Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 5th, 2017

The Correlation between the Topic and Emotion of Tweets through Machine Learning

Vincent Fiore, Kenneth Almodovar, Ange Assoumou, Debarshi Dutta, and Andreea Cotoranu
Seidenberg School of CSIS. Pace University, Pleasantville, New York

***Abstract* -- Twitter ranks as the second most popular social networking platform available on the Internet. Millions of people use the platform to express opinions on a wide variety of topics through Twitter posts (tweets). These tweets give insight into popular opinions on various topics, and gauge the public perspective and attitude towards numerous issues. Therefore, Twitter data is desirable for both scientific research and marketing. This study implements the Python SciKit-Learn library to classify tweets according to three topics and three emotions. We achieved 90% classification accuracy for tweet topic, and 87% accuracy for tweet emotion using Linear Support Vector. We further investigate the correlation of tweet topic with emotion by analyzing sample tweets from five public figures. The results confirm that there is a direct correlation between topic and emotion at the individual level, although no strong correlation is observed across the entire data set.**

***Index terms* -- Machine Learning, SciKit-Learn, Sentiment Analysis, Bernoulli Naive Bayes, Gaussian Naive Bayes, Support Vector Machine, Classification.**

I. INTRODUCTION

The aim of this study is to analyze Twitter posts (tweets) to determine the topic and the emotion associated with a tweet. Tweet-analytics has gained significant popularity over the past few years. The analysis supports both marketers and researchers to efficiently gain information about a group of people they wish to study [11], [12]. In this study tweets are categorized into three distinct topics: politics, religion, and family, and three distinct emotions: happiness, depression, and anger. Once categorization is complete, the goal is to determine where topic and sentiment overlap in order to flag specific topics for further investigation.

A tweet includes a maximum of 140 characters, and given the concise nature of the text, these messages are suitable for analysis [11]. People, those sharing personal opinions in particular, use Twitter primarily for impromptu posts. The topics of these posts vary greatly and include products, brands, food, celebrities, sports, and the daily weather, to name a few. The opinion an individual expresses about a topic in a tweet is always definite because of its concise nature. In other words, it is unlikely that a person will start a tweet praising a subject and then change their mind halfway through,. However, this type of behavior can be observed when analyzing online product reviews for example. Product reviews may start by listing the positive aspects of an item, only to expand upon negative feedback later. Additionally, the public nature of Twitter and the relatively open Twitter API makes Twitter data collection and analysis accessible.

In this study we will implement machine learning tools to classify tweets by topic and emotion. The outcomes of this study include a process that can take any tweet, analyze it, and then determine the topic and emotion categories associated with the tweet. This information is then plotted to identify correlations between the two categories. If run over a broad selection of tweets, the analysis has potential to reveal trends in the relationships between specific topics and the emotion associated with these topics.

The results of this analysis can have significant implications if genuine associations between topics and individuals' emotions are identified. The categories of religion, politics, and family are expected to display clear patterns. In many cases, revelations about public opinions on certain issues can illustrate how society feels about the topic as a whole. Furthermore, processing beyond what is analyzed in this study can reveal much about specific issues within each topic [14].

II. BACKGROUND

Extensive research has been done in the areas of text classification and tweet analysis. For the purposes of this study, the literature review is focused on three areas: general text classification, social network text classification, and tweet classification by topic. Previous works provide valuable insight into the process of text classification and serve as a starting point for this study.

The work of Dalal and Zavery [4] on general text classification describes text preprocessing. Their work includes the following steps: determine sentence boundaries, eliminate “stop-words” from the text, and “stem” the text. Stop-words are words that are common and serve no purpose in determining meaning, such as “an,” “the,” or “of.” Stemming involves trimming a word down to its root, usually by removing all suffixes and plurality. This is a common first step for all machine-learning programs. Examples of stop-words include: a, about, and, as, at, because, but, by, do, for, from, here, how, it, only, or, out, same, so, some, that, their, to, too, you, your.

Data preprocessing contributes to making searches more reliable and algorithmic processing more efficient. For example, by stemming words, different words that can take a while to process separately can be processed much faster if presented as one root word. Removing stop-words also streamlines processing. Since stop words have no significant meaning in and on themselves, removing them highlights the base meaning and context of the tweets. Although such removal would not be suitable if the text is to be read by a human, it expedites processing by computer algorithms.

