Interest Prediction via Twitter

Yutthana Srisakunkhunakorn
YS574859
YSrisakunkhunakorn@albany.edu
Computer Science

***Abstract*—Social media is the increasing research area in Computer Science. Its popularity comes from the massive use by people in sharing the information online. The information shared by users can be further utilized to classify people's interests. This information is crucial for companies in order to understand customer's decisions. This study classified people's interests based on their contents on Twitter into three categories include Music, Sport, and Politics. The total of 45 accounts were analyzed. We employed Support Vector Machine (SVM) and Logistic Regression to identify the common interests of the studied sample. The results show that our proposed model was capable of predicting people interests based on their information shared on Twitter. These findings can be further used in recommending the products or services based on users' interests shown based on their tweets.**

Keywords—Social media; Twitter; Support Vector Machine; Logistic Regression; Word cloud, data clustering

I. INTRODUCTION

Existing research on computer science has emphasized the importance of social media, especially in relating to how people communicate and exchange the information. Social media are the platforms that stimulate users' generated contents and in turn become the great sources of information. Scholars can utilize the information shared on social media to learn about people's online behaviors that may influence their decisions. Social media such as Facebook, Twitter, Linkedin have different mechanisms that serve various purposes of communication. In this study, we mainly focused on the information shared on Twitter as the representative of the other social media due to its popularity. Twitter is one of the most popular social media that people can share short SMS-like messages called tweets. On the Twitter platform, users can share opinions, thoughts, links and pictures; for example, journalists can comment on the live events, companies can promote products and engage with their customers. Twitter is very active social media platform where users tweet for approximately 500 million tweets per day. Therefore, computer science research scholars can utilize the information shared on the Twitter to analyze the behavior pattern happening in the online community.

In this study, we provide the evidence of how information shared on Twitter can be used to categorize people's preferences. Based on our sample's tweets data, we identify their common interests and therefore be able to recommend the products or services to them.

II. CATEGORIES SELECTION

Before we begin the training of the model, first thing we need to consider is the selected categories need to be diverged from each other and these categories need to tell us about the interested of the user. Then, we need to collect the training tweets from three categories. we purposively selected twitter's accounts that are related to our interested categories; Music, Sport, Politics. For music category, we recruited 15 accounts that represent music group, for example, we have Justin Bieber, Selena Gomez, Taylor Swift in our analysis. For sport category, we also selected 15 accounts that are related to sport, for example, Leister city FC, Barcelona FC, Real Madrid FC. For politic category, we analyzed 15 Twitter accounts of politicians such as Donald Trump, Hilary Clinton, and Barack Obama.

A. Music

Music is one of the categories that we selected. We can, absolutely, separate music from politic but some users who are interested in music might also interested in sport. However, if you are asked a question "Which one do you like between music and sport?" We think that most of the people can answer it without thinking.

B. *Sport*

As same as Music, Sport is one of the major interested in most of the tweet user. We can easily separate sport from politic, but, again, it is hardly to separate this category from music.

C. *Politic*

This category can be separated very easy from those two categories because of its characteristic. If we talk about politics, it is more serious topic than talking about sport or music. So, it is very to distinguish this topic from the rest.

III. MODEL

In this project, we are using two algorithms to classify the class of each tweet. Then, we compare each algorithm which give us the most accurate result.

A. *Support Vector Machine*

"Support vector machines (SVMs**)** are a set of supervised learning methods used for classification, regression and outliers-detection" [1]. It is effective in a high dimensional space. It is also very memory efficient since it uses a subset of training points in the decision function (called support vectors). However, Support Vector Machine cannot provide the probability estimate.

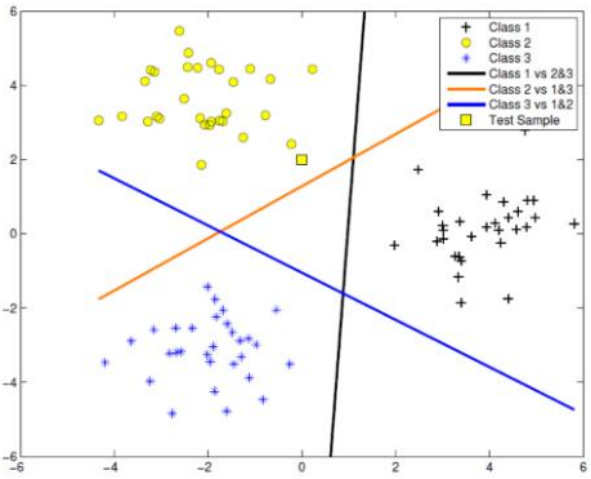

Figure 1 Multi class SVM

In our project we are using one vs rest approach. One vs rest approach will train one classifier per class by giving one class as positive and the rest are negative while one vs one will train each classifier for each difference pair of labels.

