category will become “Owl”. Category extraction allows our approach to identify tweets that generally mention a bird type or category without specifying a particular bird name. By the end of this process, this thesis collected 903 bird categories.

3.1.3. Tweet Dataset Collection

For this step, Tweepy was used to collect live tweets from Twitter with the search query containing the list of birds from Wikipedia. We collected 5000 tweets between January 1, 2018, and August 22, 2022. Data in the dataset includes the tweet owner’s name, his or her location, a short description of the tweet’s author, if the user is verified, when the tweet was created, the text of the tweet, hashtags related to the tweet, and the source of the tweet: whether it came from an Android device, an iOS device, or the Twitter web application.

3.2. Data Cleaning, Sanitization, and Preprocessing

When using supervised machine learning (ML) algorithms, data preprocessing has a significant effect on generalization performance. This step involves the cleanssing of tweets to feed to the model. To accomplish cleansing, a sequence of preprocessing steps was taken. The first step was to inspect the columns in the dataset to understand the data.

3.2.1 Data Cleaning

The process of data preprocessing in data mining and analysis involves taking raw data and converting it into a format understandable by computers and machine learning systems. The data-cleaning steps carried out in this study include the following:

1. Removing duplicate data
2. Converting all texts to lowercase
3. Removing URLs, symbols, emojis, etc., irrelevant to the analysis

4. Removing HTML tags
5. Removing boilerplate text from emails
6. Removing unnecessary blank text between words
7. Removing the hashtag symbol

3.2.2 Stop Words Removal

This step involves the removal of the most occurring common words in the tweet texts because they do not add value to the text used to develop an NLP model. Examples of stop words include “you”, “the”, “this”, “is”, “a”, “are” and “them”

3.2.2 Words Tokenization

This step involves splitting up the text columns into separate words called tokens. This was achieved using a word_tokenize function as part of a built-in python programming language package, Natural Language Toolkit (NLTK). Python's NLTK is used primarily for symbolic and natural language processing (NLP) of English written in Python. Words tokenization is an important step in data preprocessing as it helps understand the context of the text used to develop a model for NLP. This process helps interpret the meaning of a text by analyzing its sequence of words.

3.3 Filtering Tweets

The previous steps yielded 11,001 bird names and 903 categories. In this step, these lists were used to recognize the mentions of bird names or bird categories in tweets. Extracting names and categories supports this thesis’s aim of providing opinion-based recommendations of the most-liked birds and bird categories mentioned by social media users.

3.4. Sentiment Analysis

This step involves creating a polarity function using the Python package sentiment intensity analyzer to calculate the polarity scores for each tweet. The polarity scores include negative, neutral, positive, and compound sentiments. Positive sentiments indicate tweets that describe positive experiences about different birds or how the users of the tweets like or love a particular bird. The neutral sentiments refer to tweets that are neither good nor bad. Finally, negative sentiments refer tweets in which the users are dissatisfied or upset.

3.4.1. Sentiment Polarity Classification

A major subtask of natural language processing is sentiment analysis, and its primary goal is to reveal the overall opinion contained within a text. Sentiment analysis includes tasks such as subjectivity detection and polarity detection. Detecting subjectivity is removing ‘factual’ or ‘neutral’ content, i.e., non-opinionated, objective text, while detecting the polarity of an opinion is about distinguishing between 'positive' and 'negative' opinions. Sentiment intensity analyzer was used to extract the polarity of tweets. Figure 2 shows the distribution of the sentiment to positive and negative. It is observed that the tweets contained more positive sentiments than negative sentiments.

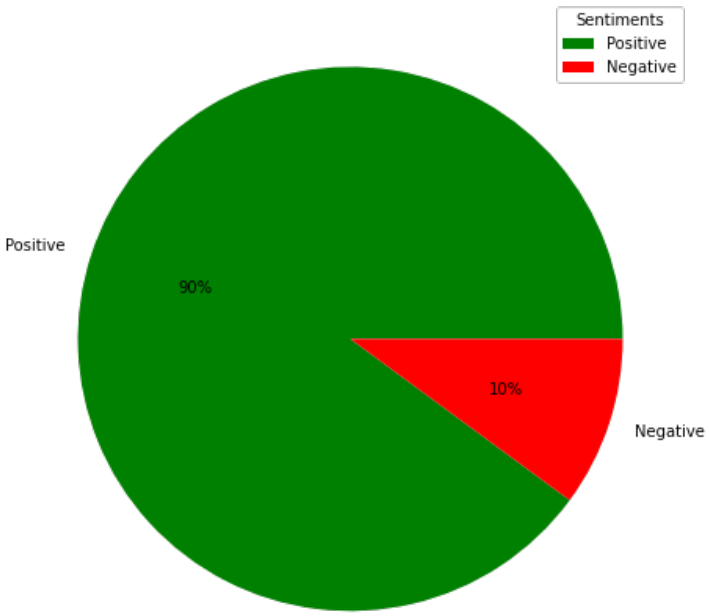

Figure 2 *Distribution of Sentiment*

3.5 Data Analysis

Exploratory data analysis includes data visualization, allowing translation of the information on the most talked about birds and categories into a visual context.