Bennett et al. [2] focused directly on creating a dictionary of words that can efficiently identify sentiment for Facebook posts. For example, by examining basic interjections, sentiment is typically highlighted without the need for further text processing. The researchers kept a list of words that were particularly harsh, positive, or negative, and then manually associated these words with an emotion that was likely for the user. This same method can also work when looking for words that may hint to a tweet’s general topic. If it is possible to create a list of words that can be associated with a particular topic, such association can be done manually, and the output can inform the learning, as part of the machine learning process.

Similarly, Bennett et al. [2] also analyze emoticons to reveal the emotion of tweets. In the proper internet lexicon, emoticons are a mix of colons and parenthesis that create the appearance of human faces. This method does not require any processing of the text beyond simply searching for the emoticons themselves. This approach reveals that certain classification methods do not process the actual text, but rather other elements of the message. In the work of Bennett et al., emoticons were the only feature that distinguished Facebook posts from other types of text, like a book for example. By searching for specific words in much the same way Bennett et al. searched for emoticons, it will make it possible to predict both topic and emotion with speed and accuracy.

Go et al. [7] focus on tweet analysis. Their work took note of possible differences in the sentiment of full sentences. According to the authors, it is much more important to understand the sentiment about a topic rather than the sentiment of the topic itself. For example, if a tweet is classified as positive for simply mentioning the product or topic under research, this may be ignoring the fact that the sentence is actually negative about the topic.

In these cases, a superficial classification of the tweet has the inherent risk of ignoring or missing the actual meaning of the tweet. This proves that it can be difficult to properly classify the emotions of tweets because of the challenges inherent to human communication. Furthermore, certain texts, like tweets, are difficult to categorize because of their sheer size. However, one of the convenient tweet features is the 140 character length, which is unlikely to allow the user to develop mixed opinions. In long form text, writers may start to develop positive opinions on a topic only to dive into much more extensive negative opinions later. This aspect introduces processing problems far beyond the scope of this research.

Go et al. [7] also use emoticons to help differentiate the general tenor of tweets. This appears to be the fastest methods to generate initial data about tweet emotion. This initial data can be further refined based on additional methods. Specifically, when combined with a large dictionary of words, these two features can help narrow down emotion to a large extent. Unfortunately, there is no similar method for posting about specific topics, so it is still necessary to use word lists. Similarly, only a few tweets in the dataset include emoticons, which made this feature not relevant for this study.

III. METHODOLOGY

A. *Data Set*

The Twitter data set used for this study is Sentiment140 [7]. This data set includes two subsets: categorized and uncategorized Twitter data. The categorized subset is comprised of roughly 1,000 tweets that have been hand classified as either positive or negative, and have the subjects of each individual tweet highlighted. While this data set can be incredibly useful for certain types of training, it does not support this study as topics are not broad enough. In addition, it is difficult to find 300 tweets to encompass each of the six categories targeted for this study.

The uncategorized data set is more suitable for this study. This corpus includes roughly 1,000,000 raw tweets. It is in this set that we identify the 300 necessary tweets by simply browsing through the data and searching for specific words.

B. *Data Processing*

The Twitter data set is preprocessed through several Python modules, according to methods described in prior research [5, 6, 7]. The intention behind preprocessing is to optimally categorize and sort the data. Scikit-learn [6], an open source Python module, is selected to perform the classification. However, the classification includes several data preprocessing steps, as described below.

First, a subset of data is preprocessed manually. As the machine learning algorithms need data to learn from, the initial process trains the algorithms on recognizing specific patterns. About 300 tweets were labeled for topic and emotion. Although Sentiment140 [7] categorizes tweets by sentiment (positive or negative), the positive/negative classification is irrelevant for this study.

