Independent Project

Report

Bee Species Prediction Using Feature Specifications in Text Format

---

Tejas Dubhir, 2018110

Shivangi Dhiman, 2018265

Advisor: Dr Swapna Purandare

7th May 2021

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Contents

1 Introduction

1.1 Problem Statement

To predict the species of a bee, given its description in text format, using Information Retrieval, Natural Language Processing and Machine Learning techniques.

1.2 Background

Bees are insects that play a significant role in the pollination of food crops and the production of honey and wax. Because of this, they are considered a crucial ecological and economic resource. There are over 20,000 known species of bees around the globe − for example, Apis mellifera (Common Honey Bee), Apis dorsata (Giant Honey bee), Bombus terrestris (Buff-tailed Bumblebee), etc. Each species has its own characteristic traits that are unique to it. One of the most common or recognizable features of bees are their striped furry bodies.

However, many species possess similar characteristics which can make the task of bee identification or bee taxonomy difficult. For example, almost all of the species of honey bees have varying dark and light striations. Bumblebees also have black and yellow stripes but are often larger than honey bees. Carpenter bees are also similar to Bumblebees but they usually have shorter hair and a black body.

Unfortunately, many species of bees are now enlisted as endangered. Many scientists believe this is due to various agricultural practices and the use of insecticides. Thus, in the past few years, it has become important to correctly classify bees as it would help in their proper surveying, studying and monitoring. But, due to such a large number of known species, it is infeasible to manually identify and classify all the bees, based on their physical traits.

To overcome this challenge, we have created an automated bee species identification system. We have employed Information Retrieval, Machine Learning and Natural Language Processing techniques to automate this process. Using a list of text-based features as queries/inputs, we aim to predict the species of the bee.

2 Literature Review

Our aim was to understand the implementation of various techniques that could help us in creating a model that would be able to predict the species of a bee using queries/keywords. For this, we read several research papers, some of which have been summarised below.

Automated Keyword Extraction Using Support Vector Machine from Arabic News Documents [3]

- Statistical feature extraction methods used:
  - Term Frequency - Inverse Document Frequency (TF-IDF)
  - First Occurrence
- Supervised learning model:
  - Support Vector Machine Classifier
- Preprocessing: the text was split into sentences, tokenized, normalised and the frequency of each word was compared to that in other documents.
- The dataset was imbalanced and thus the Downsampling Method was used to balance it.
- SVM Classifier with the Radial Basis Function (RBF) kernel was used, which gave a precision of 0.77, recall of 0.58 and an F1 score of 0.65.

Automatic Keyword Extraction from Documents Using Conditional Random Fields [38]

- A sequence labelling method – 'Conditional Random Fields' (CRF) was proposed to effectively extract keywords from 600 Chinese academic documents.
- SegTag was used to preprocess and compile the data.
- For their dataset, the most probable label sequence was determined by:

$$Y' = \arg\max P(X|Y) \qquad [Y' : \text{determined using the Viterbi Algorithm}]$$

- Because of its ability to relax the assumption of conditional independence of the observed data, CRF was able to avoid the label-bias problem.
- CRF (Precision: 0.66, Recall: 0.41, F1 score: 0.51) and SVM (Precision: 0.80, Recall: 0.33, F1 score: 0.46) significantly outperformed the rest.

Automatic Keyword Extraction from Individual Documents [29]

- Proposed a method called Rapid Automatic Keyword Extraction (RAKE): used a list of stopwords and phrase delimiters to detect the most relevant words or phrases in a text.
  1. The text was split into a list of words and the stopwords were removed to give a list of 'content words'.
  2. Created a matrix of words and stored the number of times they co-occur.
  3. The words were given a score (degree of the word in the matrix, the degree of the word divided by its frequency, or simply the word frequency).
  4. If two keywords or keyphrases appeared together in the same order more than twice, a new keyphrase was created regardless of how many stopwords the keyphrase contains in the original text.
  5. A keyword was chosen if its score belonged to the top T scores where by default, T is one-third of the content words in the document.

Identifying Offensive Posts and Targeted Offense from Twitter [39]

- The paper was divided into completing two subtasks:
  - Classifying/predicting the tweets if they are Offensive (OFF) or Not offensive (NOT)

  - Differentiating between Targeted Offense (TIN) and Untargeted Offence.
- Following techniques were applied:
  - Convolutional Neural network.
  - Bidirectional LSTM.
  - Bidirectional LSTM + GRU.
  - An ensemble of the above 3 architectures.
- The authors set up their own dataset by collecting the tweets through python libraries and manually finding the tone of the text body.
- For preprocessing they used NLTK and Keras. They completed preprocessing in the steps which were: Tokenization, cleaning and normalization.

Identifying Search Keywords for Finding Relevant Social Media Posts [35]

- The initial ranking, which is less accurate but very efficient, is to identify a large shortlist of likely keywords, or in other words, to remove those unlikely keywords. Re-ranking refines the ranking of the shortlisted words.
- Used the Double Ranking Approach.
  1. Provide keywords to be searched.
  2. Find all the keywords from the set from the dataset.
  3. Use the initial ranking algorithm to rank the found keywords and the unsuitable words should be removed from consideration.
  4. Use the Reranking Algorithm and rerank all the keywords in the obtained new set, from the dataset to get the more refined set of found words.
  5. From the new rank list, a subset is selected and checked if the dataset contains the keywords from the new set or not.
  6. If the results are satisfactory, end the process, otherwise repeat from Step 2.

Increasing the Accuracy of Discriminative of Multinomial Bayesian Classifier in Text Classification [23]

- The authors combined Discriminative Multinomial Bayesian Classifier with a feature selection technique that evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.
- Data preprocessing − texts were tokenized, stemmed and vectorized.
- Since SVMs produce good precision but poor recall, the authors tried to incorporate a Bayesian network classifier, with Frequency Estimate (FE) as the parameter.
- Using the Chi-squared method for feature ranking, it was found that an alpha of 0.01 produced the best results.

Snake Species Identification by Using Natural Language Processing [30]

- Used Weka for data extraction.
  - Data tokenization: using weka, separated words and formed vectors.
  - Stemming: bringing words to their natural form and removing -ing, -s, -es, etc.
  - Removing symbols, articles and punctuation.
- Feature extraction using TF-IDF:

$$\text{Normalized TF} = \frac{\text{Number of terms that occurred in the text}}{\text{Total number of words in the text}}$$

$$\text{IDF} = \log\left(\frac{\text{Total number of texts}}{\text{Number of texts in which selected term appeared}}\right)$$

$$\text{TF-IDF} = \text{Normalized TF} \times \text{IDF}$$

- Then the K-NN, Naive Bayes, SVM, and Decision Tree models were trained − Decision

Tree: 71.67%, SVM: 68.33%, Naive Bayes: 61.11%, K-NN: 55.56%.

