TWITTER ANALYSIS ON MARIJUANA

Guru Prasad Natarajan (A20344932)

INTRODUCTION

Twitter is one of the most popular social media and provides ways for its users to easily and instantly connect to a mass audience. The ability to engage with unknown persons is a unique appealing feature that helps drive its popularity. From survey conducted in 2013, it has been found that Twitter is viewed as the "most important social media service" which also opens doors for users to engage with adverse tweets that glamorize harmful substance behaviors.

We sought to examine the sentiment and themes of marijuana-related chatter on Twitter and describe the demographic of the users such as gender and ethnicity. We basically try to segregate the tweets based on the sentiments, as pro-marijuana, anti-marijuana, neutral marijuana and marijuana for medicinal use.

We wanted to do such analysis because marijuana-related harms may afflict some individuals; therefore, our findings should be used to inform online and offline prevention efforts that work to target individuals who are most at risk for harms associated with marijuana use. We assume that the terms used by the users would help us in classifying the tweets and the first names and last names of the users would help us in determining the gender and ethnicity respectively.

DATA

Twitter is the only source for our data collection. We used rest API to collect data. The data that we collected are based upon the following keywords "weed, marijuana, pot, pothead, cannabis, joint". We collected around 10,000 tweets over a period of two weeks and dumped it into mongoDB for later retrieval.

From the tweets, we are interested only in the following fields – text, description and user name. For gender determination we use name lists from the period 1890 – 2010 and for ethnicity determination we used a web scrapper to get the first 1000 ranks in each ethnicity.

METHODS

The first step is focused on slicing a portion of the collected data and hand-labeling them. By looking at the terms used in the tweets' text, we classify the tweets into pro-marijuana (1), anti-marijuana (-1), neutral-marijuana (0) and marijuana for medicinal use (2). This section of the data is tokenized by removing punctuations and other things that might not

help much with the classification process. We create a sparse matrix for the above and cross-validation is performed using Logistic Regression model. The average accuracy of the model is determined by repeating the process for the specified number of times. The top terms' coefficients are provided which might prove critical in the classification process. For the gender and ethnicity determination we obtained the user names of the tweeters and segregate them into first and second names. Using the name lists approach and first names of the users we determine the gender and we used k-neighbors classifier to help us in determining the ethnicity based on the last names of the users.

EXPERIMENT

The initial experiment focused on classifying the tweets. From those results it has been observed that pro-marijuana tweets tend to outnumber its counterparts. We summarize that results in Table (1), which provides the distribution among the number of tweets collected for the study. We used two models for predicting the sentiments; using Logistic Regression we obtained an average accuracy of 78.4%, whereas using SVM's LinearSVC we got only 75% as the average accuracy.

Category	**% of tweets**
Pro – marijuana	59.56
Anti –marijuana	9.8
Marijuana neutral	9.8
Marijuana for medicinal purposes	20.73

Table 1: Categorization of tweets

After the classification, we were interested in determining the gender distribution of the users who use tweet about marijuana. Since most of the identity of the users is unknown, we obtained a large number of unknowns, yet the number of male tweeters exceeds the number of female tweeters by a large amount. The data is shown below in Table (2).

Gender	**% of tweets**
Male	34.23
Female	6.04
Unknown	59.73

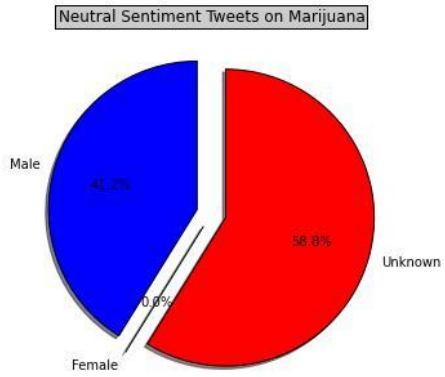
Table 2: Distribution of gender based on tweets

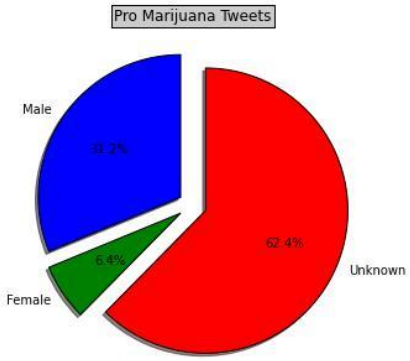
Next we summarize the number of male and female distribution in each of the classified categories in the below Table (3) and in a pie-chart.