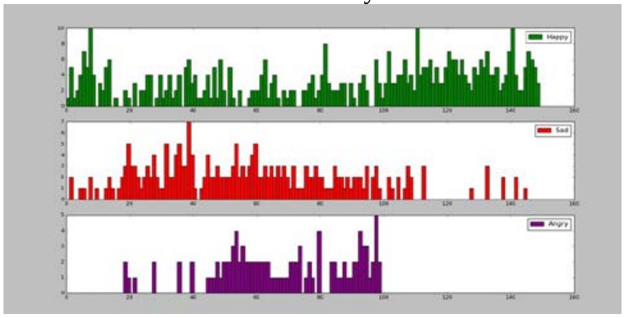

Fig. 1: A preliminary classification Tweets by emotion, based on word lists

About 1,000 tweets are classified in this manner, and in most instances these are too narrow in scope. A large number of tweets are rejected as they do not fall under the six categories targeted in this study. Although a tweet's positivity or negativity may have a profound impact on the actual topic or emotion, that information does not inform this study. Instead, the focus is on extracting specific tweets from the Sentiment140 training set. This set includes nearly one million Tweets, and allows for much finer control of the content, such as searching based on relevant words, and then extracting tweets that match one of the six target categories.

Secondly, a list of words thought to be associated with each topic and emotion is created. These are common words that appear in tweets that fall into one of the target categories: politics, religion, family, happiness, depression, and anger. According to previous research, tweets tend fall into one of three categories: posts about a subject in formal sentences, posts in a more conversational tone without formalities, or posts that are a mix of the two, including proper sentences and short thoughts or ideas [1]. Although many tweets tend to fall into the lattermost category, effective classification should alleviate any differences between the three. These categories are simply not present when properly processing tweets as each post is condensed to its root, which rarely falls into any category at all, and just appears to be a collection of related words. This next step should address this concern.

Thirdly, the data is normalized and sanitized manually: sentence boundaries are determined, stop words removed, words stemmed, and other normalization techniques performed before moving to classification. This includes compensating for repeated letters, which many Twitter users use for emphasis, and removing and categorizing punctuation where applicable.

After the text has been normalized, it is compared against the word lists. This will generate a count of the number of words in each category matched by the tweet. These numbers can be combined with the tweet text itself, variables like the length of the tweet, or any other suitable information gleaned from the text that might be significant. Once preprocessing is complete, the data is then passed to the machine learning algorithm for classification. At first, previously classified text is run to test system accuracy and a benchmark is established. Then, the algorithm is trained to identify tweets that match the search terms.

The classifier then checks for links between the topic and emotion for each post. As an outcome example, it is expected that tweets about politics might show anger whereas those about family might show happiness. This correlation can be studied further to answer various questions related to Twitter text that fall into these categories.

C. *Feature Selection and Training*

Feature selection is critical to the classification process. Although, the word lists generate features, additional features are needed [5]. These will include features that are inherent to Twitter posts before preprocessing, such as length and time of post. Leveraging Twitter's own formatting, extracting features regarding whether or not a tweet is in response to another user can also be used as a feature. "Replies," as they are called on Twitter, are delineated by a "@" followed by a username. This makes them easy to locate by a simple text search and fast to process. All together, these features are likely to be unique and independent. These qualities make these features suitable to use without any further processing, and are expected to be effective in the training process [8].

The training process is a challenging part for any classification problem [6]. It is imperative to determine not only the best method to classify the data, but also to achieve accurate results on a small scale. As the training set may not be indicative of large scale Twitter data, choosing a highly diverse training set is critical.

In this study, the training process begins with manually classifying 50 tweets from each category in both topic and emotion. The result is a set of 300 tweets to be used for training. Based on previous research [8], the size of the training set should be sufficient to achieve accurate classification results. The size of the training data set was established to meet the following criteria: diversity of content and ease of manual processing.

Scikit-learn, a Python library provides most of the tools needed to manipulate the data. Specifically, Scikit has built in processes for text manipulation that can automatically remove stop words, determine sentence endings, and stem words. Python was selected because it integrates Scikit-learn and libraries for large-scale data manipulation across different file types [10].

D. *Manual Classification*

Machine learning algorithms work by first learning from a large amount of already classified data, and then making predictions based on previous learning. One hundred and fifty tweets were chosen for both sentiment and topic. This selection was based on Fisher's Iris data set [6], which includes the same number of flowers, each manually classified into three separate categories. Although a 300 training sample set is less than what some machine learning algorithms require for training, the topic is sufficiently narrow to support this type of training.

For classification, each topic is targeted separately rather than looking for tweets that embody two or more topics simultaneously [14]. Machine learning algorithms often work best when attempting to come to only one decision at a time. This means that although the algorithm will analyze the tweets for both topic and sentiment, this will not occur simultaneously.

