**Abstract**

Due to the popularity of review services, such as yelp, and social media, like instagram and twitter, there is a vast amount of review data available to us, however, much of the data is contained forms of information that can only be interpreted by humans or an AI system, like photos or blocks of natural language. Our aim is to develop an AI system that can extract usable information from the natural language in a review or social media post. We will be focusing on a small subset of that problem. Given a review or post resulting from a search for an experience the AI system will be asked to determine if the post indeed concerns the experience in question. The experience that we will be focusing on is watching the Eagles game at a bar. Our goal will be to use the AI system to check if a review was written about an Eagles game or if the review simply contained a word or phrase that matched the search term, such as the word “Eagles”.

**1 Introduction**

Philadelphia based startup, VyB ([www.myvyb.io](http://www.myvyb.io)) has recently launched an Android and iOS application. It is a real-time discovery application for social experiences, similar to Yelp - but with data from users having experiences in real time. Our aim is to develop back-end capabilities for VyB that will allow it to classify experiences by processing the natural language found in real-time social media posts.

We have developed a custom Bayesian Classifier used to analyze review and social media post data regarding a specific topic, to help identify whether a post is regarding a general topic. Delving a bit deeper into the problem that live experiences will help in academia - we are unable to create accurate prediction models to see customer experience oscillate from multiple real time data sources. In a commercial setting, this correlates to millions of data points lost in a dark abyss. We plan to use these data points to help businesses understand consumer behavior in a timely fashion. We will do this by identifying datasets from review providers such as Yelp, Google Reviews, and Trip Advisor. We will couple this with metadata produced from products like Twitter and Instagram. After normalizing data and creating a robust digestion pipeline into our classifier, we use NLP and data modeling techniques to identify keywords from an example business. Additionally, if we are looking to see if there is a “football game” playing at a “bar”, we will isolate those topics and run regression models from our review datasets. We expect output in terms of confidence between 0 and 1. In this case, 1 might mean that a football game is definitely playing at the bar and 0 means the game is definitely not playing at the bar. We are going to test this hypothesis by live examples of known scenario and cross referencing them to the data outputted by our classifier.

This application of an advanced bayesian classifier can assist consumers in their discovery finding process as they search for specific keywords. There has been quite a bit of past work on topic modeling and understanding large datasets of reviews from sources like Yelp, Google Reviews, and more. Much of this work has been around understanding the semantics of a review, to convey the positivity or negativity. A majority of past work on this topic is related to understanding the reviews and filter them based on certain pre-set criteria. Examples of this might include seeing if certain restaurants have certain food items, or seeing if certain reviews connote an overall positive or negative tone. Continuing on, we also saw examples of classifying Yelp datasets using Bayesian scoring methods - to see the accuracy of ratings. This particular example (Bechon, et. all) came out of study at Stanford in 2010. It concluded that the relation of Bayesian scoring vs actual rating was actually closer and more accurate than anticipated.

In this paper, we identify the data collection approach, in Section 3, to help us understand the metadata available from posts that we can use to train and then eventually test our dataset. This dataset was attained using Twitter’s API, and we delve deeper into the restrictions, benefits, and structure of this raw data. In Section 4, we discuss the custom bayesian classifier we created, and the novel approach to create this from scratch. Later in Section 5, we discuss the results obtained from training our classifier with the data collected from the Twitter API, and further testing the bayesian classifier with almost 60 files we created - split into data pertaining to the “Eagles” football team, and the other half about non “Eagles” data, and data pertaining to the flying bird, Eagle. Further, in Section 6, we explain some of the challenges we faced throughout the multi-step process in data collection & optimizing our classifier. Lastly we discuss the future work that we plan to do as we continue this project as an R&D effort within VyB Technologies, LLC. with the goal to create a custom solution to keyword finding in the ratings and reviews space.

**3 Data Collection Approach**

Ultimately, to train and use our classifier, we would like to use data from multiple sources including Review and Social Media websites- Yelp, Facebook, Twitter, Instagram, and Google Reviews are all possibilities. Due to the time constraints of this project we limited our data to Twitter statuses. The data was collected using the Twitter-API Python wrapper, Tweepy[11]. Based on restrictions on the Twitter-API, tweets were only collected from a time period of seven days before the search is performed. The text in each tweet was saved as an individual text file.

Two categories of data were collected, “Eagles” and non-Eagles”- respectively referring to tweets about the professional football team and tweets about any other topic. Across both categories, 1290 tweets were collected. Tweets were categorized based on the search query used to return the tweets. Search queries were written with the aim of maximizing the accuracy of the data categorization. For example, the “Eagles” data category was filled with tweets returned when the search query included “eagles football” or other similar text strings.

**4 Bayesian Classifier & Training Data**

The Twitter data was used to construct dictionaries that kept track of the unique words in each data category along with their frequency and presence. For this project, frequency is the total number of times the word is used in the data set and presence is the total number of documents in which the word is used.

Our Bayes Classifier uses the data dictionaries to calculate the probability that any given text string belongs to either the “Eagles” of the “non-Eagles” category. The text string is categorized based on which probability is higher. If neither probability is sufficiently higher than the other, the text string is classified as neutral.