Theme-weighted Ranking of Keywords from Text Documents Using Phrase Embeddings [19]

- Used word-embedding algorithms to extract keywords from the arxiv dataset.
- Preprocessing: stemming, tokenization, removal of stopwords.
- To train this preprocessed data, two toolkits − Word2Vec and FastText − were used.
  - The Word2Vec algorithms include skip-gram and CBOW (continuous bag of words) models, using either hierarchical softmax or negative sampling.
    - In the CBOW model, the distributed representations of surrounding words are combined to predict the word in the middle.
    - In the Skip-gram model, the distributed representation of the input word is used to predict the context.
  - FastText: used for efficient learning of word representations and sentence classification.
- It was found that the models trained using Fasttext performs better than the models trained using Word2Vec on all the three tasks.

Subsequently. we also read some review articles that summarised many research papers relevant to our project. We have summarised two of those review articles below.

Automatic Keyword Extraction for Text Summarization: A Survey [7]

- Text Extraction
  1. Simple Statistical Approach
     - Training set is not required, extraction is based on statistics derived from the frequency and the position of a word and a list of keywords is generated.
     - Techniques Used: TF [18], TF-IDF [28], Word occurrences [20], PAT-tree, n-gram statistical data [13].
     - After extraction, Aprioiri techniques − support and confidence − were used to infer the strength of the keywords.
  2. Linguistics Approach
     - Linguistic features such as lexical analysis [4] (applied using electronic dictionary, tree tagger, WordNet, n-grams, POS pattern), syntactic analysis [14] (noun phrase, chunks), discourse analysis [31] etc., were incorporated.
  3. Machine Learning Approach
     - Training Data is required.
     - Commonly used models: Hidden Markov models [40], Naive Bayes, Bagging [14].
     - Keyword Extraction Algorithm or KEA [36]: Document is converted into a graph and each word is treated as a node. Whenever two words appear in the same sentence, the nodes are connected with an edge for each time they appear together. Then the number of edges connecting the vertices are converted into scores and are clustered accordingly. The cluster heads are treated as keywords.
  4. Hybrid Approach: These approaches combine the Machine Learning and Statistical approaches to find the best features.
- Text Summarization
  1. Statistical Based Approach
     - Similar to the text extraction method, statistical-based approaches make use of TF, IDF, the position of the word, etc. [11, 25, 33]
  2. Machine Learning-Based Approach
     - Labelled data is used for training.
     - Techniques used: Hidden Markov models [33] and SVM classifiers, Naive Bayes,

Hidden Markov Models [27], Maximum Entropy, Neural Network [15], Support Vector Machine [22].

Keyword Extraction: A Review of Methods and Approaches [6]

- Keyword extraction methods
  - Simple Statistics Approaches: n-gram statistics, word frequency, TF-IDF [18, 28], word co-occurrences, PAT Tree, etc. These do not give preferable results in cases where the keywords' frequency is relatively lower.
  - Linguistics Approaches: Lexical, syntactic, semantic and discourse analysis.
  - Machine Learning Approaches: Mostly supervised techniques are preferred. The dataset needs to be manually annotated for training and thus makes the tasks difficult for the researchers.
  - Supervised learning: The task is to classify whether the given text contains the keywords or not. It is a binary classification problem.
    The following are the observations from various papers:
    - Noun phrases can be used instead of frequency and n-grams, and then POS tags can be added to them.
    - Statistical associations between keyphrases and enhancing the coherence of the extracted keywords can be done with the Naive Bayes Algorithm
- Vector Space Model: It is the method where the sentences are broken into vectors and stored as word frequency.
  - Disadvantages: the meaning of the sentence gets lost in the process, the words are independent of each other which means that "not", "bad" and "not bad" have entirely different meanings, if there are sentences that have different meanings but have the same words, they would be considered the same.
- Hence, instead of VSM, graph-based models are preferred.

We have compiled the results and evaluation metrics from the above-mentioned papers in Table 1.

Paper	Method	Accuracy	Precision	Recall	F1 Score
Rusli et.al, 2019	Decision Trees	0.7167	-	-	-
Rusli et.al, 2019	SVM	0.6833	-	-	-
Rusli et.al, 2019	Naive Bayes	0.6111	-	-	-
Rusli et.al, 2019	K-NN	0.5556	-	-	-
Zhang et.al, 2019	CNN (training data)	0.8387	-	-	0.8020*
Zhang et.al, 2019	Bidirectional LSTM with Attention (training data)	0.8246	-	-	0.7851*
Zhang et.al, 2019	Bidirectional LSTM + Bidirectional GRU (training data)	0.8301	-	-	0.7893*
Zhang et.al, 2019	CNN (test data)	0.8395	-	-	0.7964*
Zhang et.al, 2019	Ensemble of CNN, BLSTM with Attention, BLSTM + BGRU (test data)	0.8407	-	-	0.8066*
Armouty et.al, 2019	SVM	-	0.77	0.58	0.64
Armouty et.al, 2019	Naive Bayes	-	0.8	0.43	0.56
Armouty et.al, 2019	Random Forest	-	0.71	0.57	0.63
Mahata et.al, 2018	Word2Vec Skip-gram	0.7658	-	-	-
Mahata et.al, 2018	Word2Vec CBOW	0.6911	-	-	-
Mahata et.al, 2018	Fasttext Skipgram	0.9627	-	-	-
Mahata et.al, 2018	Fasttext CBOW	0.9392	-	-	-
Wang et.al, 2016	Double Ranking Algorithm	0.813	-	-	-
Mouratis et.al, 2009	Multinomial Naive Bayes	0.8903	-	-	-
Mouratis et.al, 2009	Sequential Minimal Optimization	0.8196	-	-	-
Zhang et.al, 2008	Conditional Random Fields	-	0.6637	0.4196	0.5125
Zhang et.al, 2008	Logit	-	0.3248	0.5388	0.4067
Zhang et.al, 2008	SVM	-	0.8017	0.3327	0.4653

Table 1: Comparison of Evaluation Metrics (* represents F1 Micro)