As previously mentioned, the topics of politics and religion are particularly difficult to hand classify, each for their own reasons. For tweets on religion, it is important to categorize posts that are genuinely religious, and not those that simply mention a religious figure. In many cases, people use words commonly associated with religion to express shock, anger, happiness, or any number of wide-ranging emotions. This requires verification that no particular tweet is classified as religious just because the author used a colorful expletive for example.

The category of politics is also challenging for classification purposes because the data set is not diverse enough. Sentiment140 includes tweets from 2009; however, the political landscape has changed significantly over the last eight years. This meant that many tweets that were considered political at that time mentioned politicians, policies, and topics that are no longer relevant. Furthermore, it would be inaccurate to ignore the large number of topics that are currently political, but were considered non-political at the time.

The complexity of the problem demands a multi layered solution. Firstly, it became necessary to look for politics beyond simple searches of current political figures. Instead, searches based on political issues that are still relevant and those based on general political ideas were heavily utilized. Secondly, in constructing a word list, it became imperative to use a larger word set that would be relevant for 2009 and beyond. By including both the issues and names that were popular eight years ago and those that are relevant today, the issue should be almost entirely mitigated. Tables 4 and 5 include the word lists created for the topic and emotion categories.

Table 4: Sample of the word lists for the target topics, truncated for length.

Politics	Obama, Hillary, Clinton, Trump, Senate, Politics, election, elections, Conservative, Republican, Democrat, Parliament, Fox News, CNN, MSNBC, Governor, President, Mayor, liberal, Political, Democracy, Terrorist, EU
Religion	Jesus, god, buddha, zeus, catholic, catholicism, christian, jewish, muslim, allah, prayer, church, Christ, praise, religion, religious, atheist, saint, atheism, parish
Family	Family, mom, dad, mother, father, son, aunt, uncle, cousin, niece, nephew, pop, ma, pa, boy, baby, kid, parents, grandma, grandpa. grandmother, grandfather, grandparents, bro, sis, daughter

Table 5: Sample of the word lists for the target emotions, with obscenities removed and truncated for length.

Happiness	Happy, Great, Love, yay, incredible, good, wonderful, fun, joy, awesome, excited, fantastic, enjoyed, beautiful, lucky
Sadness	Sad, sorry, depressed, devastated, upset, miserable, cry, tears, worst, distraught
Anger	Stupid, Angry, Pissed, mad, horrible, hate, annoyed, irritated, miserable

IV. HYPOTHESIS

The primary assumption is that the classification of tweets by topic and emotion will work correctly. The secondary assumption is that tweets about religion and politics will lean substantially more towards anger and sadness, whereas those about family will likely exude happiness.

The lessons learned from this experiment can be applied to different problems, such as large scale sentiment analysis, especially for publically available data, and for topics of current interest. For example, researchers can apply this to politics and assess how a voting populous feels about democracy. With the continued analysis of a particular topic, trends may appear regarding these same topics.

The topics of religion and family have much broader implications for research, but are nonetheless invaluable. Determining how people feel on a large scale about religion or about their home life can reveal how society at large may be feeling about these topics. This large-scale monitoring of emotions based on broad topics can inform a wide variety of groups and is really where this research will excel. If these correlations can be monitored over a long period of time, they may reveal more about the world we live in. This type of trend analysis has already been used in other fields with promising results [15].

The common hypothesis was also that this twitter data may be useful on a more specific level regarding individual users. Although the research has obvious benefits when classifying tweets of a large population at random, if focuses on the tweets of one particular person, interesting trends may appear all the same. For example, analyzing the correlation between the topic and emotion of tweets for a political candidate over time may reveal the ways in which she or he has changed as an individual. If the classifier is directed towards a Twitter account that represents a public entity, for example, even more trends can be revealed if compared to other data. Theoretically, if analysis is linked to a stock price of a brand, for example, a correlation may be revealed between the topic and sentiment of the tweets and an increase in price.

V. RESULTS

A. *Initial Results*

Support vector classification with a linear kernel achieves roughly 50% accuracy, without any changes to the training methods. Although this result is better than random guesses and shows improvement over previous work, the accuracy is lower than acceptable. While further training with different classifiers is expected to improve accuracy, there is no inherent flaw with this overall system responsible for skewing these results significantly.