**5 Results**

We report here the words that appeared in each data set with the highest frequency and presence (figure 1). While both categories are populated with words such as “the”, “RT”, “to”, and “in” that do not indicate which category a string will belong to, other words such as “Manning”, “Dolphins”, “nature”, and “scenery” convey more meaning.

A testing data set was constructed using human-verified text strings belonging to both categories. Each text string was classified while the following statistics were tracked and reported-

Precision- 0.8334

Recall- 0.6667

F-Measure- 0.7407

In 2004, Chapman et al [3] published the results of a similar classifier. They classified free-text triage complaints into seven categories. Using advanced techniques, they were able to achieve precisions of up to 0.97. Clearly, our classifier needs some upgrades to improve our precision of 0.8334.

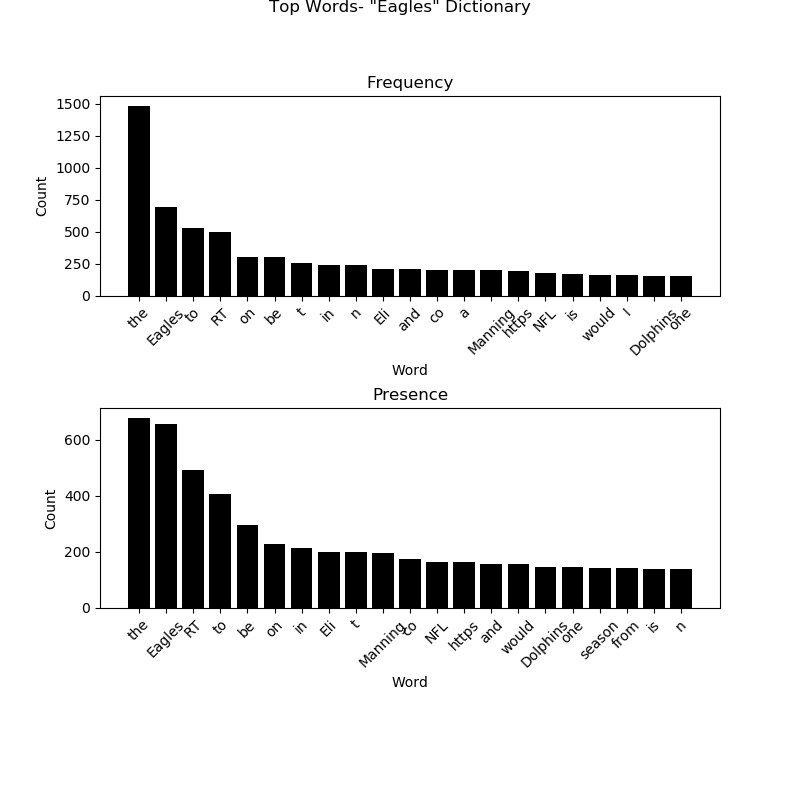
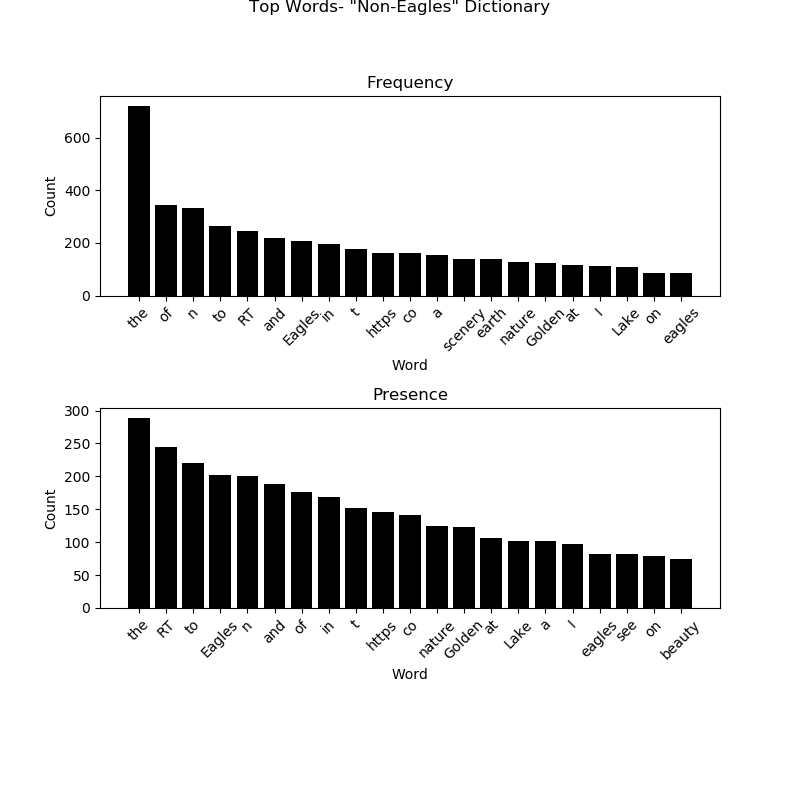


Figure 1. Frequency and Presence of the top 20 most-used words in the training data sets

**6 Challenges**

The goal of this bayesian classifier was to identify whether a given datapoint contains any reference to the sports team, Eagles or not. In doing so, we faced many challenges along the way in various aspects of this study. We faced adversities in the data collection process, the training of the data, and later during the topic model and Natural Language Understanding phase.

