Don't Take This Personally: Sentiment Analysis for Identification of "Subtweeting" on Twitter

A Senior Project submitted to
The Division of Science, Mathematics, and Computing
of
Bard College

by
Noah Segal-Gould

Annandale-on-Hudson, New York
May, 2018

Abstract

The purpose of this project is to identify subtweets. The Oxford English Dictionary defines "subtweet" as a "[Twitter post] that refers to a particular user without directly mentioning them, typically as a form of furtive mockery or criticism." This paper details a process for gathering a labeled ground truth dataset, training a classifier, and creating a Twitter bot which interacts with subtweets in real time. The Naive Bayes classifier trained in this project classifies tweets as subtweets and non-subtweets with an average F_1 score of 72%.

Contents

Dedication

I dedicate this senior project to @jack, who has willfully made numerous changes to Twitter which inevitably angered millions.

Acknowledgments

Thank you professors Sven Anderson, Keith O'Hara, and Rebecca Thomas for making this project possible through your combined efforts to teach and advise me. Thank you Benjamin Sernau '17 for enduring through three years of Computer Science courses with me and being a source of unending joy in my life. Thank you to Julia Berry '18, Aaron Krapf '18, and Zoe Terhune '18 for being my very best friends and giving me things worth caring about. Finally, thank you to my parents Tammy Segal and Emily Taylor for your constant support and patience throughout my four years at Bard College.

1
Introduction

The data gathered for this project contains unsavory language which may be considered inappropriate.

1.1 Background

This project utilizes the Twitter Application Programming Interface (API) as well as concepts from machine learning and text analysis. The ground truth dataset created for this project features 22,576 tweets from March and April of 2018. There has been significant research in the field of sentiment analysis to identify the opinions held within bodies of text, however this research does not include so-called "subtweets." The Oxford English Dictionary defines that a subtweet is a "[tweet] that refers to a particular user without directly mentioning them, typically as a form of furtive mockery or criticism." Treated as a characteristic of a tweet's sentiment, we will utilize the Naive Bayes classification algorithm to perform sentiment analysis on subtweets.

For acquisition of a ground truth dataset, we consider **true subtweets** to be tweets to which another Twitter user replied who specifically called it out as a subtweet. We consider **true non-subtweets** to be tweets to which another user replied who specifically did **not** call it out as a subtweet. Consider these examples:

	True Subtweet Data	True Non-Subtweet Data
Tweet	Talk to him again about “dropping me” and you’ll get your teeth knocked out	That’s been one of my biggest issues here; the onus is on ordinary people who, in their spare time, must campaign for the basic services of a city. This is not how progressive cities should be built
Reply	Thomas don’t subtweet me during work hours	i guess i am not as creative as i thought

We will keep these definitions and examples in mind when the classifier is used on original tweets independent of their replies. These will be called **predicted subtweets** and **predicted non-subtweets**. The following sections detail the resources and techniques utilized to acquire the labeled ground truth dataset, train the Naive Bayes classifier, and program a Twitter bot to interact with subtweets in real time.

1.2 Changes in Data Acquisition

The novel approach developed for creating a ground truth dataset relied on a particular phenomenon in which Twitter users were already calling-out the subtweets of their peers. The following pattern was observed: a user would post a subtweet which was easily recognized by a peer, and that peer would then reply to that tweet in order to complain that the original user was subtweeting or to ask if the tweet was indeed a subtweet. Initially, the program used the Twitter API’s search functionality to specifically search for replies to tweets which contained some form of the string “subtweet.” It utilized the API’s status object to access the tweet to which it was replying. For two months, each day’s alleged subtweets and their associated accusatory replies were saved.

Initially, the classifier was trained using a dataset which was half composed of these alleged subtweets and half composed of tweets randomly selected from a pre-labeled sentiment analyzed tweets dataset (Go et al., 2009). This procedure failed to make the training data representative of **true subtweets** and **true non-subtweets**. The alleged subtweets downloading program was revised and it was set to download tweets with replies which specifically did **not** contain the string “subtweet.” In both the program which downloaded subtweets and the program which downloaded non-subtweets, the assumptions about these interactions would not hold true in every case. They were intended as generalizations which would make acquiring a ground truth dataset for use in performing binary classification significantly easier and less time-consuming than finding and labeling subtweets and non-subtweets by hand. Indeed, dataset utilized by Go et al. used a similar method for acquiring labeled data. In their *Sentiment140* dataset, the labels were acquired according to emoticons present within the tweets instead of through hand-labeling by actual humans.

1.3 The Twitter API

Twitter provides a free Application Programming Interface (API) to registered users and has done so since September of 2006 (Stone, 2006). The API allows developers to programmatically access and influence tweets individually or through real time search filters, and also read and write direct messages (Twitter, 2018). The creation of a Twitter application which utilizes the API requires creation and email verification of an account, and developers are also required to agree to the terms of service (Twitter, 2016). Creation of an application provides developers with authentication tokens which can then be used to access the API.

To make creation of Twitter applications easier, Tweepy (Roesslein, 2009) is an open source library for the Python programming language which provides methods and classes used to interact with the API and its status objects (Twitter, 2018). A Twitter status object is a dictionary of key and value pairs which contains text, media, and user information associated with particular tweets (i.e. statuses). There are rate limits for both reading and writing to the API which must be kept in mind when programming for it.

1.4 Feature Extraction & Selection

Feature extraction and selection are the processes by which text is modified prior to any kind of processing such that the important features within it are easily accessible. Changes are made to preserve the characteristics of the text which are relevant to the goal of the analysis and to leave out the ones which are irrelevant. Because we will be using Naive Bayes, we must keep in mind which features in each tweet (e.g. URLs and user names) ought to influence the probabilities that an entire feature-set (i.e. that whole tweet) suits a particular class.

1.4.1 Regular Expressions

For text classification through machine learning, it is popular to modify the ground truth dataset to make features which are not important to the classification problem as **generic** as possible. For classification of subtweets, the classifier will treat URLs, mentions of usernames, and English first names generically. In other words, it will keep track of the existence of those features but specifically will not encounter the text contained within them. In identification of subtweets, there exists no syntactic or linguistic significance in the format of a URL or the name a user chooses to associate with themselves or another. However, the existence of those features within the tweet remains important. For this kind of substring searching, pattern matching through regular expressions was used to replace every occurence of URLs, usernames, and first names with special tokens which were not already in the dataset. The top 100 most common English names for both men and women over the last century were acquired from the United States Department of Social Security.

1.4.2 Tokenization

Instead of training the classifier on entire strings, **tokenization** is necessary in order to extract individual features from the text. The Natural Language Toolkit (Bird and Loper, 2004) provides a tweet tokenizer to achieve this. For some string, the tokenizer splits apart words, usernames, URLs, hashtags, and punctuating characters as individual tokens. NLTK's tweet tokenizer also appropriately distinguishes between punctuating characters and emoticons composed of punctuating characters.

1.4.3 N-Grams

An **n-gram** is a contiguous sequence of n tokens in a piece of text. For example, given a string such as “This is a test,” the bigrams ($n = 2$) for this string are “This is,” “is a,” and “a test.”

Instead of training the classifier using unigrams ($n = 1$) exclusively, we train it using unigrams, bigrams, and trigrams ($n = 3$). Thus, when the probability that some specific token within a tweet belongs to a specific class is calculated, its neighbors are also considered in combination with it. N-grams enable the classifier to treat particular groupings of tokens with some size n as importantly as it treats the individual tokens, thus identifying particular word groupings most associated with the classes.

1.4.4 Stop Words

A list of stop words typically contains the most common words in a language. For English text, the list is often composed of words such as "the," "it," and "of." Tokens matching stop words are ignored during classifier training because they are too common to help the classifier distinguish subtweets from non-subtweets.

1.5 TF & TF-IDF

The probabilities calculated for Naive Bayes are not best found using raw token counts within individual documents. Instead, TF-IDF is popularly used in order to vectorize the tokens for use in training the classifier.

