Utilizing Supporting Entities for Realtime Sentiment Classification

黄 俊[†]　岩井原 瑞穂[†]

[†] 早稲田大学大学院 情報生産システム研究科 〒808-0135 福岡県北九州市若松区ひびきの 2-7

E-mail: [†] junhuang@akane.waseda.jp, [†] iwaihara@waseda.jp

Abstract This paper proposes realtime estimation of sentiment classification via supporting entities for messages posted on Twitter. Nowadays, microblogs like Twitter have gained wide popularity for its free style of updating personal opinions in a realtime manner, which increases the difficulty for classification. We argue that supporting entities, which are members of or closely associated to a given subject, can significantly handle the above problem .These supporting entities are literally neutral, but they show the tendency of Twitter users under certain circumstances. For instance, by repeatedly mentioning “Kobe” during the NBA Playoffs, the user may have a large probability to be a fan of the Lakers. Additionally, percentage of positive tweets, namely, support rate can be calculated in a realtime way. Drawing on World Cup 2010, we collect a large amount of tweets and carry out analysis so as to extract sentiment information of the audience and go further to show the realtime support rate of the participators.

Keyword Supporting Entity, Realtime, Sentiment Classification, Microblogs

1. Introduction

As a popular micro-blogging service provider, Twitter[10] has grasped tremendous attention from the public, consisting of several tens of millions of users who are actively participating in creation and propagation of contents on topics that range from entertainment and politics to technologies and products. Moreover, given Twitter's realtime characteristics, most of these tweets have strong correlation with events that recently occurred, or those are occurring, For instance, there are a huge number of tweets talking about the new Harry Potter film and the 61st NHK Kohaku Uta Gassen.

We discovered that these tweets which involve attitudes and comments of the audience can be used to build models to accurately capture changes of people’s emotion, if properly designed. They will be more accurate than other techniques for extracting diffuse information, such as costly surveys and opinion polls. Moreover, gathering information on people’s attitudes regarding particular things will be helpful when designing marketing and advertising campaigns [2].

Extracting the attitudes of blogs is usually treated as a sentiment classification problem which is well-studied in machine learning [2], [3] and linguistics [4], with different classifiers [2] and language models [4], [5] employed in earlier work. These methods work well on static corpus, like reviews [7].

However, considering the free style of updating personal opinions and the dynamic circumstances [1], Twitter users may not express their tendency using some sentiment lexicon directly. Their comments have strong timeliness, closely related to some realtime events; sometimes these tweets are very short because users are busy focusing on what is going on. For instance, there was a tweet “Milito!” during the UEFA Champion League Final 2010 in May, when Milito broke the deadlock. Though Milito is a player name which is literally neutral, given this situation, we can know this user shows his support to Internazionale, while conventional approaches [6], [8] cannot handle these cases properly. We name entities associated to an aspect as its supporting entities, by mentioning which will make positive contribution to this aspect.

We propose utilizing supporting entities to solve the problem in sentiment classification mentioned above. First, we utilize supporting entities as the major features in our classification to extract the sentiment information of tweets discussing the given topic. We then focus on calculating percentage of positive tweets related with the topic over time, namely, support rate, to track continuous realtime change of people’ attitudes, which have not been studied extensively. Contributions of this paper are summarized as follows:

1. Our approach based on supporting entities can significantly handle sentiment classification on tweets talking about realtime topics or events, especially the ones which do not include sentiment lexicon but show obvious sentiment tendency.
2. Our support rate estimation is based on realtime data collected from Twitter. It inherits the distinctive feature of realtimeness, which means the support rate

will be updated immediately when a new tweet talking about specific topic arrives

3. With the realtime support rate curve, we can clearly explain people's reactions to important events. In our estimation, we achieved an acceptable accuracy in sentiment classification.

The rest of the paper is organized as follows: in the next section we give a formal definition of the problem we address in this paper. Section 3 illustrates our approach of extracting sentiment information and estimating the realtime support rate using machine learning method Support Vector Machine (SVM) [6,9]. Experiments and evaluations will be presented in Section 4. The final section is devoted to discussion on prospective improvements of the present work.

2. Problem Setting

Before giving a formal definition of the problem we address in this paper, we first present several definitions.

Definition1 (Topics and Aspects) Topics in our research should be sentiment-oriented, which involve several aspects $\boldsymbol{a_1}, \boldsymbol{a_2}, \dots \boldsymbol{a_i}$ in the real world or sentiment concepts.

Take the area of sports as an example. For the topic of the soccer match between England and Germany in World Cup 2010, teams England and Germany can be regarded as two opposite aspects.