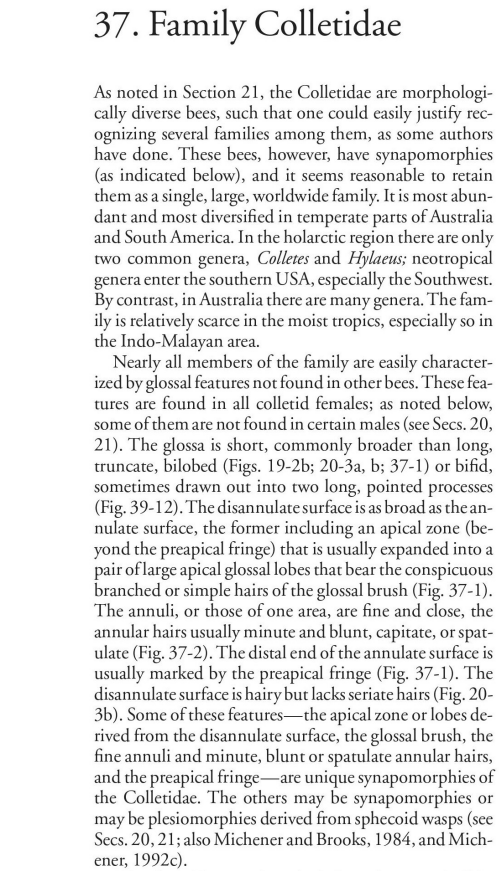
3 Work Done, Approaches & Results

3.1 Methodology

3.1.1 Data Collection

One significant issue that we faced initially, was the limited availability of a dataset that consisted of the physical traits of different species of bees. Thus, we decided to scrape and collect data from several websites, research papers, field guides, books and articles and create our own database.

We implemented Image to Text conversion methods using python's cv2 library – pytesseract and ocrmypdf, to extract text from older editions of books, for which a text format was not available. One book that we used majorly for the extraction of data was 'The Bees of the World' (Second Edition) by Charles D. Michener [21]. Figure 1 shows a snippet of a page from the book (Figure 1a) and the text extracted from it (Figure 1b). We can see that the text was extracted with desirable outcomes.



37. Family Colletidae

As noted in Section 21, the Colletidae are morphologically diverse bees, such that one could easily justify recognizing several families among them, as some authors have done. These bees, however, have synapomorphies (as indicated below), and it seems reasonable to retain them as a single, large, worldwide family. It is most abundant and most diversified in temperate parts of Australia and South America. In the holarctic region there are only two common genera, *Colletes* and *Hylaeus;* neotropical genera enter the southern USA, especially the Southwest. By contrast, in Australia there are many genera. The family is relatively scarce in the moist tropics, especially so in the Indo-Malayan area.

Nearly all members of the family are easily characterized by glossal features not found in other bees. These features are found in all colletid females; as noted below, some of them are not found in certain males (see Secs. 20, 21). The glossa is short, commonly broader than long, truncate, bilobed (Figs. 19-2b; 20-3a, b; 37-1) or bifid, sometimes drawn out into two long, pointed processes (Fig. 39-12). The disannulate surface is as broad as the annulate surface, the former including an apical zone (beyond the preapical fringe) that is usually expanded into a pair of large apical glossal lobes that bear the conspicuous branched or simple hairs of the glossal brush (Fig. 37-1). The annuli, or those of one area, are fine and close, the annular hairs usually minute and blunt, capitate, or spatulate (Fig. 37-2). The distal end of the annulate surface is usually marked by the preapical fringe (Fig. 37-1). The disannulate surface is hairy but lacks seriate hairs (Fig. 20-3b). Some of these features—the apical zone or lobes derived from the disannulate surface, the glossal brush, the fine annuli and minute, blunt or spatulate annular hairs, and the preapical fringe—are unique synapomorphies of the Colletidae. The others may be synapomorphies or may be plesiomorphies derived from sphecoid wasps (see Secs. 20, 21; also Michener and Brooks, 1984, and Michener, 1992c).

(a) Excerpt from the Book

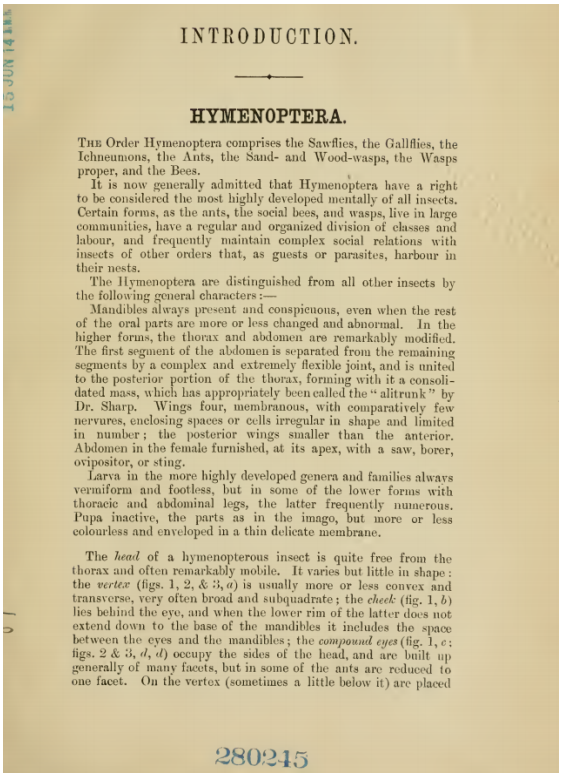
```
37. Family Colletidae

As noted in Section 21, the Colletidae are morphologi-
cally diverse bees, such that one could easily justify rec-
ognizing several families among them, as some authors
have done. These bees, however, have synapomorphies
(as indicated below), and it seems reasonable to retain
them asa single, large, worldwide family. It is most abun-
dant and most diversified in temperate parts of Australia
and South America. In the holarctic region there are only
two common genera, Colletes and Hylaeus; neotropical
genera enter the southern USA, especially the Southwest.
By contrast, in Australia there are many genera. The fam-
ily is relatively scarce in the moist tropics, especially so in
the Indo-Malayan area.

Nearly all members of the family are easily character-
ized by glossal features not found in other bees. These fea-
tures are found in all colletid females; as noted below,
some of them are not found in certain males (see Secs. 20,
21). The glossa is short, commonly broader than long,
truncate, bilobed (Figs. 19-2b; 20-3a, b; 37-1) or bifid,
sometimes drawn out into two long, pointed processes
(Fig. 39-12). The disannulate surface is as broad as the an-
nulate surface, the former including an apical zone (be-
yond the preapical fringe) that is usually expanded into a
pair of large apical glossal lobes that bear the conspicuous
branched or simple hairs of the glossal brush (Fig. 37-1).
The annuli, or those of one area, are fine and close, the
annular hairs usually minute and blunt, capitate, or spat-
ulate (Fig. 37-2). The distal end of the annulate surface is
usually marked by the preapical fringe (Fig. 37-1). The
disannulate surface is hairy but lacks seriate hairs (Fig. 20-
3b). Some of these features—the apical zone or lobes de-
rived from the disannulate surface, the glossal brush, the
fine annuli and minute, blunt or spatulate annular hairs,
and the preapical fringe—are unique synapomorphies of
the Colletidae. The others may be synapomorphies or
may be plesiomorphies derived from sphecoid wasps (see
Secs. 20, 21; also Michener and Brooks, 1984, and Mich-
ener, 1992c).
```