Term frequency (TF) is a simple method for vectorizing text in which all terms (i.e. tokens, features, words, etc.) in the corpus are featured in a vector for each document, and the frequency of each term is reflected in the number representing the corresponding term. The bag of words model essentially vectorizes features using this method. Unfortunately, TF falls short when the corpus of documents contains terms which appear frequently but do not necessarily help inform the classifier on terms that are best associated with a particular class. TF-IDF, or term frequency-inverse document frequency, is the product of the TF for a specific term and the inverse document frequency (IDF) for that same term. The TF is equal to the ratio between the number of occurrences of a term in a document, and the total number of words in that document. IDF, then, is the logarithm of the ratio between the number of documents in the corpus, and the number of documents which contain that term. Taking the logarithm means the value will be higher for rarer terms. Thus, the product of TF and IDF assigns weights which appropriately value terms which are frequent within a document but rare in the entire corpus of documents. Because the weighted feature vectors calculated using TF-IDF follow a multinomial distribution, our classification algorithm is specifically a Multinomial Naive Bayes classifier.

1.6 Naive Bayes

Naive Bayes stands out as particularly simple and common for use in text classification. A **bag of words** model typically ignores word positions in favor of keeping track of raw token frequencies, which are weighted to produce TF-IDF feature vectors. Then, Bayes Theorem is utilized to predict the probability that a given feature set (e.g. words, sentences, etc.) belongs to a particular label (i.e. a category or class). Bayes theorem states the following:

$$\text{posterior probability} = \frac{\text{conditional probability} \cdot \text{prior probability}}{\text{evidence}}$$

In all cases where we must calculate the probability for an entire feature-set, we simply take the product of all the extracted feature probabilities in the document. The **posterior probability** is the probability that a particular document belongs to a class given the observed features within that document. The Naive Bayes algorithm maximizes this probability in order to predict which class best fits a document. The **conditional probability** refers to the likelihood of encountering the features within a document given that those features belong to a particular class. The **prior probability** is otherwise known as the **class probability** and is equal to the general probability of encountering a particular class. In our case, we have chosen to keep our classes balanced (i.e. there are equal numbers of documents in both), so the prior probability will always be 50%. The evidence, then, refers to the probability of encountering the features within a document **independent** of the class label. It is often ignored in the final classification step on the basis that it acts merely as a scaling factor when trying to maximize which class produces the greater posterior probability. Consider this classification example:

$$\text{if } P(\omega = \text{subtweet} \mid \mathbf{x}) \geq P(\omega = \text{non-subtweet} \mid \mathbf{x}) \text{ classify as subtweet,}$$

$$\text{else classify as non-subtweet.}$$

Dropping the evidence term on the basis that it is constant for both classes, we expand that:

$$P(\omega = \text{subtweet} \mid \mathbf{x}) = P(\mathbf{x} \mid \omega = \text{subtweet}) \cdot P(\text{subtweet})$$

$$P(\omega = \text{non-subtweet} \mid \mathbf{x}) = P(\mathbf{x} \mid \omega = \text{non-subtweet}) \cdot P(\text{non-subtweet})$$

Assuming the posterior probability for the former is greater than or equal to the latter, the classifier predicts that the document fits that class. The naive assumption maintains that all features are treated as conditionally independent (i.e. that the presence or omission of a particular feature does not change the likelihood of encountering other features), and although this is frequently violated, Naive Bayes often performs well anyway (Zhang, 2004).

For cases in which the classifier encounters a feature absent from the features which were used to train it, a so-called **zero probability** appears. Because the probability of encountering the feature is 0, **additive smoothing** is often utilized to appropriately weight new features using an extra term α, so the probability that an entire feature-set fits into a specific class is not 0. In this project, we use **Laplace smoothing** ($\alpha = 1$). This technique smooths categorical data by including the pseudo-count α into each probability estimate. Thus, the probability of a feature given a particular class becomes:

$$P(x_i \mid \omega_j) = \frac{N_{x_i,\omega_j} + \alpha}{N_{\omega_j} + \alpha\, d} \quad (i = (1, ..., d))$$

where N_{x_i,ω_j} is the number of times feature x_i appears in samples from class ω_j, N_{ω_j} is the total count of all features in class ω_j, α is the parameter for additive smoothing, and d is the dimensionality of the feature vector $\mathbf{x} = [x_1, ..., x_d]$. Compared to datasets which contain millions of tweets such as *Sentiment140* (Go et al., 2009), our classifier has access to significantly fewer documents. We utilize Laplace smoothing because when the classifier is tested on new tweets it will likely encounter never before seen features given the limited size of the ground truth dataset.

1.7 K-Folds Cross-Validation

Instead of using the entire ground truth dataset as training data for the Naive Bayes classifier, we can split it apart into a **training set** and a **test set**. The training set is fed to the classifier, and the test set is used to observe statistics about its performance. In cross-validation using k-folds, we make random splits in the dataset k times to create several training set and test set sections. We then take statistical measurements of how well the classifier performs on the testing set for each fold. If we made one single split into training and testing sections of our ground truth dataset and only used that single testing set to gather statistics on the classifier's performance, we would not be able to confirm that those statistics were representative of all the data in the entire ground truth dataset. Instead, we perform 10-fold cross validation, choosing 90% of the data to be the training set, and 10% to be the testing set in each fold. Precision, F_1 score, and recall are calculated within each iteration of the 10 folds, thus utilizing 10 different test datasets. Finally, we use the averages of those statistics across all folds to measure the overall performance of the classifier. The following figure illustrates this process:

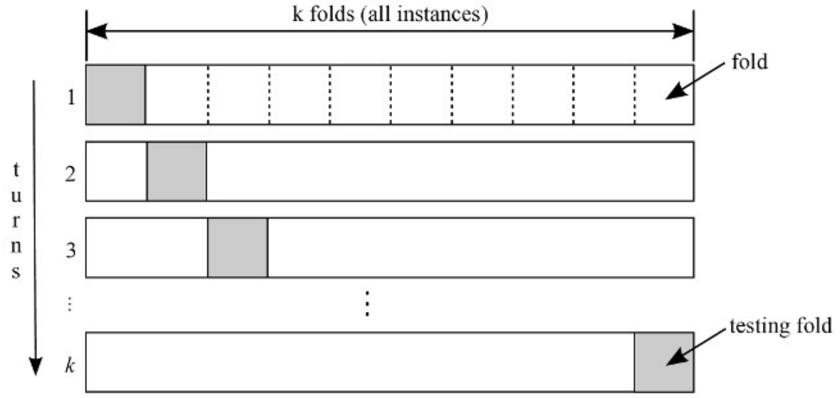

Figure 1.7.1. K-Folds Cross-Validation Example (Borovicka et al., 2012)

1.8 Statistical Considerations

In the binary classification of subtweets and non-subtweets, we consider true positives (TP) to be true subtweets which were correctly labeled as predicted subtweets, false positives (FP) to be true non-subtweets which were incorrectly labeled as predicted subtweets, true negatives (TN) to be true non-subtweets which were correctly labeled as predicted non-subtweets, and false negatives (FN) to be true subtweets which were incorrectly labeled as predicted non-subtweets. As such, there are two ways for the classifier to be wrong: it can produce false negatives and false positives.

1.8.1 Precision

Precision refers to the ratio between the true positives, and the true positives and false positives. It is also known as the positive predictive value.

$$P = \frac{TP}{TP + FP}$$

1.8.2 Recall

Recall, then, refers to the ratio between the number of true positives, and the true positives and false negatives. It is also known as the sensitivity.

$$R = \frac{TP}{TP + FN}$$

1.8.3 Accuracy

The accuracy is the ratio between the true positives and the true negatives, and the true positives, true negatives, false positives, and false negatives. Accuracy alone is a particularly bad quantifier of how well a classifier performs when working with data which is class-imbalanced (i.e. there are not equal numbers of items in each class). In our ground truth dataset, the classes are balanced so measuring accuracy will still be informative.