B. *Logistic Regression*

A logistic regression is one of the regression models used to investigate the relationship between predicting variables which can be a continuous variable, categorical variables or both, and the outcome variable which is binary or dichotomous. In addition, multiple logistic regression model exhibits the relationship between one dichotomous outcome variable with multiple independent variables. While Support Vector Machine cannot provide the probability estimate, Logistic regression can handle this problem and give each testing an estimate of probability. The following equation gives the logit of the multiple logistic regression model;

$$\mu_{y/x} = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+.....\beta_n x_m)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+.....\beta_n x_m)}}$$

after making the logit transformation as follows,

$$\mu_p = \log = \left[\frac{\mu_{y|x}}{1-\mu_{y|x}}\right],$$

we obtain the multiple linear regression model:

$$\mu_p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_m.$$

The logistic regression is pursued with more robust where the set of independent variables is distributed multivariate normally and not violating the common variable-covariance matrix. Unlike the linear discriminant approach, logistic regression method can be used without the strict distributional assumptions [2]. Therefore, it is expected that logistic regression technique can do better when there is evidence of multivariate normality as is the case where there are some dichotomous or zero/one variables or where distributions are highly skewed or heavy-tailed [2].

Prior literature argued that although both logistic regression and discriminant analysis could be interchangeably used for most of the analysis, but the former technique has two major advantages [3]:

- it is more robust to violations of assumptions of multivariate normality and equal variance-covariance matrices across groups; and

- it is similar to regression with its more easily interpretable diagnostic statistics

Further benefits of the analysis that raise its popularity come from the following assumptions:

- logistic regression does not assume a linear relationship between the dependent and independent variables;
- independent variables need not to be interval;
- logistic regression does not require that the independents be unbounded; and
- normally distributed error terms are not assumed.

IV. Data Preprocessing

After we collected the list of the tweet from various fields of categories, we then eliminate all the stop words. The stop word is the word that appear a lot in the content but does not provide any significant meaning to the content [4]. For example, the preposition word (in, on, above, and so on). We are using the list of the stop words provide in class in our project. After we remove all the stop word, we then left with the stemmed words of each tweet.

The stemmed words that we got after removing all stop words can then be used in our project. We can perform a simple analysis by using word count. In this way, we can see which words commonly used the most in each category.

Figure 2 Music clustering

From Figure 2, we use clustering and word cloud to visualize the highest number of word that appear in music category. The words that appear the most are new, get and see. These words reflect the significant contribution to music industry. For example, the word "new" would be likely to appear a lot when the new albums or songs were introduced to the market. Key music people would often use the word "new" to get audience attentions. Moreover, the word "get" would also be used as the call-to-action word in asking people to buy their products such as get it now or get the new product today. We also noticed that the word "see" would be used for music video posted on social media. For example, a musician would provide the YouTube links on their Twitter's timeline in which contain the word see before the URL links. Moreover, the word "see" might also be used for introducing the new products as well. In addition, we can also see the pattern of less frequent words that also reflects the key themes of music industry such as "live", "sale", "soon", and "wait." For example, the word "live" would be relating live music or the word "soon" would refer to the new albums or singles that are coming soon.

From Figure 3, the highest frequent words shown in word cloud such as fans and see show us that this list of tweet relates to sport. These keywords can be similar to many industries that having many audiences and therefore, the word "fans" and "see" were used quite often in social media. For the sports industry, audiences also refer themselves as big fans for their favorite sport teams. Furthermore, the words such as "win" or ranking terms such as "3rd" or "4th" manifest the uniqueness relating to Sport. People would share the story of winning, losing, or the achievement their teams received. We might also notice the words such as "celebrates", "life", "hero" to represent the nature of sport industry where the fans celebrate the achievement of their teams and also becoming their heroes.

Figure 3 Sport Clustering

Figure 4 shows the clustering of the politic category. According to the figure, the word "vote" was used most frequently in this sample, and it also represents this category. In democracy system, voting is the crucial term for people to choose their leader. Therefore, in our politician sample, it is understandable that they would use the word "vote" quite frequently. Moreover, during the time of our data collection, it was also the election period. We assume that the politician in our sample would encourage people to vote and therefore result in using the word "vote" a lot on their Twitter. Another interesting keyword is "can." Politicians make a lot of promises to people in trying to get their votes. In providing the promises, politicians would use the word "can" quite often to give people hope and trust. Furthermore, we realize that they used a lot of keywords that represent the national development issues such as "tax," "race," "abuse," "bad" and "act." These discovered keywords can be tied logically in laying the rationale representing people contributing to the politics.