Definition2 (Supporting Entities) Aspect $\boldsymbol{a_i}$ may be accompanied by particular types of entities $\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, \dots \boldsymbol{e_{ij}}$, which are members of or closely associated to $\boldsymbol{a_i}$. By mentioning these entities in their tweets, there will be a large chance that people show their sentiment tendency toward $\boldsymbol{a_i}$.

With the example mentioned above, supporting entities to the aspect England should include the players and head coach of England and the positive events they are involved in, typically, the goals they have scored. In the case of aspect Germany, the situation is same

Definition3 (Sentiment Value) Given an aspect $\boldsymbol{a}$ of a specific topic and a tweet $\boldsymbol{t_i}$ talking about *a*, the *polarity* $\boldsymbol{s_i}$ of $\boldsymbol{t_i}$ with respect to *a* is an integer -1 or 1. Tweet $\boldsymbol{t_i}$ is said to be *positive* if the overall sentiment expressed in $\boldsymbol{t_i}$ is positive, with $\boldsymbol{s_i = 1}$ to denote positive emotions; while $\boldsymbol{t_i}$ is nonpositive if the overall sentiment expressed in $\boldsymbol{t_i}$ is neutral or negative, with $\boldsymbol{s_i = -1}$ to denote neutral and negative emotions. Table 1 shows examples of tweets during the game England vs. Germany and the game England vs. Slovenia in World Cup 2010.

However, our research excludes the usage of irony which in most cases, positives words are deployed to express negative feelings. Also tweets with mixed polarities are excluded from our discussion. Only tweets with apparent polarities are used in our training data.

Definition4 (Labeled Tweet/Unlabeled Tweet) Given an aspect *a* of a specific topic *p*, suppose that $\boldsymbol{t_i}$ is a tweet talking about *p* and $\boldsymbol{s_i}$ is the polarity of $\boldsymbol{t_i}$with respect to *a*. The pair $(\boldsymbol{t_i}, \boldsymbol{s_i})$ of $\boldsymbol{t_i}$ and $\boldsymbol{s_i}$ is called a *labeled tweet*. If $\boldsymbol{s_i}$ is not assigned to $\boldsymbol{t_i}$, it is called an *unlabeled tweet*

Definition5 (Support Rate) Given an aspect $\boldsymbol{a_i}$ of a topic *p*, and all the related labeled tweets $(\boldsymbol{t_i}, \boldsymbol{s_i})$ collected in the time period *T*, we define the *support rate* of the aspect $\boldsymbol{a_i}$ during *T* as the percentage of positive tweets $\boldsymbol{p_T}(\boldsymbol{a_i})$,

$$p_T(a_i) = \frac{\#of(\boldsymbol{t_i}, 1)}{\#of(\boldsymbol{t_i}, 1) + \#of(\boldsymbol{t_i}, -1)} * 100\% \qquad (\mathbf{1})$$

where in this formula, $\#of(\boldsymbol{t_i}, 1)$represents the number of positive tweets with respect to aspect a_i during *T*, while $\#of(\boldsymbol{t_i}, -1)$ denotes the number of nonpositive tweets, respectively.

Based on the definitions above, we now define the problem we try to address in this paper as follows:

Problem Definition (Realtime Sentiment Classification for Tweets) Given tweets collected continuously with specific topics, the task is to assign a proper sentiment value $\boldsymbol{s_i}$ to a unlabeled tweet $\boldsymbol{t_i}$ and update the support rate of current period $\boldsymbol{p_T}(\boldsymbol{a_i})$ immediately.

3. Approach

To explain our approach clearly, we consider soccer game as the target topic for the following discussion. It is important to know that the process of designing features is topic-dependent. For other domains, we need to modify features according to the different aspects and supporting entities. But required characterization of different

Table 1. Example tweets during World Cup 2010

Sentiment	Sentiment Value	Query word	Tweet
Positive	1	England	Cummm onnn #ENGLAND do us proud! http://tweetphoto.com/28693928
Neutral	-1	England	Watching the US vs Algeria and England vs Slovenia games with the team
Negative	-1	England	London bridge has fallin down .. So does england!!!

Table 2. Examples of preprocess

Before	After
Cummm onnn	Cum onn
Gooooooooal	Gooal
Yeahhhhhh	Yeahh

domains is shallow, so that the modification should be minor.

We have focused on soccer games in this study for two main reasons.

- The topic of soccer games is of considerable interest among the Twitter users, characterized both by large number of users discussing the games, as well as a substantial variance in their comments.
- The emotion of Twitter users are strongly influenced over the game. Posting simple tweets without sentiment lexicon is a common phenomenon when users are getting excited.