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

1.8.4 F1 Score

The F_1 score is a weighted average of the precision and recall. Thus, it takes both false positives and false negatives into account.

$$F1 = \frac{2 * (P * R)}{P + R}$$

1.8.5 Null Accuracy

The null accuracy is just the accuracy which is obtained by always predicting the most frequent class. Because there are two classes and the tweets within the ground truth dataset equally compose both, the null accuracy will always be 50%.

1.9 Review of Literature

“Subtweet” was coined in December of 2009 by Twitter user Chelsea Rae (Rae, 2009) and was entered into Urban Dictionary the following August (Urban Dictionary, 2010). In “To tweet or subtweet?: Impacts of social networking post directness and valence on interpersonal impressions” (Edwards and Harris, 2016), Edwards and Harris sought to analyze student participants’ perceptions of known subtweeters. In the news, too, subtweets have garnered attention in *The Atlantic* (Madrigal, 2014), *The Washington Post* (Dewey, 2016), and *Slate* (Hassler, 2016). In news media, subtweets garner attention for their prevalence among government officials as well. Following President Donald Trump’s inauguration, The Washington Post compiled its “A running list of all the possible subtweets of President Trump from government Twitter accounts,” (Ohlheiser, 2017) cementing subtweets as particularly newsworthy.

There were over 140 million active Twitter users who sent 340 million text-based tweets to the platform every day by March of 2012 (Twitter, 2012). Since Twitter-founder Jack Dorsey sent the first Tweet in March of 2006 (Dorsey, 2006) social scientists, political scientists, and computer scientists have applied machine learning techniques to understand the patterns and structures of the conversations held on the platform. Through sentiment analysis, we are able to use machine learning to identify patterns within natural language which indicate particular feelings both broadly (e.g. positive, neutral, or negative) and toward topics (e.g. politics, terrorism, brands, etc.).

On Twitter, the most common way to publicly communicate with another user is to compose a tweet and place an "@" before the username of that user somewhere in the tweet (e.g. "How are you doing, @NoahSegalGould?"). Through this method, public discussions on Twitter maintain a kind of accountability: even if one were to miss the notification that they were mentioned in a tweet, one's own dashboard keeps a running list of their most recent mentions.

If an individual sought to disparage or mock another, they could certainly do so directly. But the targeted user would probably notice, and through the search functions of the platform, anyone could see who has mentioned either their own or another's username. Instead, a phenomenon persists in which users of the platform deliberately insult others in a vague manner by making complaints while omitting the targets of those complaints.

Although the OED's definition states that a subtweet "...refers to a particular user without directly mentioning them, typically as a form of furtive mockery or criticism," it is perhaps too restrictive. Some individuals believe subtweets abide by this definition, but others expand it to allow inclusion of others' real names (especially if that individual does not own a Twitter account), and some do not even require that a particular user be the target of the tweet. Because subtweeting is colloquial in nature, we will expand the definition of subtweet to permit these less restrictive features.

Sentiment analysis on social networking services such as Twitter has garnered attention within seemingly distinct fields of interest. In "Text mining for market prediction: A systematic review," Nassirtoussi et al. surveyed varied methods for text-mining social media for sentiment analysis of financial markets and approached that problem with both behavioral and economic considerations in mind (Nassirtoussi et al., 2014). Following a terrorist event in Woolwich, London in 2013, Burnap et al. analyzed the immediate Twitter response following the attack to inform statistics on how long it takes for responses from official sources to disseminate during crises (Burnap et al., 2014). Prior research of these kinds utilizes sentiment analysis techniques on tweets, but no known research exists which specifically performs any sentiment analysis on subtweets.

Long before Twitter, psychologist Gordon Allport wrote about "antilocution" in *The Nature of Prejudice* (Allport, 1954). For Allport, antilocution was the first of several degrees of apathy which measure prejudice in a society. It represented the kind of remarks which target a person, group, or community in a public or private setting but do not address the targeted individual directly. Different from both hate speech and subtweeting, antilocution necessitates that an in-group ostracize an unaware out-group through its biases.

The most germane research available focuses on sentiment analysis of figurative language. Determining sentiment based on features of text which are distinctly separate from their literal interpretations presents difficulties for human readers as well as computer programs. In *SemEval*, the International Workshop on Semantic Evaluation, analysis of figurative language on Twitter has been a core task for their competition since 2015 (Ghosh et al., 2015) and returns this year with a specific focus on ironic tweets (Van Hee et al., 2018). In this year's description for "Task 3: Irony detection in English tweets," Van Hee et al. touch upon online harassment as a potential point of significance for sentiment analysis of ironic tweets.

We pursue sentiment analysis of subtweets in order to challenge the hypocrisy of utilizing a service which presents itself as a public forum to speak in distinctly private ways. Toward this end, these are our goals: this project will provide a framework for collecting examples of subtweets, train a classification algorithm using those examples, and finally utilize that classifier in real time to make tweets which were intended to be unseen by specific parties easily accessible to all parties. In presenting covertly hurtful content as obviously hurtful in a public fashion, perhaps it will promote a particular awareness that tweets posted by public accounts are indeed publicly accessible, and that Twitter's Terms of Service (Twitter, 2016) allows for this kind of monitoring.

Using a machine-learning approach to perform sentiment analysis, syntactic and linguistic features are typically utilized in probabilistic (e.g. Naive Bayes and Maximum Entropy) and linear (e.g. Support Vector Machines and Neural Networks) classification algorithms. The probabilistic approach is sometimes called *generative* because such models generate the probabilities of sampling particular terms (Medhat et al., 2014). Linear classification utilizes the vectorized feature space of words, sentences, or documents to find a separating hyperplane between multiple classes. In this project, we approach the problem of identifying subtweets using the probabilistic Naive Bayes classification algorithm.

2
Implementation

Given access to the Twitter API, the programs developed for this project download a ground truth dataset of **true subtweets** and **true non-subtweets**, clean the data and select its features, train and test a Naive Bayes classifier using that dataset, and interact with **predicted subtweets** in real time. The following sections are sequentially ordered and contain brief explanations of the process at each step.

2.1 Searching for Tweets Using the Twitter API

The first step in this process is to acquire data for the ground truth dataset. After the API credentials are loaded, the following code is used to download tweets with replies which **do** and **do not** call them out as subtweets:

```
def get_subtweets(max_tweets=2500,
                  query=("subtweet AND @ since:2018-04-01"
                         "exclude:retweets filter:replies")):
    subtweets_ids_list = []
    subtweets_list = []
    for potential_subtweet_reply in tweepy.Cursor(api.search, lang="en",
                                                  tweet_mode="extended",
                                                  q=query).items(max_tweets):
        potential_subtweet_original = first_tweet(potential_subtweet_reply)
        if (not potential_subtweet_original.in_reply_to_status_id_str
            and potential_subtweet_original.user.lang == "en"):
            if (potential_subtweet_original.id_str in subtweets_ids_list
                or "subtweet" in potential_subtweet_original.full_text
                or "Subtweet" in potential_subtweet_original.full_text
                or "SUBTWEET" in potential_subtweet_original.full_text):
                continue
            else:
                subtweets_ids_list.append(potential_subtweet_original.id_str)
                subtweets_list.append({"tweet_data": potential_subtweet_original._json,
                                       "reply": potential_subtweet_reply._json})
                with open("../data/other_data/subtweets.json", "w") as outfile:
                    json.dump(subtweets_list, outfile, indent=4)
    return subtweets_list
```

Tweepy (Roesslein, 2009) provides the `Cursor` object which is used to iterate through different sections of the Twitter API. In this function, we use it with Twitter's search API to find tweets matching our query. This particular version finds **true subtweets** with their accusatory replies, but it requires only modification of its query argument to download **true non-subtweets**. Within the `for` loop, we confirm that the following qualities hold true:

- The alleged subtweet or non-subtweet is not in reply to any other tweet.
- The user who posted it used English as their primary language.
- We do not download duplicate tweets.
- It does not contain the string "subtweet" (with variations in capitalization).