3.5.1 Word Cloud

Word clouds, also known as text clouds or tag clouds, are used to visualize unstructured text data and discover word patterns. The more often these words appear in a text or dataset, the larger they appear. This thesis has built a model to show the word cloud for the dataset used, which appears in Figure 3. A second word cloud was built to display the dataset's positive sentiments, which can be found in Figure 4. The figure also shows the larger texts that indicate how often they are mentioned in the dataset.

Figure 3 *Word Cloud for All Bird Names in the Dataset*

Figure 4 *Word Cloud of Bird Names Mentioned with Positive Sentiments*

3.5.2 Popularity of Birds

Table 1 shows individual birds' popularity based on the number of times people mentioned it or talked about it in their tweets.

Table 1 *Top 10 most frequently mentioned birds in tweets*

Bird Name	**Popularity Score**
Great Horned Owl	1484
House Sparrow	752
Belted Kingfisher	538
Red-Bellied Woodpecker	375
Mourning Dove	304
Northern Flicker	266
Downy Woodpecker	217
Pileated Woodpecker	172
Eastern Phoebe	133
Common Grackle	72

Performing data analysis on Table 1 above gives a visual representation on the top 10 most talked about birds in the dataset. Figure 5 shows the most frequently mentioned birds in the dataset, with the “great horned owl” being at the top of the list.

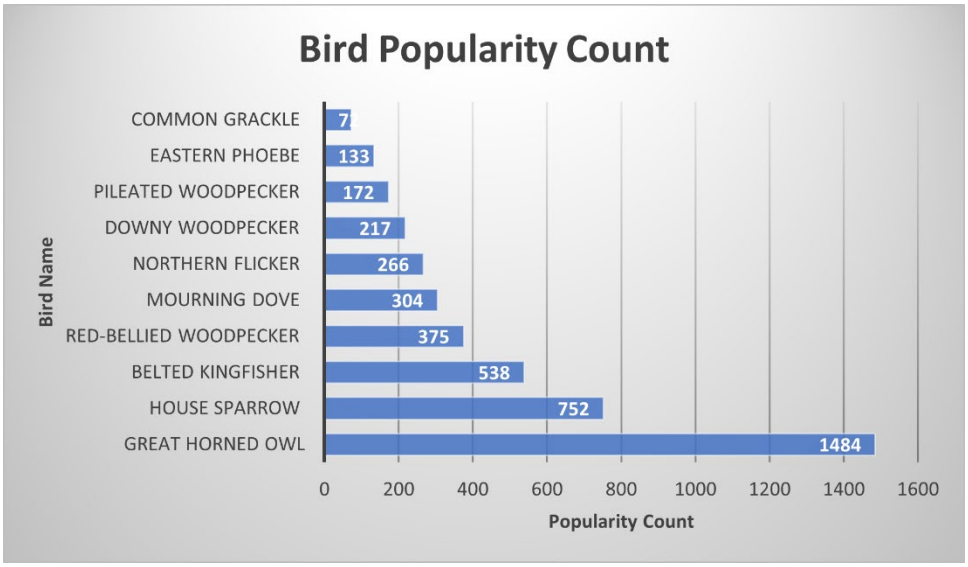

Figure 5 *Popularity Count of The Top 10 Birds*

3.5.3 Popularity of Bird Categories

Table 2 shows the bird category popularity based on the number of times it appeared in the tweets contained in the dataset. Figure 6 gives a visual representation of the dataset's top 10 most talked about bird categories.