3.1 Preprocess

Twitter users post their messages via variety of devices, and the language used in a tweet is usually informal. This results in a high frequency of misspelling and full of repeated letters to express their emotions. In Table 1, the tweet “Cummm onnn #ENGLAND do us proud! http://tweetphoto.com/28693928” is a typical example. “Cummm onnn” should actually be “Come on”. In the tweets we collected, words with several repeated letters, like “Goooooooal” and “Yeahhhhhhh”, are very common, yet such repeated characters should be removed. Also, letters “h”, “w” and “l” occurred repeatedly in the tail of a word need to be removed. After preprocessing there will be no more than two consecutive occurrence of a letter. Table 2 shows some examples of preprocess.

Notice that, even after preprocess, there still exists misspellings, and this will cause a matching problem when we transform a tweet into vectors indicating the presence of features. In the case that follows, a non-strict matching function is needed, which we will discuss in the end of Section 3.2. Also, users often include links in their tweets; such kinds of URLs are often little correlation with sentiments, so we remove these links to reduce the calculation of feature matching in our approach.

3.2 Classification via Support Vector Machine

Support Vector Machine is a popular classification method. By preparing positive and nonpositive tweets as training data, it automatically produces a classifier that classifies tweets into two categories (positive and nonpositive). A particular type of Support Vector Machine, LibSVM2 [9] is used in our experiments.

Before defining features we designed for classification, we need to explain several elements which are related with the sentiment values.

- **Aspects:** As defined in Section 2, aspects in the topic of a football game are the names of the two teams. We send the team names to Twitter as query words. The number of occurrences and the position of the query words have high weight for deciding sentiment values. It is easy to understand the number of occurrences of a certain aspect has high weight, because the repeated words can be treated as a symbol that the author is getting excited. The position of the aspect in a tweet is taken into consideration under the assumption that during an exciting game, people would not express their opinion in the usual manner, like “I like X’s performance today.”(X means the query word) Instead of that, users may prefer to use a shorter and more straightforward sentence to deliver their feelings and in most cases, these sentences begin with the aspect or have the aspect in the tail. For instance, “Come on! X!”, “GoGoGo, X” and “X is awesome!”
- **Supporting entities:** As defined in Section 2, each aspect may be accompanied by particular types of entities. By mentioning these entities, people show their sentiment tendency, such as new coming events and people involved in these events. In the case of a football game, a goal can be treated as the most obvious event which will arouse people’s desire to express their support to one of the two teams, and the player who just scores will also be mentioned in the tweets. This results from the behaviors of the Twitter users when they get excited for the game, especially, when there is a goal or a beautiful shot. They will post their feelings to such an event and usually a player related with the event will be mentioned. There are a large amount of such examples in the tweets we have collected, like “GOAL! Milito breaks the deadlock! 1-0 to Inter!”, “Blimey, Inter being outplayed by Bayern, now 0-1 up with a lovely goal from Milito”, where Diego Milito is a famous forward in Inter. By mentioning his name, the audience show their support to Inter. Although there would be occurrences of supporting entities in negative contexts, positive contexts can be more often than negative ones. Therefore we can assume that occurrences of supporting entities such as player names have a positive contribution to the support rate of their team.
- **Numerical indicator:** Numbers in the tweets sometimes also contain sentiment information, especially in the tweets posted during sports competitions. In the case of a football game, we use the

score report as our numerical indicator, in the forms of "x-x", "x vs. x" or "x to x", caught by regular expression. Based on our observation, most of these score reports emerge after a goal which is an obviously positive event to the fans of the team just scored. Therefore we consider numerical indicators affect sentiment.

Utilizing these above elements, we can design the features for a football game. Table 3 shows twelve features used in our estimation.

Table 3. 12 Features used in classification

Feature	Type
Tweet Length	Integer
Number of occurrence of first aspect (SE)	Integer
Position of first occurrence of first aspect	Integer
Position of last occurrence of first aspect	Integer
Number of occurrence of second aspect (SE)	Integer
Position of first occurrence of second aspect	Integer
Position of last occurrence of second aspect	Integer
Number of occurrence of players of first team(SE)	Integer
Number of occurrence of players of second team(SE)	Integer
Occurrence of event symbol	Integer
Current scores of first team, if mentioned	Integer
Current scores of second team, if mentioned	Integer

Algorithm 1 describes the process of extracting feature values.