(b) Converted Text

Figure 1: Image to text conversion on a section of the book [21] by C.D. Michener

Another book that we wanted to extract our data from was 'The Fauna of British India, Including Ceylon and Burma. Hymenoptera (Vol. 1) Wasps and Bees' [8] by Charles Thomas Bingham. This book was only available in the scanned form of the physical copy, which made it a bit more challenging to extract the text accurately, as compared to The Bees of The World. Data cleaning was required to reach the desirable results. Figure 2 shows the comparison

between the original text (Figure 2a) and the extracted and processed text (Figure 2b).



INTRODUCTION.

HYMENOPTERA.

THE Order Hymenoptera comprises the Sawflies, the Gallflies, the Ichneumons, the Ants, the Sand- and Wood-wasps, the Wasps proper, and the Bees.

It is now generally admitted that Hymenoptera have a right to be considered the most highly developed mentally of all insects. Certain forms, as the ants, the social bees, and wasps, live in large communities, have a regular and organized division of classes and labour, and frequently maintain complex social relations with insects of other orders that, as guests or parasites, harbour in their nests.

The Hymenoptera are distinguished from all other insects by the following general characters :—

Mandibles always present and conspicuous, even when the rest of the oral parts are more or less changed and abnormal. In the higher forms, the thorax and abdomen are remarkably modified. The first segment of the abdomen is separated from the remaining segments by a complex and extremely flexible joint, and is united to the posterior portion of the thorax, forming with it a consolidated mass, which has appropriately been called the "alitrunk" by Dr. Sharp. Wings four, membranous, with comparatively few nervures, enclosing spaces or cells irregular in shape and limited in number; the posterior wings smaller than the anterior. Abdomen in the female furnished, at its apex, with a saw, borer, ovipositor, or sting.

Larva in the more highly developed genera and families always vermiform and footless, but in some of the lower forms with thoracic and abdominal legs, the latter frequently numerous. Pupa inactive, the parts as in the imago, but more or less colourless and enveloped in a thin delicate membrane.

The *head* of a hymenopterous insect is quite free from the thorax and often remarkably mobile. It varies but little in shape: the *vertex* (figs. 1, 2, & 3, *a*) is usually more or less convex and transverse, very often broad and subquadrate; the *cheek* (fig. 1, *b*) lies behind the eye, and when the lower rim of the latter does not extend down to the base of the mandibles it includes the space between the eyes and the mandibles; the *compound eyes* (fig. 1, *c*; figs. 2 & 3, *d*, *d*) occupy the sides of the head, and are built up generally of many facets, but in some of the ants are reduced to one facet. On the vertex (sometimes a little below it) are placed

280245

(a) Excerpt from the Book

(b) Converted Text

Figure 2: Image to text conversion on a section of the book [8] by C.T. Bingham

To scrape data from online resources, we used python's BeautifulSoup library. Data was scraped and compiled from several websites [1, 9, 10, 17, 24, 26, 32, 37].

After compiling all the resources the format of the dataset was decided − the title i.e. the name of the species followed by the description of the species containing the physical characteristics of the bees − which would later be compared with the input query text.

Finally, we prepared a dataset containing 102 species of bees and their detailed description in order to create the algorithm and validate it using those samples. Screenshot of the dataset is given in Figure 3.

On analysis of our dataset, we found that the words 'species', 'bee' and 'genus' were the most frequently occurring words. The 25 most common words have been shown in the form of a wordcloud in Figure 4.

3.1.2 Data Preprocessing

The obtained text descriptions of the species were then preprocessed using the steps mentioned below (also see Figure 5):

- **Tokenization:** the string containing the text was broken down into smaller strings by separating them based on the space character (' '). The obtained description was then

Mining Bees (Andrena)	Some of the first bees to emerge in spring, members of the genus Andrena vary greatly in size and appearance. Females can be recognized by patches of velvety hairs between the eyes. Mining bees carry pollen on their hind legs and on hairs between the abdomen and the thorax. Preferred Crops: Apple, cherry, peach, and pear. Nesting Behavior: Solitary. Andrena nest in small tunnels in the ground.
Mason Bees (Osmia)	Like leafcutter bees, Osmia have large jaws and big heads. They range in color from metallic blue to green, occasionally black. Their abdomens often have a rounded appearance. Osmia are called mason bees because they use mud to make their nest cells. Several species are managed for agricultural production. Mason bees carry pollen on specialized hairs on the abdomen.
Small Carpenter Bees (Ceratina)	Certina are small mostly hairless bees that vary in color from dark metallic blue to green. They emerge in the spring and stay active until fall. Small carpenter bees have rudimentary pollen-carrying hairs. They may transport pollen by swallowing it and regurgitating it back at the nest. This behavior has been observed in primitive bees. Preferred Crops: Apple, cane berries, cherry, pear, and strawberry.
Long-Horn Bees (Eucerini)	They are solitary bees with about 500 species in 32 genera in the tribe Eucerini. Long-horn bees have hairy bodies and legs with black and tan markings. One common distinguishable feature is their long antennae. Long-horn bees are commonly found feeding on pollen on sunflowers. These bees don't produce honey and live a solitary existence where they nest in small tunnels. Long-horn bees generally have pale bands on black fuzzy bodies and two long antennae. Their six legs are hairy and a dark tan color.

Figure 3: Snippet of the Compiled Dataset

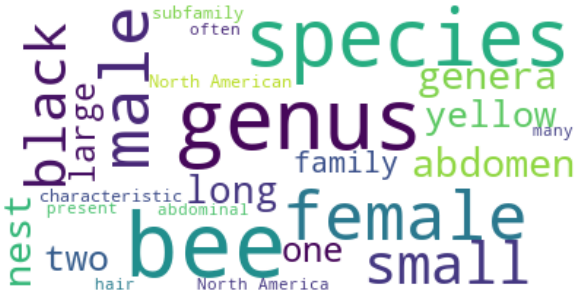


Figure 4: 25 of the most frequently used words in our dataset

stored in an array.

- **Lower casing:** the words in the array were then converted to the lowercase in order to maintain uniformity among the words which were the same but differed in their capitalisation. For instance, "NorthWestern", "Northwestern" and "northwestern".