Table 2 *Top 10 most frequently mentioned Bird Categories in tweets*

Bird Category	**Popularity Score**
Owl	1563
Sparrow	864
Woodpecker	794
Tern	754
Kingfisher	548
Dove	463
Flicker	274
Crow	155

Phoebe	137
Jay	126

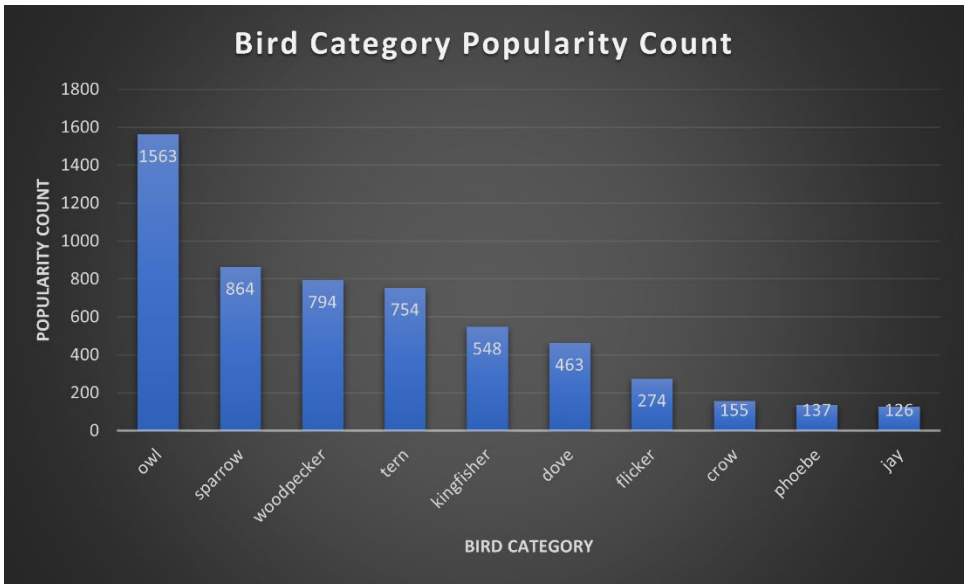

Figure 6 *Popularity Count of The Top 10 Bird Categories*

3.5.4 Average polarity score for the most frequently mentioned birds

Table 3 shows the average polarity score for each of the top 10 popular birds in the dataset. Performing data analysis on Table 3 gives a visual representation of the average polarity score of the dataset's top 10 most talked about birds, as shown in Figure 7.

Table 3 *Average Polarity Score for Top 10 Birds (1 means highly positive sentiments)*

Bird Name	**Polarity Score**
Great Horned Owl	0.96
House Sparrow	0.95
Belted Kingfisher	0.84
Red-Bellied Woodpecker	0.81
Mourning Dove	-0. 34

Northern Flicker	0.90
Downy Woodpecker	0.90
Pileated Woodpecker	0.42
Eastern Phoebe	0.86
Common Grackle	0.97

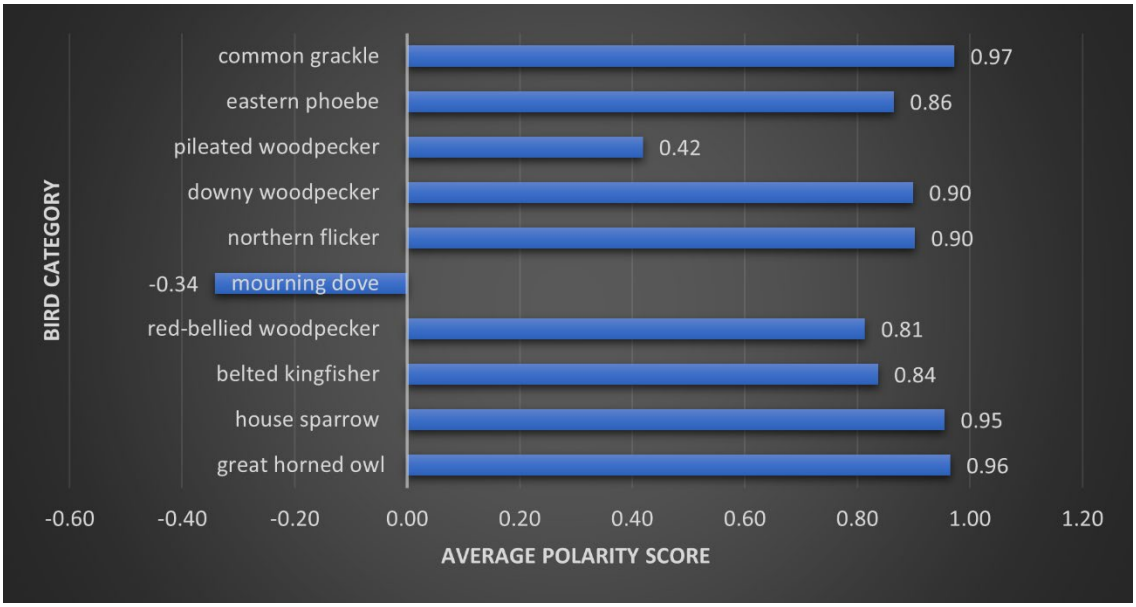

Figure 7 *Average Polarity Score for Top 10 Birds*

3.5.5 Average Polarity Score for Popular Bird Categories

Table 4 shows the average polarity score for each of the dataset's top 10 popular bird categories. Performing data analysis on Table 4 gives a visual representation of the average polarity score of the top 10 most talked about bird categories in the dataset, as shown in Figure 8.