Tweet status objects downloaded through the Twitter API contain pairs of keys and values just like Python dictionaries and JSON files. For acquisition of both subtweets and non-subtweets, this function can be modified to accept different queries. Respectively, these are the queries which were utilized in acquiring so-called subtweet-accusations (i.e. these replies typically claim that the tweet to which they reply was a subtweet) and non-subtweet-accusations (i.e. these replies are essentially normal replies to normal tweets):

```
"subtweet AND @ since:2018-04-01 exclude:retweets filter:replies"

"-subtweet AND @ since:2018-04-01 exclude:retweets filter:replies"
```

The only difference between the two is that the former searches for tweets which contain the string "subtweet" and the latter searches for tweets which exclude it. Both access the API as far back as it will allow (typically one week for free API users) and exclude retweets (i.e. unoriginal tweets which are reposted) while specifically searching for tweets which were in reply to other tweets. Unfortunately, tweet status objects representing replies to tweets do not contain the object data of the tweet to which they reply. In order to obtain this object, we utilize the following function:

```python
def first_tweet(tweet_status_object):
    try:
        return first_tweet(api.get_status(tweet_status_object.in_reply_to_status_id_str,
                                          tweet_mode="extended"))
    except tweepy.TweepError:
        return tweet_status_object
```

This code essentially goes up a chain of recursively finding the original tweet to which a reply was made until the API can no longer find another tweet object with an `in_reply_to_status_id_str` attribute, indicating that the tweet is an original **true subtweet** or **true non-subtweet**. The two versions of this program for acquiring **true subtweets** and **true non-subtweets** ran for three weeks between March and April of 2018, acquiring over 20,000 tweets which compose our ground truth dataset.

2.2 Cleaning and Preparing the Data

Although **true subtweets** and **true non-subtweets** are identified using characteristics of the Twitter API's tweet status objects, we utilize Naive Bayes exclusively for text classification on the unicode text contained within each alleged subtweet and non-subtweet status object. Thus, we ignore the replies to these tweets. Consider the following:

```python
def load_data(filename, threshold=0.1):
    data = [(urls_pattern.sub("GENERIC_URL",
             at_mentions_pattern.sub("GENERIC_MENTION",
             names_pattern.sub("GENERIC_NAME",
             t["tweet_data"]["full_text"])))
             .replace("\u2018", "’")
             .replace("\u2019", "’")
             .replace("\u201c", "\"")
             .replace("\u201d", "\"")
             .replace("&quot;", "\"")
             .replace("&amp;", "&")
             .replace("&gt;", ">")
             .replace("&lt;", "<"))
            for t in json.load(open(filename))
            if t["tweet_data"]["lang"] == "en"
            and t["reply"]["lang"] == "en"
            and t["tweet_data"]["user"]["lang"] == "en"
            and t["reply"]["user"]["lang"] == "en"]
    new_data = []
    for tweet in data:
        tokens = tokenizer.tokenize(tweet)
        english_tokens = [english_dict.check(token) for token in tokens]
        percent_english_words = sum(english_tokens)/len(english_tokens)
        if percent_english_words >= threshold:
            new_data.append(tweet)
    return new_data
```

The `load_data` function genericizes URLs, mentions of user names, and mentions of English first names. Using regular expressions for pattern matching, these substrings are replaced with special identifiers. The tweets are also cleaned to make HTML characters and unicode characters more consistent. The list comprehension intentionally excludes **non-English** tweets and those which were **not** posted by accounts which list their primary language as English. NLTK's tweet tokenizer is utilized at the end to check for tweets which contain at least 10% English tokens. This language detection is performed on each token in the tweet using the pyEnchant library (Kelly, 2016) which primarily serves as a spell-checker. The resulting dataset of either **true subtweets** or **true non-subtweets** is returned as a list.

After text cleaning, we remove tweets which are present in both the dataset of **true subtweets** and **true non-subtweets**. Duplicates may have appeared because one user thought a tweet was a subtweet but another did not. Consider the following Python code:

```
subtweets_data = load_data("../data/other_data/subtweets.json")
non_subtweets_data = load_data("../data/other_data/non_subtweets.json")
subtweets_data = [tweet for tweet in subtweets_data
                  if tweet not in non_subtweets_data]
non_subtweets_data = [tweet for tweet in non_subtweets_data
                      if tweet not in subtweets_data]
```

With the duplicates gone, we limit the size of the larger dataset to be the same as the smaller of the two. Thus, both the **true subtweets** and **true non-subtweets** compose the entire final ground truth dataset equally:

```
smallest_length = len(min([subtweets_data, non_subtweets_data], key=len))

subtweets_data = sample(subtweets_data, smallest_length)
non_subtweets_data = sample(non_subtweets_data, smallest_length)

subtweets_data = [(tweet, "subtweet") for tweet in subtweets_data]
non_subtweets_data = [(tweet, "non-subtweet") for tweet in non_subtweets_data]

training_data = subtweets_data + non_subtweets_data
```

Python's `random` library provides the `sample` function which randomly returns a list of n items from a list. After both datasets are made the same length, the lists of strings are made into lists of tuples. For each tuple, the **true subtweet** or **true non-subtweet** is associated with its label (i.e. subtweet and non-subtweet). Finally, these two lists of tuples are put together for use in training the classifier.

2.3 Training & Testing Naive Bayes

We utilize Scikit Learn's API for machine learning (Pedregosa et al., 2011) to create a pipeline. In Scikit, pipelines make managing machine learning algorithms easy by consolidating their parts into one object with configurable attributes. Our pipeline contains a vectorizer and a classifier:

```
sentiment_pipeline = Pipeline([
    ("vectorizer", TfidfVectorizer(tokenizer=tokenizer.tokenize,
                                   ngram_range=(1, 3),
                                   stop_words="english")),
    ("classifier", MultinomialNB())
])
```

We change the default arguments for Scikit's TF-IDF vectorizer to use NLTK's tweet tokenizer, and specify that we want to calculate our TF-IDF vectors using unigrams, bigrams, and trigrams. Then, we set the vectorizer to use Scikit's default English language stop words. The Multinomial

Naive Bayes Classifier we implement using Scikit includes **Laplace smoothing** ($\alpha = 1$) by default.

Scikit also has a convenient `KFold` object which we utilize to perform cross-validation on our classifier. The goal of K-Folds Cross-Validation is to train the classifier and acquire statistics on its performance while treating k different parts of the ground truth dataset as test sets. Within a single iteration for 10 iterations, the test set will always be 10% of the entire ground truth dataset, however the next iteration will use a different section. We ultimately average those statistics to understand the overall performance of the classifier. Consider the following Python code:

```python
def confusion_matrices(training_data, num_folds=10):
    text_training_data = np.array([row[0] for row in training_data])
    class_training_data = np.array([row[1] for row in training_data])
    kf = KFold(n_splits=num_folds, random_state=42, shuffle=True)
    cnf_matrix_test = np.zeros((2, 2), dtype=int)
    for train_index, test_index in kf.split(text_training_data):
        text_train, text_test = (text_training_data[train_index],
                                 text_training_data[test_index])
        class_train, class_test = (class_training_data[train_index],
                                   class_training_data[test_index])

        sentiment_pipeline.fit(text_train, class_train)
        predictions_test = sentiment_pipeline.predict(text_test)
        cnf_matrix_test += confusion_matrix(class_test, predictions_test)
```

In each iteration of the 10 folds, the program splits apart a training set and a test set. The classifier is trained on the training set using `sentiment_pipeline.fit`, and the classifier's predictions for the test set in that fold are used to add toward a confusion matrix which will visualize the performance of the classifier. We also calculate the precision, recall, and F_1 score for each fold's individual test set.