- **Removal of extra spaces:** the extra space between words needed to be removed as it could have led to confusion in subsequent processes (ex: whether the white space character is a part of the word or not). We felt that this could also later help us in bi-gram indexing, which is a future aspect of this project.

- **Removal of stop-words:** irrelevant words which do not add to the meaning of the sentence (such as: the, a, then, etc.) had to be removed so that only the meaningful keywords remained in the dataset.

- **Lemmatization:** the remaining words were then converted to their root form. This means that words like 'pollinated', and 'pollinating' were represented by one word - 'pollinate' - and hence were considered the same.

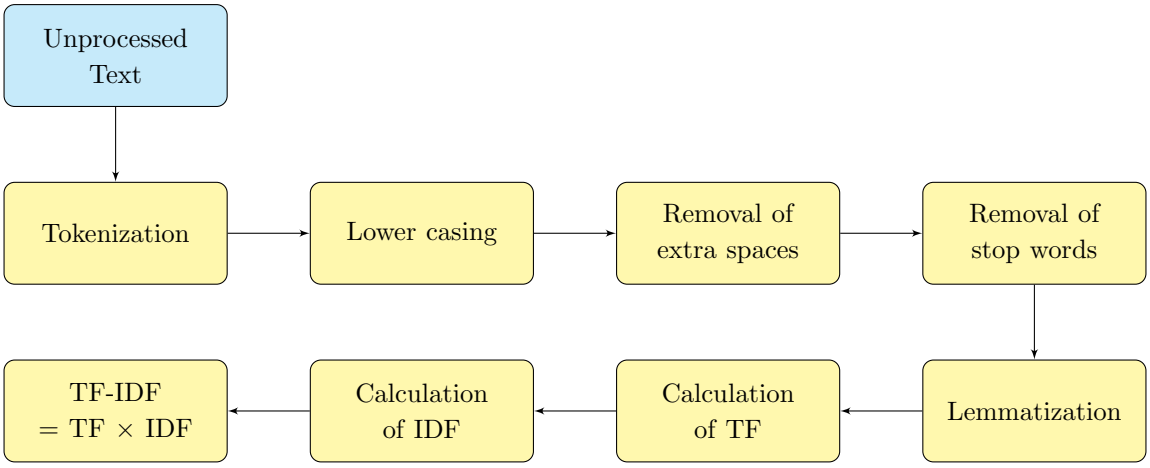


Figure 5: Data Preprocessing and Score Calculation Steps

Our final cleaned/processed vocabulary list consisted of 1842 words. The IDF distribution of these words is shown in Figure 6. We see that a large majority of the words have their IDF score between 4.0 and 4.5. This shows that a majority of the words in our dataset were rare terms.

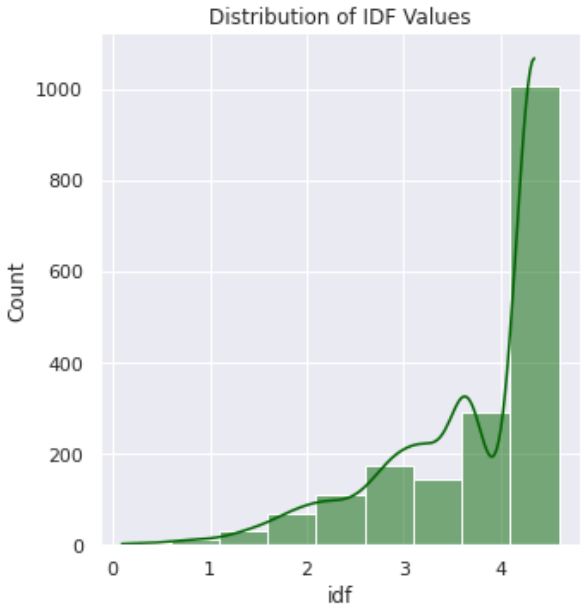


Figure 6: Distribution of IDF

3.1.3 Rank Prediction

Term Frequency - Inverse Document Frequency (TF-IDF) is a numerical statistic that is used to represent the importance of a word in a document in a collection or corpus.

Term frequency is the measurement of how often a term occurs within a document, whereas Inverse Document Frequency is the measure of the rareness of a term. Thus, works as an important toolkit for scoring and ranking a document's relevance given a user query (in our

case, features of the bee).

In order to create a predicted rank list, we applied three ranking algorithms to the list of queries and merged them all into one, to finally return the top 15 most likely species. The three rank lists were individually made by raw TF-IDF extraction, binary TF-IDF extraction and log-normalized TF-IDF extraction, respectively.

1. **Raw TF-IDF**

   Once the preprocessed data was obtained, the next step was to calculate the term frequency for each of the terms in each of the sample classes. The measure of the originality of a word can be compared by comparing the number of times a word appears in a text with the number of samples the word appeared in. It shows the importance of the word in the document as well. To compute the Term Frequency,

   $$\text{TF} = \frac{\text{Number of times the term occurred in the text}}{\text{Total number of words in the text}}$$

   On the other hand, IDF i.e. the Inverse Document Frequency describes the amount of information a word is providing across the sample documents.

   $$\text{IDF} = \log\left(\frac{\text{Total number of texts}}{\text{Number of texts in which the selected term appeared}}\right)$$

   Therefore,

   $$\boxed{\text{TF-IDF} = \text{TF} \times \text{IDF}}$$

   Now, the TF-IDF score of a word shows how significant it is to the text and here, it will tell us how relevant the word is in describing the physical characteristics of the specified bee species.

2. **Binary TF-IDF**

   After the vanilla TF-IDF calculation, in order to reduce the spontaneity in the algorithm, we calculate the rank list through Binary TF-IDF extraction.

   In this method, the TF calculation is different from the raw TF-IDF calculation. If the text contains the word, the binary TF is 1, otherwise, the binary TF is 0. Thus, no matter how many times the word occurs in a document, its binary TF will always be 1.

   $$\text{Binary TF} = \max(1, \text{Number of times the term occurred in the text})$$

   The IDF is calculated in the same way as above, and thus, the final scores are calculated by the formula:

   $$\boxed{\text{Binary TF-IDF} = \text{Binary TF} \times \text{IDF}}$$

3. **Log Normalised TF-IDF**

   In this method, 1 is added to normalise the term and prevent the value from reaching large extremes. The log normalized TF is calculated by the following formula:

   $$\text{Log Normalised TF} = \log(1 + \text{Number of times the term occurred in the text})$$

   The final score is calculated by multiplying the log normalised TF with IDF, as done in the previous methods.

$$\boxed{\text{Log Normalized TF-IDF} = \text{Log Normalized TF} \times \text{IDF}}$$

Thus, the final rank list is prepared by voting the candidates of all the above-obtained rank lists and merging them.