Steps were taken to try to identify the issue behind low accuracy. Word lists are often difficult to properly train, so this became the main focus of the investigation. Upon running through the classification process while printing all of the words from the tweets that matched with the word list, clear problems began to present themselves. Before progressing, it was thought that the uneven length of the word lists may have had a negative effect on the frequency that words fell into any given category. However, it became clear that this did not have a major impact on the classification.

Instead, patterns began to emerge around words in the lists that caused dozens of false positives on each runs. In particular, these types of words usually fell into one of two categories. Since each list was many hundreds of words long, it was impossible to confirm that every word was equally relevant to the topic, had similar appearance frequencies, and did not repeat across any of the other lists. In particular, this was especially important for the topics of sadness and anger. This came as a surprise, but ultimately was rather clear upon further investigation. In many circumstances, words such as “hate” or “cry” can be classified in either list. This required further analysis of the lists themselves and decisions based on whether a particular word better represented one group or the other.

Some of the more troublesome words were “bad, cry, dead, die, feel, hate, kill, lie, mad, tears, upset.” These words appeared in the lists for both sadness and anger, and provided variation in classification. As they appeared in both groups of words, the classifier would increase the word frequency for both emotions, and could ultimately incorrectly predict the emotion for these tweets if there was not a sufficient enough number of other word list words in the tweet. Although these words could be classified as falling into one or more of these topics, they were removed due to their ambiguity. The context clues needed to properly identify the real category of these words fell beyond the scope of this research.

Furthermore, a second large group of trouble words began to present itself. This list came completely unexpected and likely had a more negative impact on the classification than the previous group. These words, who will now be referred to as “segment words,” or “seg words,” for short, appeared due to a flaw in the selected Python word matching algorithm. In order to ensure that this potentially difficult process was completed in a reasonable amount of time, a relatively simple selection method was used. This method, which utilized Python’s built “in” method, simply searched each tweet and determined whether or not the characters from each selected word appeared. Since there was no real analysis behind what words actually matched those in our lists, many of the matching words were not representative of the desired emotion at all.

These seg words usually consisted of small words that, while carrying meaning in their own right, were usually apart of many larger words with often conflicting definitions. One of the most common examples of seg words that were encountered also illustrates the difficulty in trying to remove them from the lists. “Sad,” in and of itself has an obvious meaning, but tends to be a part of many larger words. In one of the more comical encounters of this phenomenon, a tweet about a Mexican restaurant was classified as sad because it contained the word "quesadilla," which itself contained the word "sad." Although it was originally assumed that this would not be an issue because any word that contained a larger word would likely have the same meaning, this was often not the case.

Remedying this issue required careful consideration of the most common offenders and was often not a universal solution for each word. For example, words like "sad" needed to have a space appended to them, to ensure that the method would only consider a tweet to match if it actually contained the word "sad" and did not have "sad" as part of a larger word. For some of the other common seg words, deletion was a simple solution. Again, due to the large nature of the lists, some words that were almost entirely irrelevant began to slip through. Words like "aid," may have technically been considered to present themselves in happy tweets, but provided so many false positives that they needed to be removed.

A third and less prominent normalization technique utilized involved the removal of many nouns across all of the list emotion lists. Although these words provided relatively accurate data for the topic lists where nouns were a common delineator between each type of Tweet, these words in the emotion list only added confusion. In many cases, the nouns present in the list simply were not accurate or specific enough to reasonably fall into any of the categories or, at worst, could fall into more than one category. Although not even noun was removed across these three lists, this major elimination process helped in making the classification process substantially more accurate.

Further appearances of this phenomenon were observed with the following words: "aid, art, being, cool, family, feel, fine, give, hot, many, sun, trust, well, win, zing, away, down, fed, ill, low, mad, man, out, bad, badly, break, leave, love, and rob." In each case, these words were too frequently either vague or subsections of other words. Much like the previous words that caused false positives, these words could reasonably be classified into one of the various categories, but their inclusion was not worth the issues that they caused.

B. Support Vector Classification

Linear Support Vector Classification (SVC) was used for this study. Throughout multiple tests, SVC achieved roughly 90% accuracy in determining the emotion, and roughly 87% accuracy in determining the topical. As the base accuracy of each classifier would be 33%, a one in three chance of correctly predicting a tweet's category, these results are evidence of a working algorithm.