2.4 Creating the Twitter Bot

The Tweepy library we use to interact with the Twitter API provides a convenient object for streaming Twitter data in real time. The `StreamListener` class can track tweets by searching for specific users, locations, and keywords. In the following code, we extend it to track subtweets:

```
class StreamListener(tweepy.StreamListener):
    def on_status(self, status):
        id_str = status.id_str
        screen_name = status.user.screen_name
        created_at = status.created_at
        retweeted = status.retweeted
        in_reply_to = status.in_reply_to_status_id_str
        text = status.full_text

        # Genericize extra features and clean up the text
        text = (urls_pattern.sub("GENERIC_URL",
                at_mentions_pattern.sub("GENERIC_MENTION",
                names_pattern.sub("GENERIC_NAME",
                text)))
                .replace("\u2018", "'")
                .replace("\u2019", "'")
                .replace("\u201c", "\"")
                .replace("\u201d", "\"")
                .replace("&quot;", "\"")
                .replace("&amp;", "&")
                .replace("&gt;", ">")
                .replace("&lt;", "<"))

        tokens = tokenizer.tokenize(text)

        english_tokens = [english_dict.check(token) for token in tokens]
        percent_english_words = sum(english_tokens)/float(len(english_tokens))

        # Make sure the tweet contains some English
        is_english = False
        if percent_english_words >= 0.1:
            is_english = True

        # Calculate the probability using the pipeline
        subtweet_probability = sentiment_pipeline.predict_proba([text]).tolist()[0][1]
```

This part of the program is for live classification of subtweets and gathers information on tweet status objects as they are encountered. We extract data from the tweet object including the ID of that tweet and the text contained within it. We utilize the same techniques for cleaning and genericizing the tweet which we used in preparing our ground truth data for the classifier. The pipeline has a `predict_proba` method which takes as its input a list of strings and outputs an array of probabilities for each class. `subtweet_probability`, then, uses that method to predict the probability that the tweet fits the "subtweet" class according to the Naive Bayes classifier and the vectorizer we used in our pipeline. We also check that the potential subtweet meets specific requirements before the Twitter bot will interact with it:

```
if all([subtweet_probability >= THRESHOLD,
        "RT" != text[:2], is_english,
        not retweeted, not in_reply_to])
```

Included in this conditional statement is a comparison to determine if the probability that the tweet is a subtweet meets a specific threshold. We do not want to call out subtweets unless the probability is high enough. Following this check, we can interact with the tweet in several ways:

```
# Quote the tweet
api.update_status(("Is this a subtweet? {:.2%} \n" +
                   "https://twitter.com/{}/status/{}").format(subtweet_probability,
                                                              screen_name,
                                                              id_str))
# Like the tweet
api.create_favorite(id_str)

# Reply to the tweet
api.update_status("@{} Is this a subtweet? {:.2%}".format(screen_name,
                                                          subtweet_probability),
                  id_str)
```

To quote the potential subtweet means that the tweet being referenced is embedded within our own tweet with a caption above it. We instantiate our custom `StreamListener` class and use it to filter through tweets in real time:

```
stream_listener = StreamListener()
stream = tweepy.Stream(auth=api.auth, listener=stream_listener, tweet_mode="extended")

stream.filter(follow=user_ids, stall_warnings=True, languages=["en"], async=True)
print("Streaming has started.")
sleep(DURATION)
stream.disconnect()
```

The `user_ids` list contains strings of Twitter user IDs. Every time a user whose ID is in the list sends a tweet, the program will classify that tweet and use the predicted probability that it is a subtweet to call it out on Twitter from the account linked to the application's API credentials. We filter the stream asynchronously in order to use the `sleep` function from the `time` Python library to run the Twitter bot for a limited number of seconds.

3
Results

3.1 Ground Truth Dataset

We obtained 12,169 **true subtweets** and 21,411 **true non-subtweets** to be used in our ground truth dataset. After confirming they were English and removing duplicates, there were 11,288 **true subtweets** and 19,289 **true non-subtweets** remaining. In order to preserve a balance between the classes, we limited the size of **true non-subtweets** to be the same as the set of **true subtweets**. Thus, our final ground truth dataset contains 22,576 tweets. The programs for acquiring this data were run for three weeks between March and April of 2018. The following table features 10 random examples from each category:

True Subtweets	True Non-Subtweets
I know who I want to take me home	He's followed the putrid smell of GENERIC_MENTION which has led him to GENERIC_MENTION's whereabouts. Some odd neighbor boy was watching him as he approached the house. Liam didn't hesitate to drink the young one dry. "HONEY, I'M HOME!" He calls, kicking in the door.
One bad chapter doesn't mean your story is over	YG so low? I guess the only group that sold physicals well is Bigbang GENERIC_URL
What do you offer someone who doesn't like tea, coffee, or hot chocolate? YOU'RE A GUEST, YOU NEED A HOT DRINK, IT IS THE LAW.	is everything ok?? GENERIC_URL
on the flip side, i agree that "big" accounts have an obligation to do their due diligence to not sic their followers on others	The parade of falsehoods about CIA nominee Gina Haspel GENERIC_URL
im bout to drop like 20lbs and then become a rapper.	Very first day of fortnite got 2 second place games and a third place gg not bad for a trash player :p
It's hilarious that your only "godly" when it's convenient for you or when you want to put it on twitter.	Rosie the Corpse Flower bloomed at Tucson Botanical Gardens: See pix allery here GENERIC_URL #Tucson GENERIC_MENTION GENERIC_URL
It really bothers me when people try to take advantage or assume I do shoots for free GENERIC_URL	What is Uncle's GREATEST prediction?
I just don't get people lol	Didi Gregious is my new favorite player.
y'all kids today with your "waifu this" and "waifu that," back in my day our waifus had metal jaws and carried high-frequency blades	GENERIC_MENTION we need to get you drawn as Dekamaster Doggy Krueger
I follow this gorgeous dog account on insta and his owner has such an annoying voice. SHUT UP AND LET ME ENJOY YOUR MAGNIFICENT DOG! I'M NOT HERE FOR YOU!	"Huge Caravan" Of Central Americans Is Headed For The U.S. Border In Hopes Of Asylum GENERIC_URL

In this table, GENERIC_URL, GENERIC_MENTION, and GENERIC_NAME indicate that those replacements stand in place of actual URLs, mentions of user names, and English first names in the original tweets. With tens of thousands of tweets contained within the ground truth dataset, understanding the characteristics of these tweets by example alone is insufficient. To improve our overall understanding of these characteristics, we gathered statistics on the distributions of lengths, punctuating characters, stop words, and unique words in the dataset:

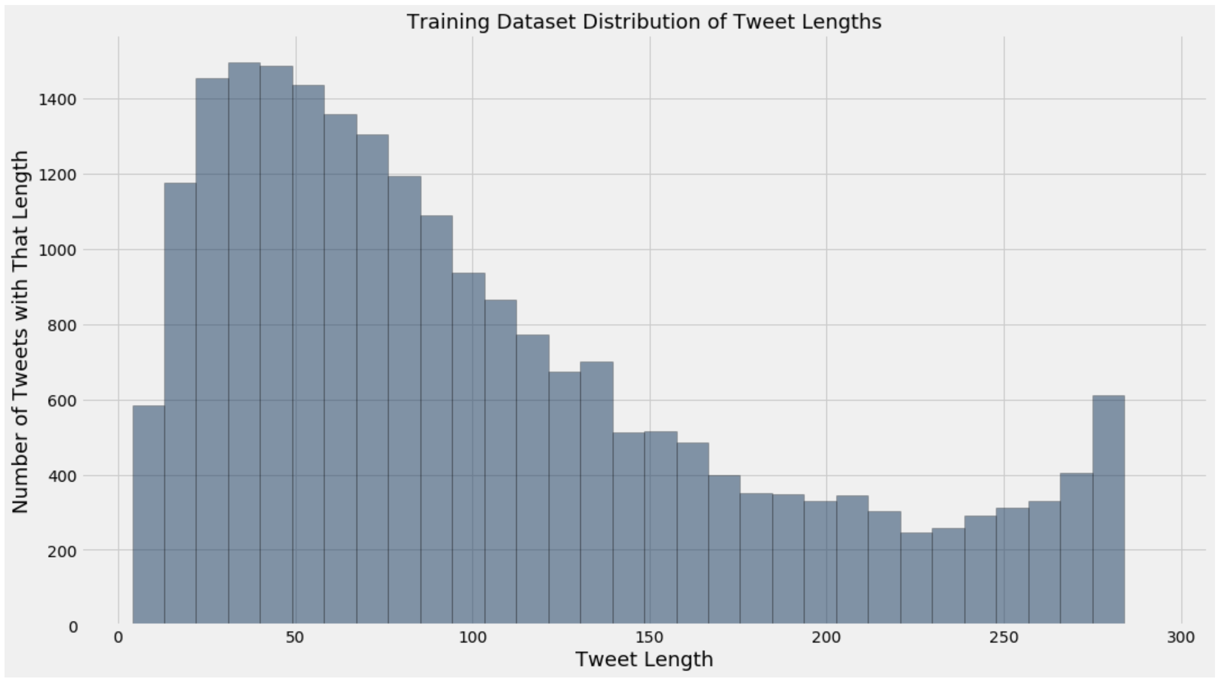

Figure 3.1.1. Histogram of Ground Truth Tweet Lengths ($\mu = 104.41, \sigma = 73.68$)