3.1.4 Machine Learning Models

For text/feature classification we also ran the following models on our dataset, using python's sklearn library – Multinomial Naive Bayes, Support Vector Classifier, Random Forest Classifier, Logistic Regression, K-Nearest Neighbours Classifier and Multilayer Perceptron Classifier. The accuracy obtained with each model has been reported under Results in Table 2.

3.2 Graphical User Interface

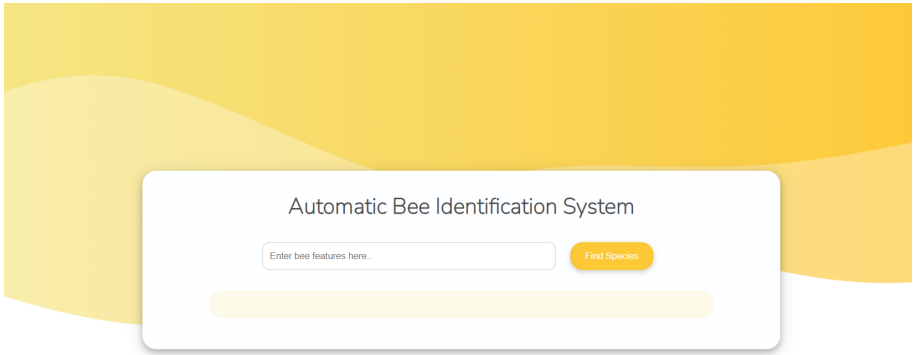


Figure 7: Design of the GUI

To display the working of our model, we created a Graphical User Interface (GUI) using Flask. It consists of a textbox where the user can enter the features/queries of the bee they want to identify the species of. On clicking on the 'Find Species' button, a list of 15 most likely species, that fit the query description, will be displayed. Figure 7 shows the GUI. The app can be viewed here: bees-identifier.herokuapp.com

3.3 Results

We tested our model against four queries. The queries entered for ranking were selected from the sample dataset itself and the terms in the description were randomly selected and then shuffled in order to introduce some randomness while retrieval. Four such queries were created and all the 3 ranking algorithms were run for each of them.

Ex: One of the queries was:

```
query = 'Some of the first bees to emerge in spring, vary greatly in size and
appearance.  Females have patches of velvety hairs between the eyes.'
```

We ran the three algorithms, which returned the following results:

```
Raw TF-IDF Algorithm
['Sawflies and wood-wasps (Symphyta)',
 'Asian honey bee (Apis cerana)',
 'Family: Oxaeidae',
 'Striped Green Sweat Bees (Agapostemon)',
 'Mason Bees (Osmia)',
 'Sweat Bees (Lasioglossum)',
 'Carpenter bees ',
 'Hylaeus',
 'Microthurge',
 'Genus: Colletes Latreille',
 'Subfamily: Dasypodinae ',
 'Small Carpenter Bees (Ceratina)',
 'Long-horned Bees (Melissodes)',
 'Miner bees (Andrenidae)',
 'Mining Bees (Andrena)']

Boolean TF-IDF Algorithm
['Sawflies and wood-wasps (Symphyta)',
 'Asian honey bee (Apis cerana)',
 'Family: Oxaeidae',
 'Striped Green Sweat Bees (Agapostemon)',
 'Mason Bees (Osmia)',
 'Sweat Bees (Lasioglossum)',
 'Carpenter bees ',
 'Hylaeus',
 'Microthurge',
 'Genus: Colletes Latreille',
 'Subfamily: Dasypodinae ',
 'Small Carpenter Bees (Ceratina)',
 'Long-horned Bees (Melissodes)',
 'Miner bees (Andrenidae)',
 'Mining Bees (Andrena)']

Log-Normalised TF-IDF
['Sweat Bees (Lasioglossum)',
 'Striped Green Sweat Bees (Agapostemon)',
 'Sawflies and wood-wasps (Symphyta)',
 'Mason Bees (Osmia)',
 'Dufoureinae',
 'Genus: Colletes Latreille',
 'Family: Oxaeidae',
 'Subfamily: Hylaeinae',
 'Hylaeus',
 'Genus: Dianthidium Cockerell',
 'Subfamily: Dasypodinae ',
 'Small Carpenter Bees (Ceratina)',
 'Long-horned Bees (Melissodes)',
 'Miner bees (Andrenidae)',
 'Mining Bees (Andrena)']
```

Here, the most probable species are the ones at the bottom. As we can see, all the rank lists have the same number one contender – Mining Bess (Andrena). This meant that the final rank list must also have this as the number one too. The rest of the ranks were decided by merging the three lists and weighing the elements on the basis of their comparative ranks.

Therefore the final ranking comes out to be:

```
['Mining Bees (Andrena)',
 'Miner bees (Andrenidae)',
 'Long-horned Bees (Melissodes)',
 'Small Carpenter Bees (Ceratina)',
 'Subfamily: Dasypodinae ',
 'Genus: Colletes Latreille',
 'Hylaeus',
 'Microthurge',
 'Carpenter bees ',
 'Mason Bees (Osmia)',
 'Sweat Bees (Lasioglossum)',
 'Family: Oxaeidae',
 'Genus: Dianthidium Cockerell',
 'Subfamily: Hylaeinae',
 'Striped Green Sweat Bees (Agapostemon)']
```

It is now confirmed that the number one element on the list is the same in the final list as in the intermediate lists.

We checked this for all the four sample queries that were generated from our dataset by jumbling the text content. For each of these queries, the most probable species was predicted correctly. In Figure 8 we have shown the output using our graphical interface.