These high accuracy rates were surprising in that SVC is not typically the most efficient classifier for this type of data [5], [13]. According to SciKit's own documentation, the classifiers that perform well with this type of data are Bernoulli Naive Bayes and Gaussian Naive Bayes [13]. These algorithms are considered ideal for specific text inputs, and in cases where there is a smaller than average set of training data. Since these were thought to most likely provide the best results, accuracy averages were taken and compared against all three classifiers.

These results showed SVC to be the most accurate classifier when compared against the other two. In testing, both Bernoulli Naive Bayes and Gaussian Naive Bayes provided accuracy levels that were roughly 3-5% lower than SVC. Although the training data was split and randomized to provide more accurate testing, the various testing was done after the randomization and split occurred. In other words, each classifier was tested under the exact same data. These methods allowed for the following testing to be completed with a high level of confidence.

C. Final Results

The results on the correlation of emotion and topic for Twitter messages are promising. Five users, public figures, were selected and a sample of their tweets was analyzed.

This method was chosen to help in the analysis of the results and to confirm the accuracy of the classifier on real world data. These users were chosen because simple generalizations were easy to make beforehand regarding the content of their tweets. For politicians, for example, it was a clear sign that the topic analysis was correct if the classified returned a majority of tweets as falling into the politics category. Similarly, for certain users, the emotion of the tweets would be simple to ascertain before classification began. For this reason, the data received was particularly promising, as it all but mirrored the original assumptions.

A small setback was experienced because of the nature of the topics selected. More often than not, tweets failed to fall into a specific enough category to be reasonably classified as either family, politics, or religion. For the most part, these tweets, began to fall into the category of religion.

The quality and quantity of the dataset, the list of words are major factors that affect the result. The more words added to each list the more accurate is the result. Some words overlap between two topics or two emotions. The amount of words for each topic can make the difference. A political debate on the health care for example tends to classify the topic into "Family" even the main topic is "Politic". We collected tweets from 124 senators and 90% of them fall within "Family". The political tweets on the family welfare does create an ambiguity between the "Politic" and "Family". This issue does not appear on the emotions. It is necessary but not compulsory to compare the number of unique users to the number of tweets. Tweets from random users may show different languages. It is preferable to collect tweets amongst selected users within different social class.

For reasons that are not entirely clear, the classification algorithm tends to assume that most indeterminate tweets are of religious nature. Although the classification of emotion is correct, data skewed towards being highly religious is disregarded as anomalous.

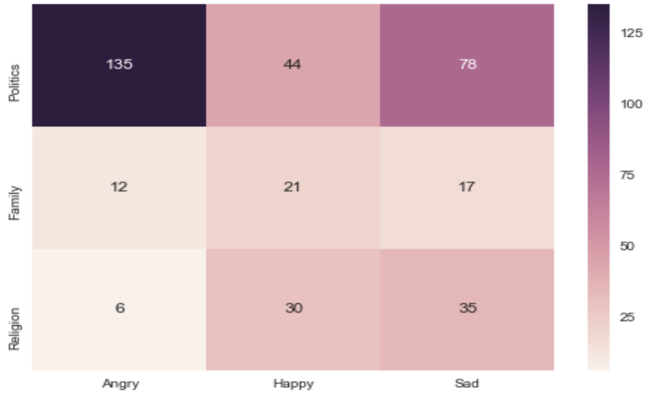

Fig. 3: The classification of Donald Trump's sample tweets.

In the categories where clear identification was able to occur, some patterns began to reveal themselves. For religion, ignoring the tweets that were most likely incorrectly identified, the majority of the tweets were classified as happy. This is most easily viewed for data produced for both the Dalai Lama and Pope Francis. These two cases serve as the most accurate of the classifications produced through this process, due in part to the relatively obvious nature of the topic that was expected and the way in which the results very closely mirror this theory. Each man's tweets provided almost identical heat maps. Although some sad and angry outliers were seen, these tweets rarely embodied the stronger levels of either anger or sadness that was common in our hand-classified tweets.