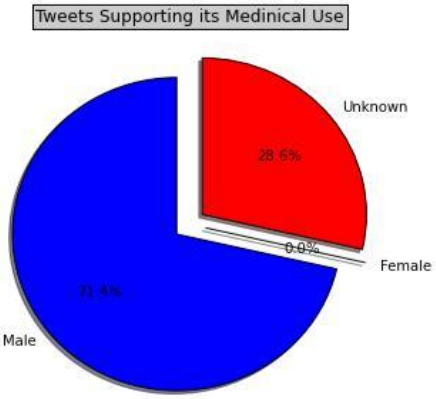
	Gender		
Category	**%Male**	**%Female**	**%Unknown**
Pro-marijuana	31.2	6.4	62.4
Anti-marijuana	37.37	11.11	51.85
Marijuana neutral	41.68	0.0	58.82

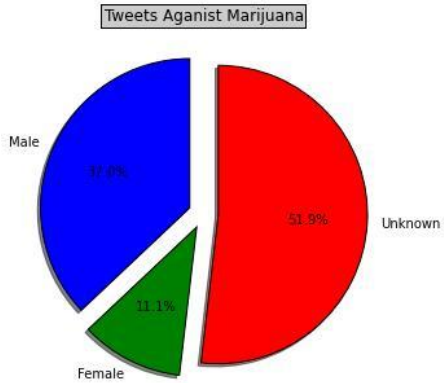
Marijuana for medicinal use	71.42	0.0	28.57

Table 3: Categorical distribution of tweets among genders

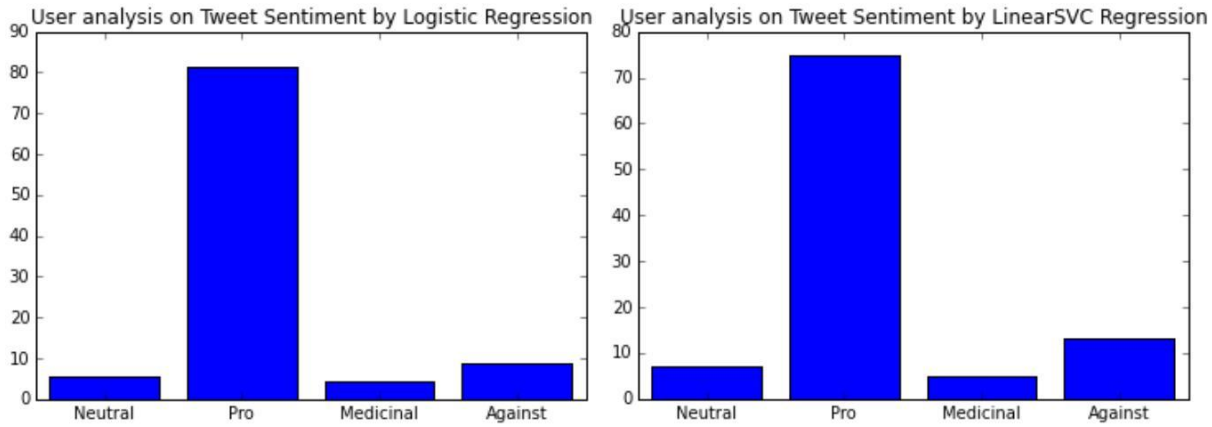


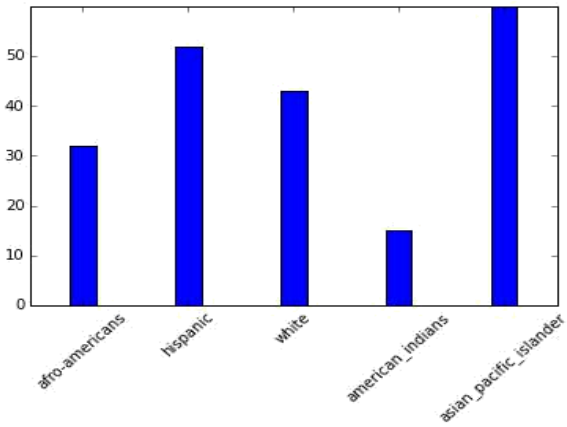






Comparisons of the classifiers are shown in the below histogram.





RELATED WORK

Our work concentrated on collecting tweets based on the keywords used in the tweets. In the related work we referred to, tweets are obtained from a particular handle and the analysis is made from it. This formed the basis for our work. Rather than collecting the tweets from a handle we opted to collect tweets based on the keywords and performed our analysis. This analysis helped to find the distribution of tweets in every category that we anticipated. Also the referenced work did not have a way to categorize tweets that support marijuana for medicinal purposes. But in our approach, we are trying to classify even the tweets that support medicinal marijuana.

CONCLUSION AND FUTURE WORK

Our work concentrated mainly on categorizing the tweets into four categories and the results show that most of the twitter chatters are inclined towards positive sentiment towards marijuana. Furthermore results show that majority of the tweets are from male profiles which proves the point that the usage of marijuana is common among males. However in order to have a more detailed analysis, it might require at least six months to collect quality data. Since with the data collected within two weeks we were able figure out that most of the people did not prefer to reveal their personal identity which would provide an in-depth analysis of the tweeters.

In future, we are planning to determine the age distribution of the tweeters and their location information to help determine which region has got the most number of marijuana tweets and what percentage of the people are for and against marijuana because ago factor plays a significant role towards marijuana usage. Also, since the current methodology of determining the gender and ethnicity was not that much accurate we are planning to have a much more intelligent system to determine those details.