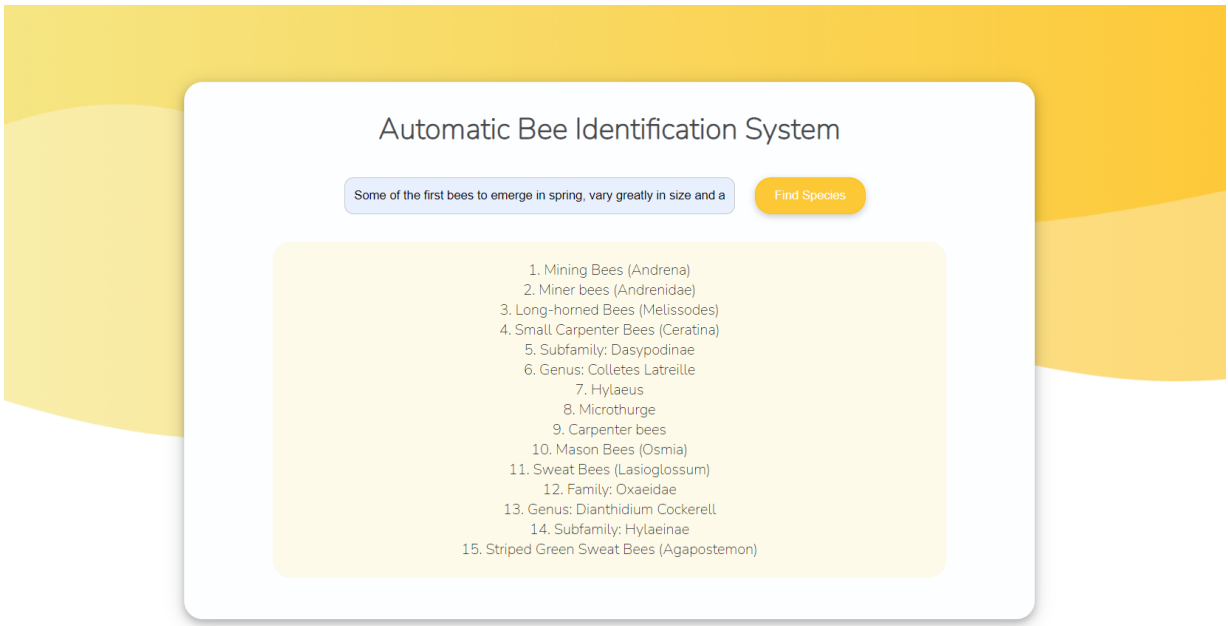


Figure 8: GUI for the the given query

On running the models mentioned in Subsection 3.1.4, we obtained the following accuracies:

Model	Accuracy (%)
Multinomial Naive Bayes	77.78
Random Forest Classifier	55.54
K-Nearest Neighbours	11.12
Logistic Regression	88.89
SVM Classifier	77.78
MLP Classifier	77.78

Table 2: Accuracy for the Models

Figure 9 helps us visualise our results.

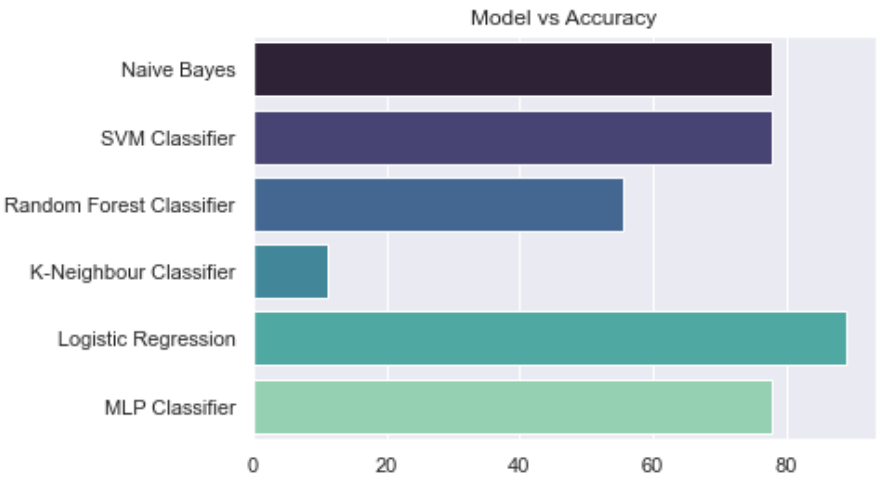


Figure 9: Model vs Accuracy

We see that Logistic Regression gives the highest accuracy of **88.89%** and K-Nearest Neighbours Classifier gives us the lowest accuracy (11.12%). These results are comparable with the evaluation metrics seen in the reviewed literature.

Thus, we were able to classify the query of features.

4 Source Code

The source code and dataset are available in this GitHub repository.

Instructions to set up the environment and perform predictions are present in the README present in the same repository.

5 Future Work

- Compilation of a larger dataset for better results.
- Application of Bi-grams/multi-grams indexing for ranked retrieval.
- Computer Vision techniques to identify the species of bees from an image
  - In this project, we have only taken up text-based queries to output a list of possible species. However, in the future one can also take the image of a bee as the input and then predict the list of possible species.
- Use deep learning models like LSTM and RNNs for better predictions.

References

[1] *About Bees*. URL: https://idtools.org/id/bees/exotic/bees_classification.php.

[2] Alper. *NLP: Classification and Recommendation Project*. July 2020. URL: https://towardsdatascience.com/nlp-classification-recommendation-project-cae5623ccaae.

[3] B. Armouty and S. Tedmori. 'Automated keyword extraction using support vector machine from Arabic news documents'. In: *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE. 2019, pp. 342–346.

[4] R. Barzilay and M. Elhadad. 'Using lexical chains for text summarization'. In: *Advances in automatic text summarization* (1999), pp. 111–121.

[5] G. Bedi. *Simple guide to Text Classification(NLP) using SVM and Naive Bayes with Python*. July 2020. URL: https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34.

[6] S. Beliga. 'Keyword extraction: a review of methods and approaches'. In: *University of Rijeka, Department of Informatics, Rijeka* (2014), pp. 1–9.

[7] S. K. Bharti and K. S. Babu. 'Automatic keyword extraction for text summarization: A survey'. In: *arXiv preprint arXiv:1704.03242* (2017).

[8] C. Bingham. 'The Fauna of British India, Including Ceylon and Burma. Hymenoptera-Vol. 1. Wasps and Bees.' In: (1897).

[9] S. Bob. *HYMENOPTERA Bees, Wasps, Sawflies &; Ants*. May 2014. URL: http://www.bobs-bugs.info/2014/01/02/hymenoptera-bees-wasps-ants-etc.

[10] O. M. Carril, T. Griswold, J. Haefner and J. S. Wilson. *Wild bees of Grand Staircase-Escalante National Monument: richness, abundance, and spatio-temporal beta-diversity*. Nov. 2018. URL: https://riversedgewest.org/sites/default/files/resource-center-documents/Carril%5C%20et%5C%20al%5C%202018%5C%20Bees%5C%20GSENM%5C%20%5C%281%5C%29.pdf.

[11] M. Chandra, V. Gupta and S. K. Paul. 'A statistical approach for automatic text summarization by extraction'. In: *2011 International Conference on Communication Systems and Network Technologies*. IEEE. 2011, pp. 268–271.

[12] A. Chauhan. *Natural Language Processing: Intelligent Search through text using Spacy and Python*. Sept. 2020. URL: https://towardsdatascience.com/natural-language-processing-document-search-using-spacy-and-python-820acdf604af.