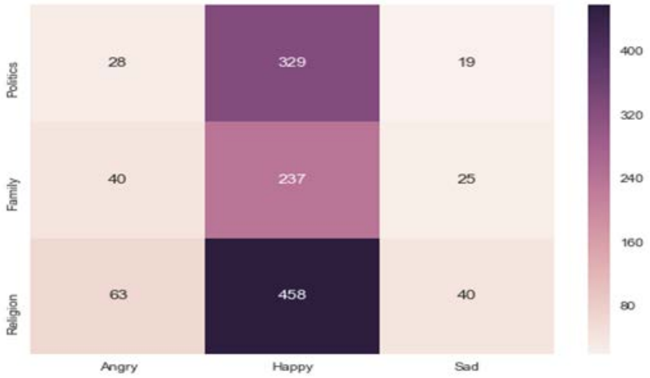

Fig. 4: The classification of the Dalai Lama's sample tweets.

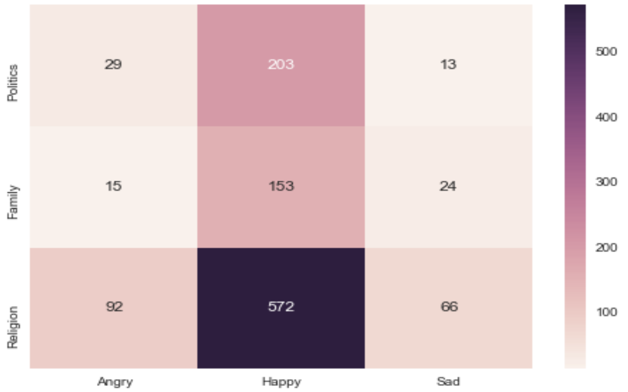

Fig. 5: The classification of Pope Francis's sample tweets.

For politics, a wider range of emotions was observed, but many tweets are clearly into either anger or sadness. For this study two political figures were selected: President Trump and Nancy Pelosi.

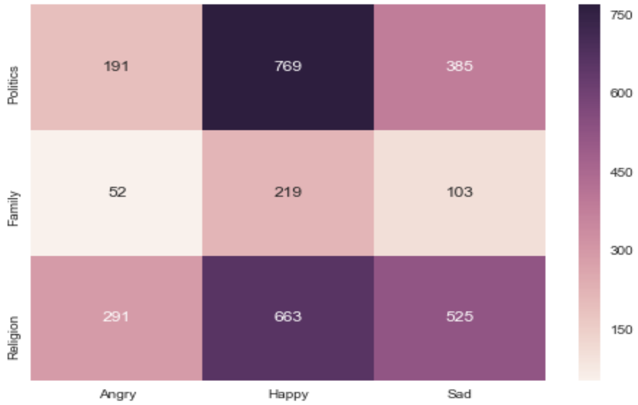

Fig. 6: The classification of Nancy Pelosi's sample tweets.

The President's tweets there are clear trends of negativity and sadness in tweets. Similarly, the majority of his tweets can be classified as political, although all three topics are fully represented. For Nancy Pelosi, her tweets tended to show happiness. It is worth noting that Pelosi's tweets are classified as overly religious regardless of their actual content. This result is an outlier and requires further investigation.

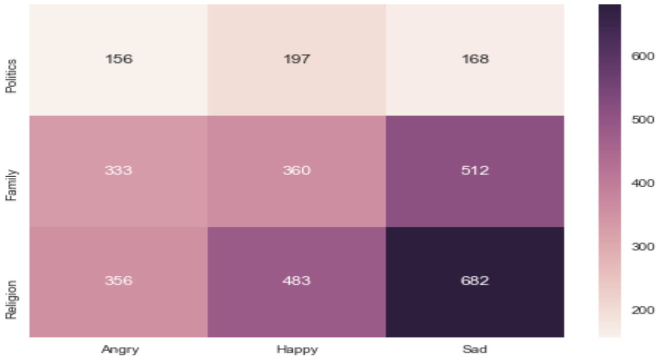

Fig. 7: The classification of Dr. Phil's sample tweets.

VI. CONCLUSION

The most interesting data trends in this study reveal more about the individual users than it does about the entire data set. For the most part, no overarching trends are observed related to topic and emotion. However, such trends are observed at the individual level, and as such, shift from user to user. Although it was hypothesised that certain topics, like religion, would tend to always show happiness, this was mainly seen when the user was positive with all of her or his tweets.