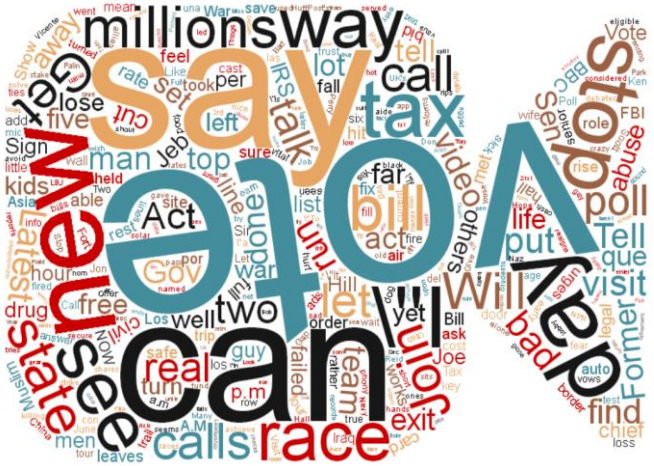

Figure 4 Politic Clustering

V. EXPERIMENT

After we eliminate all the stop words from the training list of twitter, we then count the number of each word. We only consider the word that has more than twenty time appearance in the context and the word that has more than two character. We end up getting 1410 words in our vocab list. So, our Support Vector Machine and Logistic Regression will have 1410 dimensions and 3 class label, Music as class "0", Sport as class "1", and Politic as class "2". Then we use the python package called "sklearn" to create Linear Support Vector Machine with one vs. rest approach. The reason we are using one vs the rest approach because it has a better in computation time than one vs one. One vs rest approach will train one classifier per class by giving one class as positive and the rest are negative while one vs one will train each classifier for each difference pair of labels. Also, to avoid overfitting of the model, we used 10 cross validation. From table 1, the best c parameter in our project would be 1, which give us the highest mean, 0.634.

mean: 0.63433,	std: 0.05757,	params:	{'C': 1}
mean: 0.63382,	std: 0.05791,	params:	{'C': 2}
mean: 0.63377,	std: 0.05776,	params:	{'C': 3}
mean: 0.63377,	std: 0.05780,	params:	{'C': 4}
mean: 0.63373,	std: 0.05792,	params:	{'C': 5}
mean: 0.63386,	std: 0.05765,	params:	{'C': 6}
mean: 0.63394,	std: 0.05777,	params:	{'C': 7}
mean: 0.63386,	std: 0.05764,	params:	{'C': 8}
mean: 0.63377,	std: 0.05765,	params:	{'C': 9}

Table 1. SVM 10 Cross Validation Parameter

We used the same python package "sklearn" to create the training model of Logistic Regression with 10 cross validation. From table 2, the best c parameter is 2, which give us the highest mean, 0.635.

mean: 0.63527,	std: 0.05688,	params:	{'C': 1}
mean: 0.63548,	std: 0.05697,	params:	{'C': 2}
mean: 0.63540,	std: 0.05703,	params:	{'C': 3}
mean: 0.63501,	std: 0.05768,	params:	{'C': 4}
mean: 0.63476,	std: 0.05740,	params:	{'C': 5}
mean: 0.63493,	std: 0.05739,	params:	{'C': 6}
mean: 0.63471,	std: 0.05754,	params:	{'C': 7}
mean: 0.63454,	std: 0.05768,	params:	{'C': 8}
mean: 0.63437,	std: 0.05781,	params:	{'C': 9}

Table 2. LR 10 Cross Validation Parameter

Based on the result from these table, we can see that Logistic regression gives a better result than Support Vector Machine.

After we trained the model, we then test the model with three well-known twitter accounts that are related to each categories. For music, we use Kanye West. For sport, we use Lionel Messi. For Politic, we use politico, which is the twitter account that tweet about the politic.

Kanye West	Logistic Regression	Support Vector Machine
Music	418	429
Sport	62	58
Politic	120	113

Table 3 Kanye West testing to be Music

From table 3, we can see that Kanye West clearly has interested in music.

Lionel Messi	Logistic Regression	Support Vector Machine
Music	206	209
Sport	245	242
Politic	149	149

Table 4 Lionel Messi Testing to be Sport

Lionel Messi has interested in sport the most but he also has interested in music too from figure 4.

Politico	Logistic Regression	Support Vector Machine
Music	159	167
Sport	80	70
Politic	361	363

Table 5 Politico testing to be Politic

Since politico is the twitter account that tweet about politics. So, we can see that it has the highest number of tweet related to politic for both Support vector Machine and Logistic Regression.

VI. Work Distribution

We mostly work together in every part start from come up with an idea on this project, working on collecting dataset, and testing. Table below shows the work distribution for each of us.

Name	Responsible
Yutthana	Creating the models Support Vector Machine and Logistic regression with 10 cross validations and testing the model.
Manish	Extracting data from list of twitter and testing.
Mounika	Data collection and testing the model.

References

[1] "1.4. Support Vector Machines — Scikit-Learn 0.17.1 Documentation". *Scikit-learn.org*. N.p., 2016. Web. 13 May 2016.

[2] Green, H., Boze, B.V., Choundhury, A.H. and Power, S. (1998), "Using logistic regression in classification", Marketing Research, Vol. 10 No. 3, pp. 5-31.

[3] Dawes, P.L., Patterson, P.G. and Midgley, D.F. (1997), "Involvement of technical consultants in high technology business markets", Journal of Business & Industrial Marketing, Vol. 12 No. 2, pp. 83-102.

[4] Marco, View. "Mining Twitter Data With Python (Part 1: Collecting Data)". *Marco Bonzanini*. N.p., 2015. Web. 14 May 2016.