Initially, Tying to collect data proved to be challenging due to the various restrictions on both the data availability and data transfer. In this process, we tried to discover data from multiple sources including Google Reviews API, Yelp API and Twitter API. Some restrictions that we came across stemmed from the various schemas that were not optimal for the Bayesian classifier. After going through various API options, we picked Twitter due to the convenience the data source provided. Twitter’s API provided its own challenges including a restriction on the number of results given per search query (15). This proved to be challenging due to the time consuming process of paginating each individual query to make sure we could get a plethora of training data. This data collection process created time restrictions that were initially not expected. In order to mitigate this process, we created a script to

accurately retrieve the data in organized schemas that can be seen in Figure 2.



Figure 2. A single query from the Twitter API Call, searching for “Eagles” keyword in chronological order.

As we can see that each query item contains quite a bit of data that may not be useful for our purposes. Thus, we had to manually and meticulously parse the data, retrieving only the parameters that are necessary for the proper training of the bayesian classifier.

Additionally, another challenge came with the Natural Language Understanding aspect of this study. We try to understand ia sentence structure and its relevance given a topic. Topic modeling to a deeper level is quite difficult and we tried our best to understand abnormalities within data. While training the data we took instances of “Eagles” and “Not Eagles” data - however, there were instances where test data was difficult to understand due to the varied usage of the word “Eagle”. In the next section we describe additional work we hope to implement in the future to improve the efforts of this study.

**7 Future Work**

While we created a robust bayesian classifier, there are many possible areas of improvement that can help the overall initiative of this study. First and foremost, increasing the dataset can certainly increase the complexity and accuracy of the classifier. Our hope is to understand how to obtain data from multiple types of datasets such as Google Reviews, Yelp, Trip and Trip Advisor. Additionally, we would like to normalize these data schemas into a singular form to better digest the data into the classifier. In order to do this we will have to improve our data query scripts, and create better normalization methods.

Future work on this project may also need for human effort to classify the data correctly in the training files. Doing so will increase the accuracy of the classifier. In the scope of data collection, we also hope to find ways to query data from a limitation of more than 7-days to vastly improve our dataset. The Twitter API also only allows us to use data from the past 7 days, we can create a better training set by using data from all year, for example, when they play the dolphins, “dolphins” will have a big impact, whereas later in the year, “dolphins” will not show up.

Lastly, we would like to dive deeper into how we can create a graphical user interface (GUI) for users to test our tool, and benefit from our bayesian classifier.

**References**

[1] Bechon, P., Grimaldi, L., & Merouchi, Y. CS 229 Project: Improving on Yelp Reviews Using NLP

and Bayesian Scoring.

[2] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python.

O’Reilly Media Inc.

[3] Chapman, Wendy & Christensen, Lee & Wagner, Michael & Haug, Peter & Ivanov, Oleg & Dowling, John & Olszewski, Robert. (2005).

Classifying free-text triage chief complaints into syndromic categories with

natural language processing. Artificial intelligence in medicine. 33. 31-40. 10.1016/j.artmed.2004.04.001.

[4] Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., … Centers for Disease

Control and Prevention (CDC) (2014). Using online reviews by restaurant patrons to identify unreported cases of foodborne illness - New York City, 2012-2013. MMWR. Morbidity and mortality weekly report, 63(20), 441–445.

[5] Joshi, P. (2019). A NLP Approach to Mining Online Reviews using Topic Modeling. Retrieved

from https://www.analyticsvidhya.com/blog/2018/10/mining-online- reviews-topic-modeling-lda/.

[6] Martin, L., & Pu, P. (2014). Prediction of Helpful Reviews Using Emotions Extraction. In AAAI

Conference on Artificial Intelligence. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8562/8607

[7] Ranard BL., Werner RM., Antanavicius T., et al. Yelp reviews of hospital care can supplement

and inform traditional surveys of the patient experience of care. Health Aff (Millwood). 2016;35(4):697–705.

[8] Solangi Y. A., Solangi Z. A., Aarain S., Abro A., Mallah G. A., & Shah A.(2018). Review on

Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. In 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS) - Bangkok, Thailand. 1-4. 10.1109/ICETAS.2018.8629198

[9] Wang L., Feng M., Zhou B., Xiang B., & Mahadevan, S. (2015). Efficient Hyper-parameter

Optimization for {NLP} Applications. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing - Lisbon, Portugal. Association for Computational Linguistics. 2112-2117. 10.18653/v1/D15-1253

[10] Yu, B., Zhou, J., Zhang, Y., & Cao, Y. (2017). Identifying Restaurant Features via Sentiment

Analysis on Yelp Reviews. arXiv preprint arXiv:1709.08698.

[11] https://www.tweepy.org