For example, the tweets of Pope Francis and the Dalai Lama are generally positive, regardless of topic. For users like President Trump, the negativity, either expressed through anger or sadness, persisted regardless of topic. In other words, it can be argued that topic and emotion do show clear correlation, but at the individual level.

Although it can be argued that a user's tweets should be consistent over time, this study points out that this is not always the case. Politicians and religious figures especially do tend to stay "on brand" when it comes to the topic and emotion of their tweets. This research confirms what a researcher can already infer is most likely the subject of a tweet and to get the emotion behind said tweet.

When analyzing political or religious figures, trends regarding their emotion can be illuminating in terms of their own state of mind or, over larger groups of these types of users, the general state of a topic. Unfortunately, with users who do not naturally fall into any specific topic, this information is obfuscated. This is due, in part, to the shortcomings of the classifier which tends to over classify tweets as religious, but can also be attributed to personal writing styles and personalities. Similarly, the inherent nature of classifying two separate data sets increases the error rate. Individually, each classifier achieved around 90% accuracy, but when combined the error rate is multiplied.

It can be very difficult to see any real trends develop unless massive samples of tweets are taken from large user bases. This research is ultimately limited in scale in the processing and data collection, and the limits of the researchers. If this research was carried out over a longer period of time, this would be the clear direction that it would need to take. If the sampling of tweets could come directly from a multitude of users, these sorts of trends may reveal themselves much differently. Instead of analyzing the correlations on a micro scale of individual users, macro analysis may reveal greater trends in the corpus at large.

Similarly, long term development might allow for the classification process to be altered to further remove the occurrences of the anomalous classifications. At the time of this research, narrowing down the possible cause of this issue proved challenging. Machine learning itself is not always an exact science, and it can be almost impossible to determine what causes a classifier to come to one particular conclusion beyond simply analyzing the trends in the feature set that it is fed. The consistency of the words lists and the datasets play a significant role in the classification. The features that define both topics and emotions are natural languages that evolve rapidly and need to change regularly. The generic steps in the process doesn't change and may be replaced by a software with a graphical user interface to upload and submit the file, then display the report with the chart.

Further enhancements to machine learning will contribute to gaining a deeper understanding into the reasoning process behind the classification, which will only help to improve these types of classifiers. As the data grows more accurate, so too will the results it can produce.

REFERENCES

[1] Chris Allen, M., Ming-Hsiang Tsou, Anoshe Aslam, Anna Nagel, Jean-Mark Gawron: "Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza", PLOS ONE, 2016

[2] Kristin P. Bennett, E.P.-H.: 'The Interplay of Optimization and Machine Learning Research', Journal of Machine Learning Research,, 7, pp. 1265-1281, July, 2006.

[3] M Bouazizi and T. Ohtsuki, "Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets", Global Communications Conference, Feb. 2016

[4] M. Dalal and M. Zaveri, "Automatic Text Classification: A Technical Review," Semantics Scholar, 2011.

[5] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, no. 10, pp. 78, Jan. 2012.

[6] R.A. Fischer, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, pp. 179-188, Sep. 1936.

[7] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," pp. 1-6, Jan. 2009.

[8] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003.

[9] S.B. Hamouda and J.Akaichi.: "Social Networks' Text Mining for Sentiment Classification: The case of Facebook' statuses updates in the "Arabic Spring" Era", International Journal of Application or Innovation in Engineering & Management, vol. 2, no. 5, pp. 1-9 May 2013.

[10] McKinney, W.: 'Python for Data Analysis', Sebastopol: O'Reilly, 2014.

[11] A Mollett, D.M., Patrick Dunleavy, "Using Twitter in university research; Teaching and impact activities," pp. 1-11, 2011.

[12] S. Moon, H. Park, C. Lee, and H. Kwak, "What is Twitter, a Social Network or a News Media?", pp. 1-10, 2010.

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," Journal of Machine Learning Research, vol. 12, no. Oct, pp. 2825–2830, 2011.

[14] L. B. Shyamasundar, P.J.R., "Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms," India Conference, 2016.

[15] J. Zaldumbide and R. Sinnott, "Identification and validation of real-time health events through social media," IEEE International Conference, Dec. 2015.