Table 4 *Average Polarity Score for Top 10 Bird Categories*

Bird Category	Polarity Score
Owl	0.96
Sparrow	0.92
Woodpecker	0.76
Tern	0.94
Kingfisher	0.84
Dove	0.06
Flicker	0.91
Crow	0.97
Phoebe	0.87
Jay	0.32

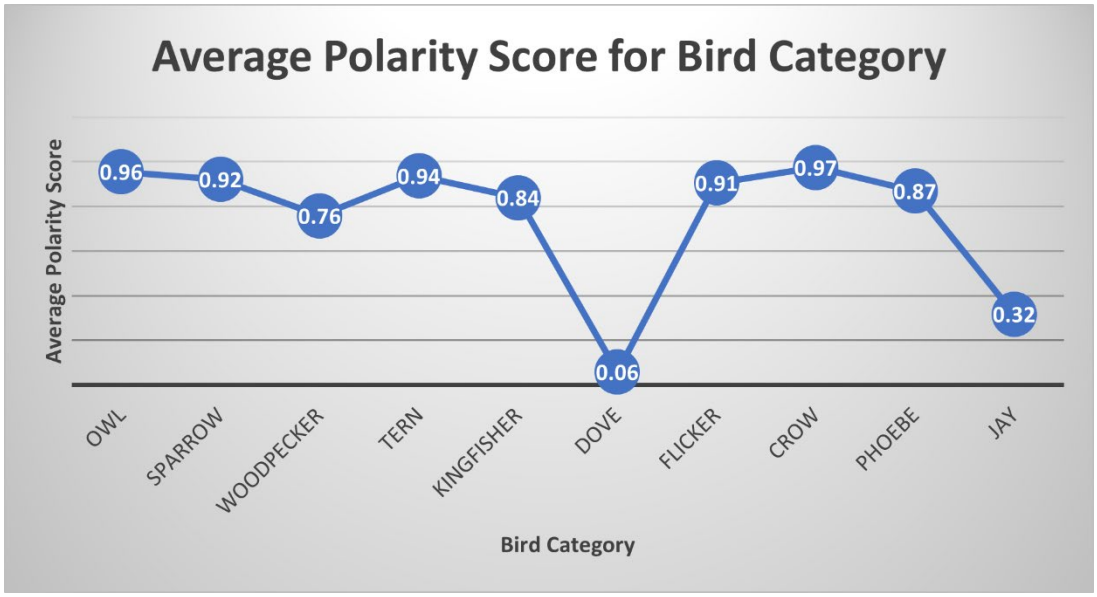

Figure 8 *Average Polarity Score For Top 10 Birds*

Chapter 4. Discussion and Limitations

Wildlife parks can use results from this thesis to acquire birds and categories most liked and frequently mentioned by people on social media. This study collected 5000 tweets mentioning different bird names in North America, and then based on that, we provide the top 10 most liked birds. This study can also be adjusted to a particular area or country. Bird names and categories might change if different bird names and tweets are collected for a different country. Areal limitations can also help inform people about the kind of birds others find desirable.

4.1 Casual vs. Interesting Bird Mentions

One important discussion is considering other interpretations of the most frequently mentioned birds. A different viewpoint to analyze and interpret frequently mentioned birds is that these birds are casually and frequently seen in nature. Therefore users may frequently create posts mentioning these birds. This could mean that acquiring these birds in wildlife parks might be more casual than interesting to consumers. In this case, it could be more beneficial to parks to acquire less frequently mentioned birds with highly positive sentiments to provide a more interesting bird collection that attracts more customers who seek to view interesting birds.

4.2 Limitations

In this thesis, we use the polarity for positive and negative sentiments, but one limitation is that the NLP sentiment analysis module sometimes misinterprets words. Negative sentiments could be extracted from bird names like “the mourning dove”. The word “mourning” is mistakingly interested by the sentiment analysis model as “sad”. This is evident in Figure 7, which shows “mourning dove” as the least desired bird. One possible solution to this limitation is to conduct a sentiment analysis on bird names and types to detect bird names that get falsely interpreted as “negative” sentiment. Then replace these names with a neutral name (e.g.,

mourning dove = LE_m dove). The original bird name can be inserted again after completing the sentiment analysis. Such a workaround can be implemented to avoid false sentiment interpretation due to similar bird names.