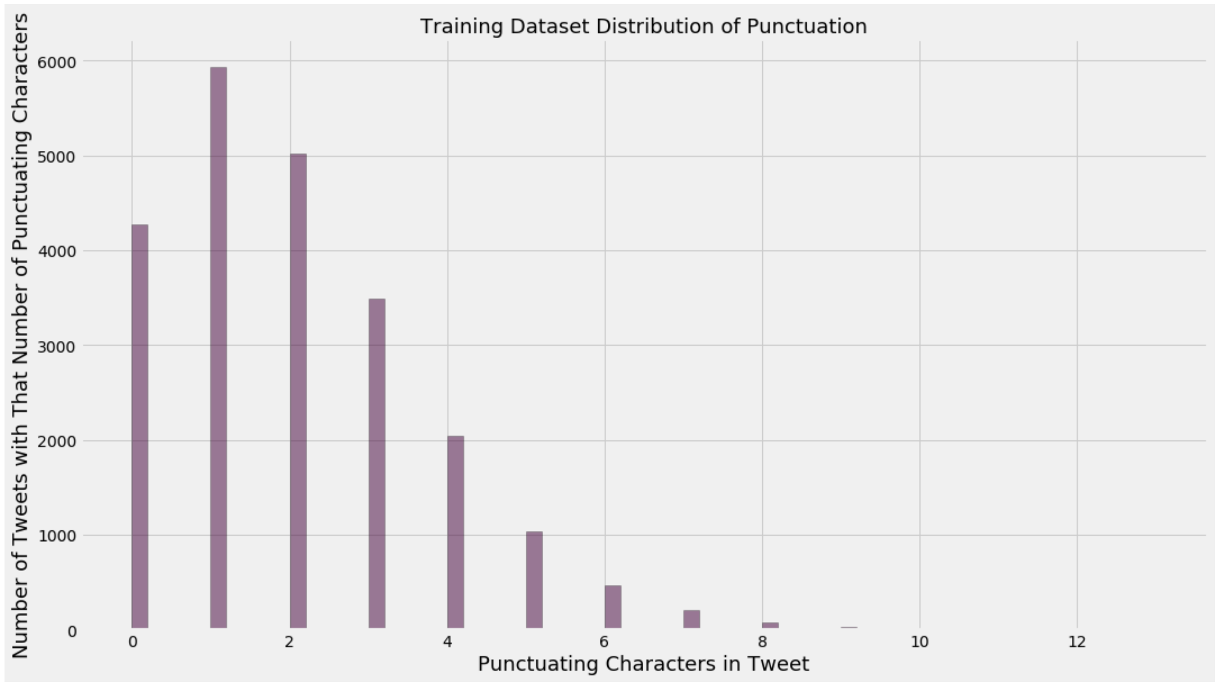

Figure 3.1.2. Histogram of Ground Truth Tweet Punctuation ($\mu = 1.99, \sigma = 1.64$)

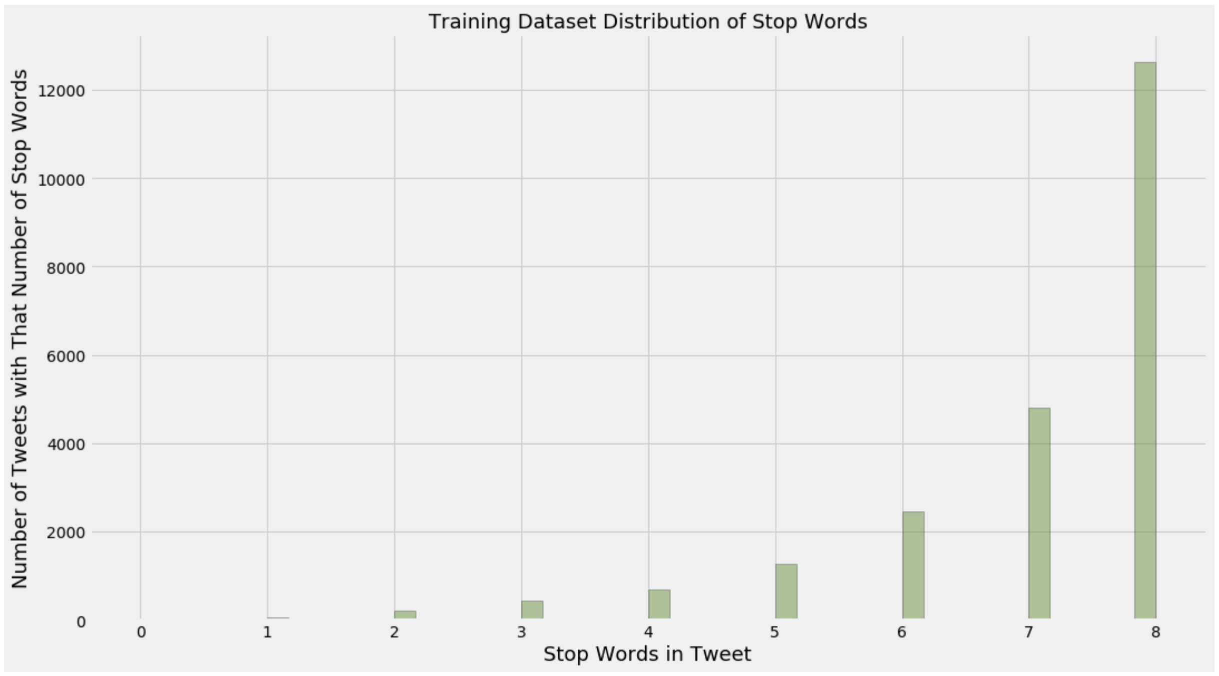

Figure 3.1.3. Histogram of Ground Truth Stop Words ($\mu = 7.10, \sigma = 1.35$)