[13] J. D. Cohen. 'Highlights: Language-and domain-independent automatic indexing terms for abstracting'. In: *Journal of the American society for information science* 46.3 (1995), pp. 162–174.

[14] A. Hulth. 'Improved automatic keyword extraction given more linguistic knowledge'. In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003, pp. 216–223.

[15] T. Jo. 'NTC (Neural Text Categorizer): neural network for text categorization'. In: *International Journal of Information Studies* 2.2 (2010), pp. 83–96.

[16] *Keyword Extraction: A Guide to Finding Keywords in Text*. URL: https://monkeylearn.com/keyword-extraction.

[17] S. Kincaid. *Common Bee Pollinators of Oregon Crops*. 2017. URL: oregon.gov.

[18] H. P. Luhn. 'A statistical approach to mechanized encoding and searching of literary information'. In: *IBM Journal of research and development* 1.4 (1957), pp. 309–317.

[19] D. Mahata, R. R. Shah, J. Kuriakose, R. Zimmermann and J. R. Talburt. 'Theme-weighted ranking of keywords from text documents using phrase embeddings'. In: *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE. 2018, pp. 184–189.

[20] Y. Matsuo and M. Ishizuka. 'Keyword extraction from a single document using word co-occurrence statistical information'. In: *International Journal on Artificial Intelligence Tools* 13.01 (2004), pp. 157–169.

[21] C. D. Michener. *The bees of the world*. Vol. 1. JHU press, 2000.

[22] L. N. Minh, A. Shimazu, H. P. Xuan, B. H. Tu and S. Horiguchi. 'Sentence extraction with support vector machine ensemble'. In: (2005).

[23] T. Mouratis and S. Kotsiantis. 'Increasing the accuracy of discriminative of multinomial bayesian classifier in text classification'. In: *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*. IEEE. 2009, pp. 1246–1251.

[24] A. Müller. *Palaearctic Hoplitis bees of the subgenera Chlidoplitis and Megahoplitis (Megachilidae, Osmiini): biology, taxonomy and key to species*. URL: https://www.biotaxa.org/Zootaxa/article/view/zootaxa.3765.2.4.

[25] M. R. Murty, J. Murthy, P. P. Reddy and S. C. Satapathy. 'Statistical approach based keyword extraction aid dimensionality reduction'. In: *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012*. Springer. 2012, pp. 445–452.

[26] *Palaearctic Osmiine Bees*. URL: https://blogs.ethz.ch/osmiini/phylogeny-and-classification.

[27] L. Rabiner and B. Juang. 'An introduction to hidden Markov models'. In: *ieee assp magazine* 3.1 (1986), pp. 4–16.

[28] J. Ramos et al. 'Using tf-idf to determine word relevance in document queries'. In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[29] S. Rose, D. Engel, N. Cramer and W. Cowley. 'Automatic keyword extraction from individual documents'. In: *Text mining: applications and theory* 1 (2010), pp. 1–20.

[30] N. L. I. Rusli, A. Amir, N. A. H. Zahri and R. B. Ahmad. 'Snake species identification by using natural language processing'. In: *Indonesian Journal of Electrical Engineering and Computer Science* 13.3 (2019), pp. 999–1006.

[31] G. Salton, A. Singhal, M. Mitra and C. Buckley. 'Automatic text structuring and summarization'. In: *Information processing & management* 33.2 (1997), pp. 193–207.

[32] *The Most Beneficial Types of Bees (With Identification Guide and Pictures)*. Apr. 2021. URL: https://leafyplace.com/types-of-bees.

[33] J. R. Thomas, S. K. Bharti and K. S. Babu. 'Automatic keyword extraction for text summarization in e-newspapers'. In: *Proceedings of the international conference on informatics and analytics*. 2016, pp. 1–8.

[34] S. Vivek. *Automated Keyword Extraction from Articles using NLP*. Dec. 2018. URL: https://medium.com/analytics-vidhya/automated-keyword-extraction-from-articles-using-nlp-bfd864f41b34.

[35] S. Wang, Z. Chen, B. Liu and S. Emery. 'Identifying search keywords for finding relevant social media posts'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016.

[36] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning. 'Kea: Practical automated keyphrase extraction'. In: *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI global, 2005, pp. 129–152.

[37] *Wood Wasps and Sawflies*. URL: https://www.amentsoc.org/insects/fact-files/orders/hymenoptera-symphyta.html.

[38] C. Zhang. 'Automatic keyword extraction from documents using conditional random fields'. In: *Journal of Computational Information Systems* 4.3 (2008), pp. 1169–1180.

[39] H. Zhang, D. Mahata, S. Shahid, L. Mehnaz, S. Anand, Y. Singla, R. R. Shah and K. Uppal. 'Identifying offensive posts and targeted offense from twitter'. In: *arXiv preprint arXiv:1904.09072* (2019).

[40] K. Zhang, H. Xu, J. Tang and J. Li. 'Keyword extraction using support vector machine'. In: *international conference on web-age information management*. Springer. 2006, pp. 85–96.

Resources Used for Data Scraping

[1] *About Bees*. URL: https://idtools.org/id/bees/exotic/bees_classification.php.

[9] S. Bob. *HYMENOPTERA Bees, Wasps, Sawflies &; Ants*. May 2014. URL: http://www.bobs-bugs.info/2014/01/02/hymenoptera-bees-wasps-ants-etc.

[10] O. M. Carril, T. Griswold, J. Haefner and J. S. Wilson. *Wild bees of Grand Staircase-Escalante National Monument: richness, abundance, and spatio-temporal beta-diversity*. Nov. 2018. URL: https://riversedgewest.org/sites/default/files/resource-center-documents/Carril%5C%20et%5C%20al%5C%202018%5C%20Bees%5C%20GSENM%5C%20%5C%281%5C%29.pdf.

[17] S. Kincaid. *Common Bee Pollinators of Oregon Crops*. 2017. URL: oregon.gov.

[24] A. Müller. *Palaearctic Hoplitis bees of the subgenera Chlidoplitis and Megahoplitis (Megachilidae, Osmiini): biology, taxonomy and key to species*. URL: https://www.biotaxa.org/Zootaxa/article/view/zootaxa.3765.2.4.

[26] *Palaearctic Osmiine Bees*. URL: https://blogs.ethz.ch/osmiini/phylogeny-and-classification.

[32] *The Most Beneficial Types of Bees (With Identification Guide and Pictures)*. Apr. 2021. URL: https://leafyplace.com/types-of-bees.

[37] *Wood Wasps and Sawflies*. URL: https://www.amentsoc.org/insects/fact-files/orders/hymenoptera-symphyta.html.