Another limitation in this domain is the bias in the data. The data used for this study only captures tweets in English. This means the data is mostly from English-speaking countries (e.g., the United States of America). Other locations can be considered for future work by including more languages to extract insights about birds and bird types in different countries.

In addition to the limitations mentioned above, emojis were not translated prior to the development of the approach. For instance, a crying emoji refers to sad emotion. Similarly, a smiling emoji translates to a happy emotion. As a result, some tweets may represent a different sentiment from the sentiment analysis output.

Chapter 5. Conclusion

This thesis proposed an end-to-end approach to collect tweets that mention bird names or types. Then, the approach performed sentiment analysis on these user-generated texts to extract how users feel about these birds. Finally, the approach provided aggregated scores and ranked lists of the most frequently mentioned bird and bird types and the most desired birds and their types based on the sentiment implied by the users' language.

5.1 Future Work

We believe that the bird popularity and sentiment analysis pipeline presented in this thesis is a starting point for additional research and analysis. In this part, we provide future directions that can improve the proposed approach to extract useful information from user-generated content more effectively.

One future direction is to use a similar solution and apply it to other wildlife species to discover how people feel about them. This could provide parks with more information about desired wildlife (other than birds), which can be helpful for the stakeholders to make more relevant decisions driven by consumer-based content.

Another possible direction is to gather user-generated content from other social media platforms and bird-watching platforms for this study. In addition, collecting posts in different languages and utilizing translation modules (e.g., Google translate) can enrich this work to include desired wildlife in different countries. Considering these factors allows us to better understand the user desirability of birds and wildlife.

Finally, this study could also collaborate with wildlife experts to determine ecological processes such as animal migration over time or hibernation patterns.

References

[1] S. Brown, "Machine learning explained," [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[2] "Natural Language Processing Thesis," [Online]. Available: https://www.phddirection.com/natural-language-processing-thesis/.

[3] S. Gupta, "Sentiment Analysis: Concept, Analysis and Applications," [Online]. Available: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

[4] T. Contributor, "Sentiment analysis (opinion mining)," [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining.

[5] "5 Text Analytics Approaches: A Comprehensive Review," [Online]. Available: https://getthematic.com/insights/5-text-analytics-approaches/.

[6] "How to Make a Twitter Bot in Python With Tweepy," [Online]. Available: https://realpython.com/twitter-bot-python-tweepy/#:~:text=Tweepy%20is%20an%20open%20source,Data%20encoding%20and%20decoding.

[7] "Tweepy Documentation," [Online]. Available: https://docs.tweepy.org/en/stable/getting_started.html.

[8] P. Kaur, "Sentiment analysis using web scraping for live news data with machine learning algorithms," *Materials Today: Proceedings,* pp. 3333-3341, 2022.

[9] S. Padjama, S. S. Fatimah and S. Bandu, "Analysis of sentiment on newspaper quotations: A preliminary experiment," *013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT),* pp. 1-5, 2013.

[10] B. Luciano and F. Junlan, "Robust sentiment detection on Twitter from biased and noisy data," *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10). Association for Computational Linguistics,,* pp. 36-44, 2010.

[11] R. T. Pooja and G. Prof Narendra, "Identification of Student's Behavior in Higher Education from Social Media by using Opinion based Memetic Classifier," *IJRITCC,* 2015.

[12] P. Barnaghi, P. Ghaffari and J. G. Breslin, "Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment," *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService),* pp. 52-57, 2016.

[13] A. S. Raghuwanshi and S. Pawar, "Polarity Classification of Twitter Data using Sentiment Analysis.," 2017.

[14] C. Shofiya and S. Abidi, "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data.," *Int J Environ Res Public Health,* 2021.

[15] D. Mihail and G. Dilek, "Polarity Classification of Twitter Messages using Audio Processing," *Information Processing & Management,* 2020.

[16] "List of birds by common name," [Online]. Available: https://en.wikipedia.org/wiki/List_of_birds_by_common_name.

[17] M. A. Alanezi and N. M. Hewahi, "Tweets Sentiment Analysis During COVID-19 Pandemic," *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI),* pp. 1-6, 2020.

VITA

ISIWAT ADENOPO

Education: M.S. in Computer Science, East Tennessee State University, Johnson City, Tennessee, 2022

B.S. in Engineering, University of Lagos, 2017

Professional Experience: Graduate Assistant, East Tennessee State University, Department of Computing, 2021 - 2022

Data Analyst, Technology Distributions Limited, 2019 - 2020