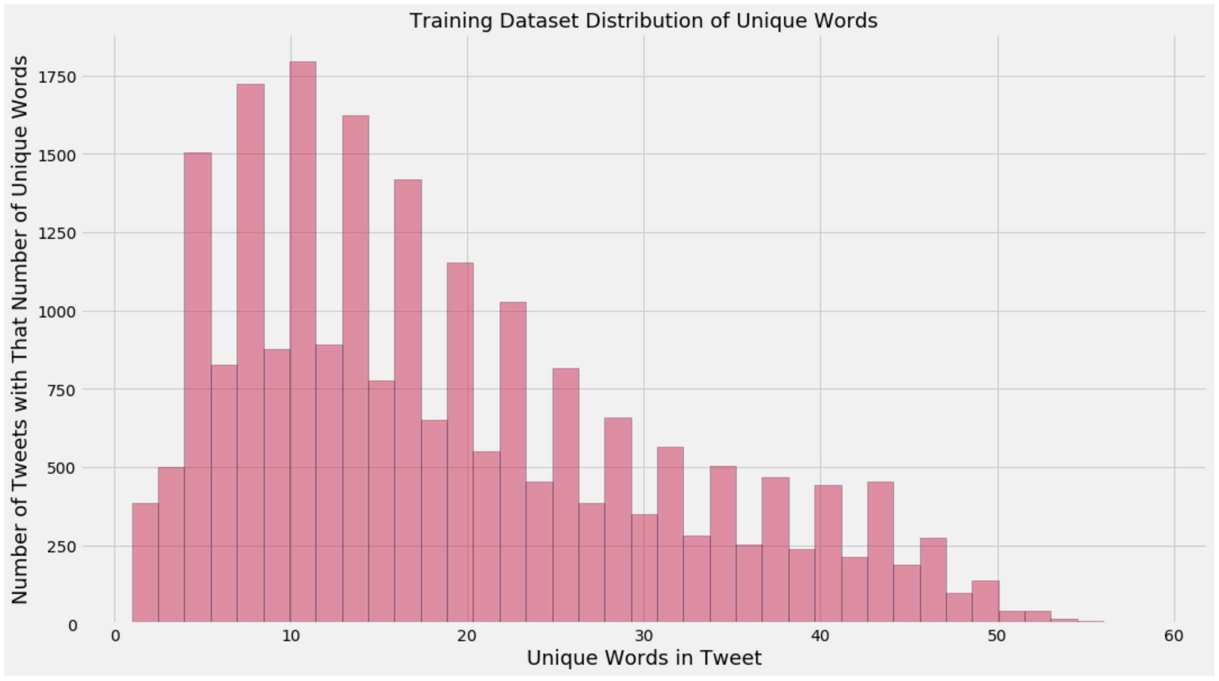

Figure 3.1.4. Histogram of Ground Truth Unique Words ($\mu = 18.89, \sigma = 12.07$)

3.2 Confusion Matrices

A confusion matrix is a table which visualizes the performance of a classifier. For our binary classification of subtweets and non-subtweets, it illustrates the accumulated true positives, true negatives, false positives, and false negatives across all 10 folds of the K-Folds cross-validation:

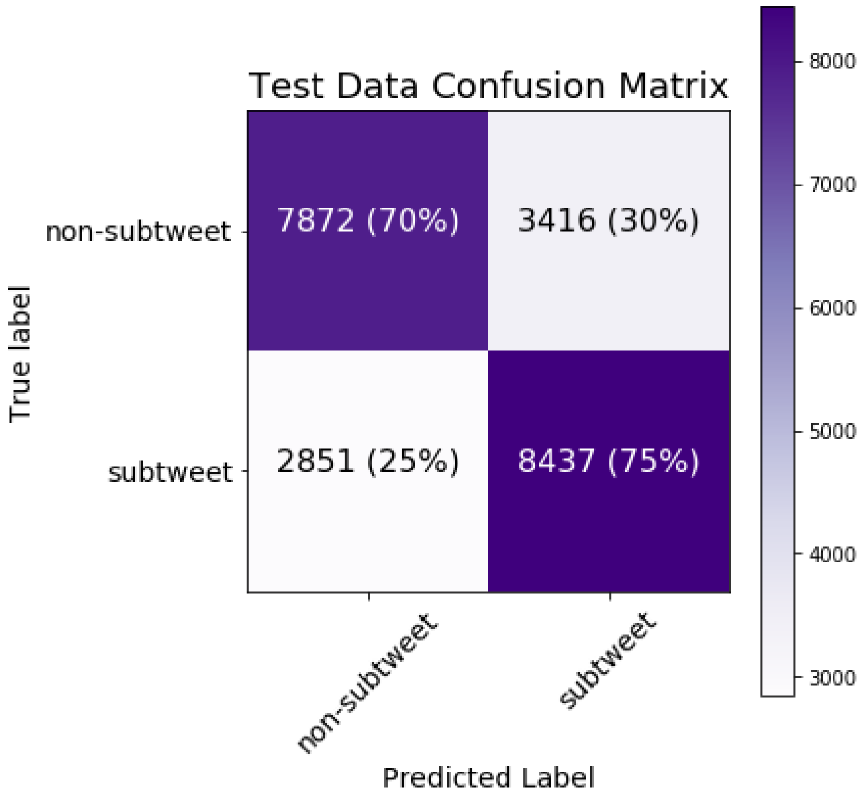

Figure 3.2.1. Confusion Matrix on Test Data from Each Fold on the Ground Truth Dataset

We read the confusion matrix by matching rows with columns to see how well the classifier performs on **true subtweets** and **true non-subtweets** in making classifications to produce **predicted subtweets** and **predicted non-subtweets**. These results indicate that the Naive Bayes classifier performs 5% better when classifying true positives than it does when classifying true negatives. In comparison, it is 5% **worse** at classifying false positives than it is at classifying false negatives.

3.3 Statistical Analyses

The K-folds cross-validation we utilized to measure the statistical performance of our classifier was averaged across all 10 folds. The following table depicts these statistics:

	Precision	Recall	F_1 Score
Non-Subtweet	0.7342	0.6963	0.7146
Subtweet	0.7113	0.7480	0.7290
Averages	0.7230	0.7220	0.7218

Thus, our method for classifying subtweets has an average F_1 score of 72%.

3.4 Application of the Classifier on Tweets from Known Subtweeters

In order to test how the classifier performed on tweets from known subtweeters, we acquired all publicly available tweets from 14 different accounts from which users had previously subtweeted. Of these accounts, the following histogram illustrates the distribution of subtweet probabilities for 3 random users:

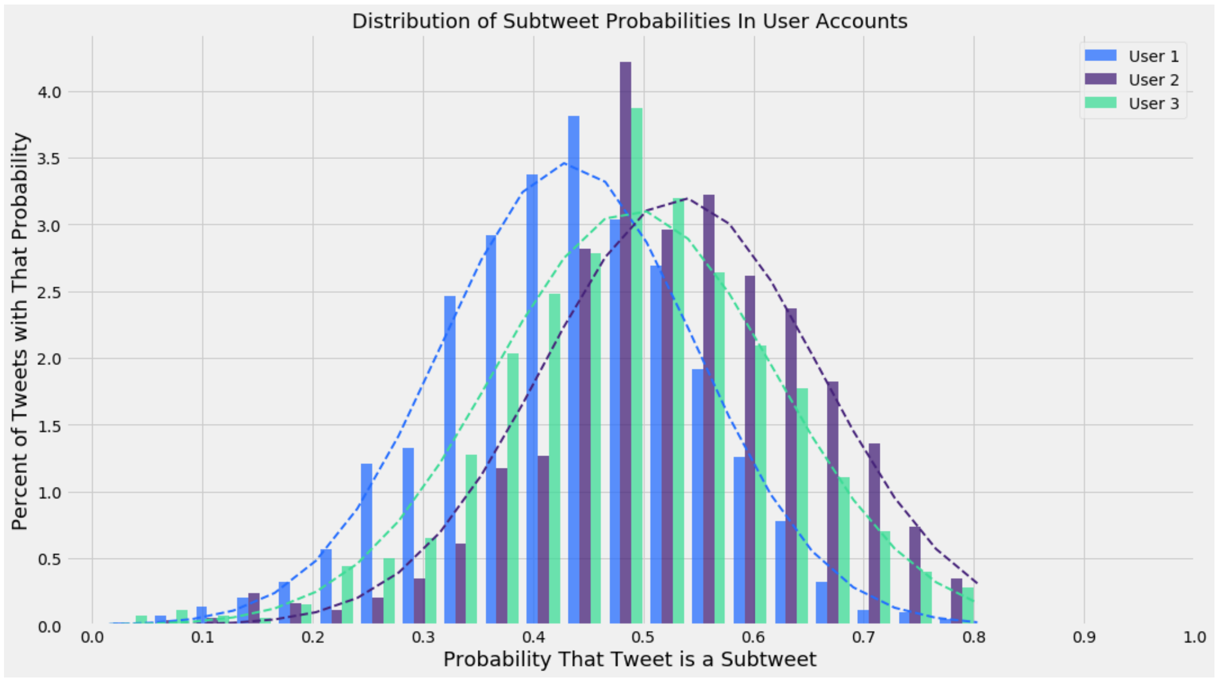

Figure 3.4.1. Distribution of Subtweet Probabilities for 3 Random Known Subtweeters $(\mu_1 = 0.432, \sigma_1 = 0.115), (\mu_2 = 0.534, \sigma_2 = 0.125), (\mu_3 = 0.492, \sigma_3 = 0.128)$

In comparison, consider the distribution of predictions across all 14 accounts:

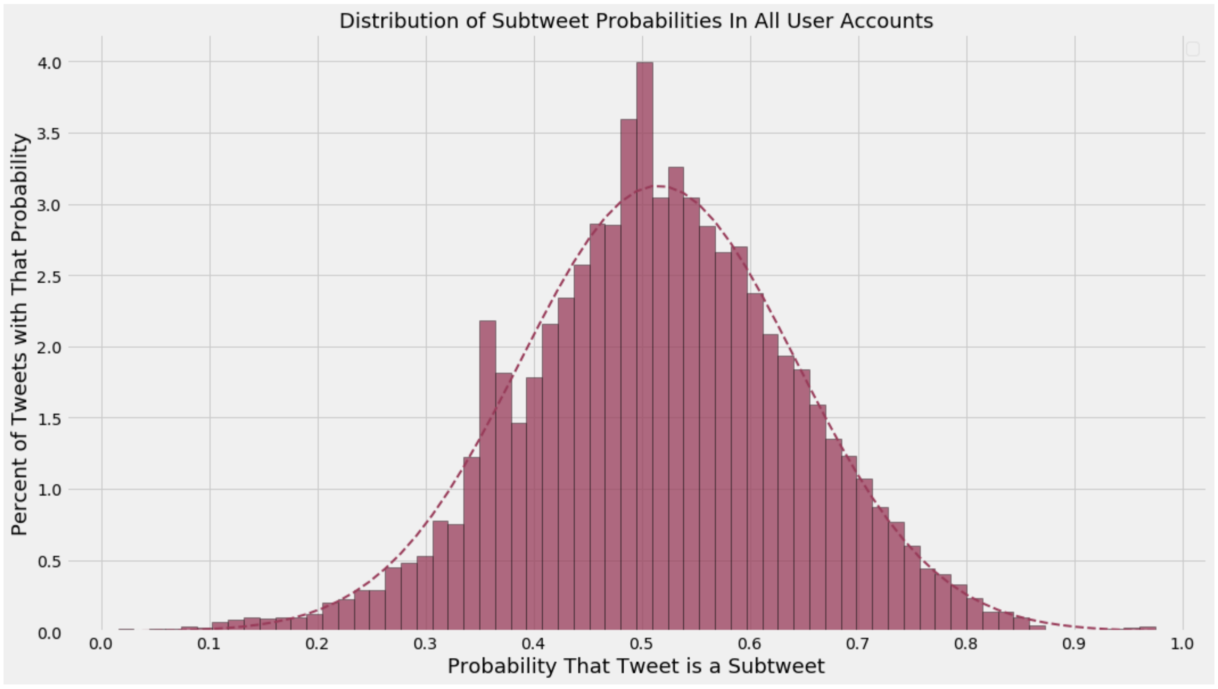

Figure 3.4.2. Distribution of Subtweet Probabilities for 14 Known Subtweeters ($\mu_1 = 0.516, \sigma_1 = 0.138$)

3.5 Most Informative Features

By iterating through the entire vocabulary of features on which the classifier was trained, and sorting those features by the logarithm of the estimated probability that the feature belongs in the subtweet class, we are able to see which features are most informative for classifying tweets as **predicted subtweets**. Consider this table:

Feature	Log Probability
.	-7.655317955479811
,	-8.022304498271094
GENERIC_URL	-8.048671040237405
”	-8.148437784201970
people	-8.427022837390009
?	-8.552068283434405
like	-8.672840985124703
don’t	-8.682709277848670
just	-8.744231174479117
i’m	-8.832921046724476
!	-8.863926249605123
it’s	-9.079518769840222
. GENERIC_URL	-9.100407613290404
:	-9.109587924777987
know	-9.183559311149267
...	-9.204228807382403
you’re	-9.210332638698153
twitter	-9.283961663904783
love	-9.328945234571764
*	-9.404169039956173
friends	-9.418393807734482

The only bigram in the top 20 most informative features is “. GENERIC_URL,” which surely occurs when a URL follows the end of a sentence.

3.6 The Twitter Bot

After training and testing our classifier, we utilized it in creation of a Twitter bot which interacts with **predicted subtweets** in real time.

3.7 Discussion

[Discussion of results.]

4
Conclusion

4.1 Summary of Project Achievements

[Ground truth data, classifier, cleaning, Twitter bot.]

4.2 Future Work & Considerations

[Test other classification algorithms. Utilize more training data. Clean the text differently.]

Bibliography

Go, Alec, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification using Distant Supervision*, CS224N Project Report, Stanford, 12.

Stone, Biz. 2006. *Introducing the Twitter API*, available at `https://blog.twitter.com/official/en_us/a/2006/introducing-the-twitter-api.html`.

Twitter. 2018. *Twitter API Docs*, available at `https://developer.twitter.com/en/docs`.

______. 2016. *Twitter Terms of Service*, available at `https://twitter.com/en/tos`.

Roesslein, Joshua. 2009. *tweepy Documentation* **5**, available at `http://docs.tweepy.org/en/v3.5.0/`.

Twitter. 2018. *Tweet objects*, available at `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object`.

Bird, Steven and Edward Loper. 2004. *NLTK: the natural language toolkit*, Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 31.

Zhang, Harry. 2004. *The optimality of naive Bayes*, AA **1**, no. 2, 3.

Borovicka, Tomas, Marcel Jirina Jr, Pavel Kordik, and Marcel Jirina. 2012. *Selecting representative data sets.*

Rae, Chelsea. 2009. *I hate when i see people...*, available at `https://twitter.com/Chelsea_x_Rae/status/6261479092`.

Urban Dictionary. 2010. *Subtweet*, available at `https://www.urbandictionary.com/define.php?term=subtweet`.

Edwards, Autumn and Christina J Harris. 2016. *To tweet or subtweet?: Impacts of social networking post directness and valence on interpersonal impressions*, Computers in Human Behavior **63**, 304–310.

Madrigal, Alexis C. 2014. *Behind the machine's back: How social media users avoid getting turned into big data*, available at `https://goo.gl/h36jxx`.

Dewey, Caitlin. 2016. *Study confirms what you always knew: People who subtweet are terrible*, available at `https://goo.gl/SeV3mx`.

Hassler, Chelsea. 2016. *Subtweeting looks terrible on you. (you know who you are.)*, available at `https://goo.gl/NCz27z`.

Ohlheiser, Abby. 2017. *A running list of all the possible subtweets of President Trump from government Twitter accounts*, available at `https://goo.gl/hFh81R`.

Twitter. 2012. *Twitter turns six*, available at `https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html`.

Dorsey, Jack. 2006. *inviting coworkers*, available at `https://twitter.com/jack/status/29`.

Nassirtoussi, Arman Khadjeh, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. *Text mining for market prediction: A systematic review*, Expert Systems with Applications **41**, no. 16, 7653–7670.

Burnap, Pete, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. *Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack*, Social Network Analysis and Mining **4**, no. 1, 206.

Allport, Gordon W. 1954. *The nature of prejudice.*

Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. *Semeval-2015 task 11: Sentiment analysis of figurative language in twitter*, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 470–478.

Van Hee, Cynthia, Els Lefever, and Vronique Host. 2018. *Semeval-2018 Task 3: Irony detection in English Tweets*, Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018).

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal **5**, no. 4, 1093–1113.

Kelly, Ryan. 2016. *PyEnchant: a spellchecking library for Python.*

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. *Scikit-learn: Machine learning in Python*, Journal of machine learning research **12**, no. Oct, 2825–2830.