Algorithm. 1 Feature Value Extraction Algorithm

```
1. Obtain a new tweet $t$ discussing the topic;
2. Use regular expression to catch the current scores if mentioned. Else, mark the corresponding feature values as -1;
3. Split this tweet into a series of tokens $w_1, w_2, \ldots w_n$, denote n as the length of the tweet;
4. For each $w_i$ do:
    (a) For each aspect a do:
        1) compare $w_i$ with a using edit distance to check whether there is a match;
        2) check whether $w_i$ contains a;
        3) if the results of 1) or 2) is true add 1 to corresponding features and update the feature stands for the position of the aspect;
    (b) For each entity $e$ do:
        1) compare $w_i$ with $e$ using edit distance to check whether there is a match;
        2) check whether $w_i$ contains e;
        3) if the results of 1) or 2) is true add 1 to corresponding features or update the feature stands for event symbol;
End
```

Aspects here stand for two team names and supporting entities include the name list of the players and coaches, country names and event symbols like "Goal". We obtained these supporting entities from an ontology extracted from the website of fifa[1]. Tokens here can be a single word or a combination of several words; this depends on how to split the tweets. In our case, we use a blank character as the separator.

When a comparison is carried out between the current token and the aspect or supporting entity, because of casual language used in Twitter and frequent misspelling mentioned before, edit distance between two given phrases should be calculated. If it is less than one, presence of the current target word can be confirmed. In the example of Table 2, after preprocess we obtain a token "Gooal", and a comparison is carried out between the event symbol "Goal" and "Gooal". Since the chance of misspelling occurring in the first and last letter in a word is tiny, a higher weight is set to these letters while calculating the edit distance. In this case, after we set higher weight to "G" and "l", the distance turns out to be zero and the event symbol is confirmed.

After executing the feature value extraction process, the vector for the input data is obtained. By applying the classifier to the vector, a sentiment value of 1 or -1 will be sent to the support rate calculation program as the output.

4. Experiments and evaluations

In this section we use tweets collected during World Cup 2010 and UEFA Champion League Final 2010 to conduct our experiments. We extracted tweets over frequent intervals using the Twitter Search API [10], thereby ensuring we had the timestamp, author, tweet id and tweet text for our analysis. About 1.1 million tweets referring to 65 matches (WorldCup-64, Champion League-1) released from May to July. Our evaluations mainly focus on the following aspects:

- **Accuracy of support rate**: Percentage of correctly classified tweets;
- **Correlation between supporting entities and support rate**: Statistics on the occurrences of supporting entities.
- **Feasibility of realtime calculation**: Due to the time limitation of World Cup, we collect tweets first, then carry out off-line analysis. We need to evaluate whether our approach is fast enough to deliver results in realtime.

4.1 Training Data

Tweets collected during UEFA Champion League Final are used for training; we use different games for training and test. We collected 11,250 tweets during the game between Inter and Bayern. After removing retweets and duplicates, we randomly pick up 200 tweets from the rest

[1] http://www.fifa.com/worldcup/archive/southafrica2010/teams/team=43942/index.html

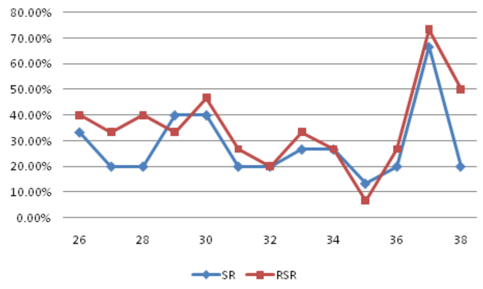

Fig. 1(a) SR and RSR for England

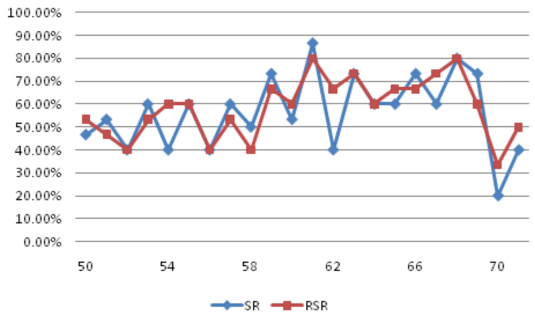

Fig. 2(a) SR and RSR for Brazil

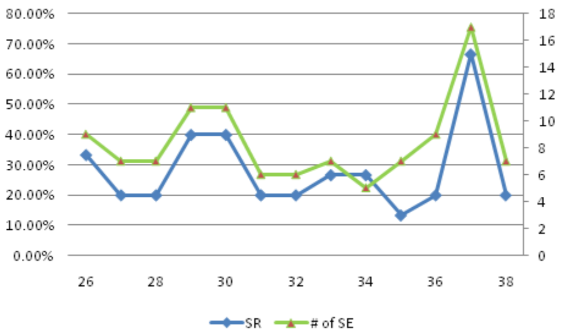

Fig. 1(b) SR and # of SE for England

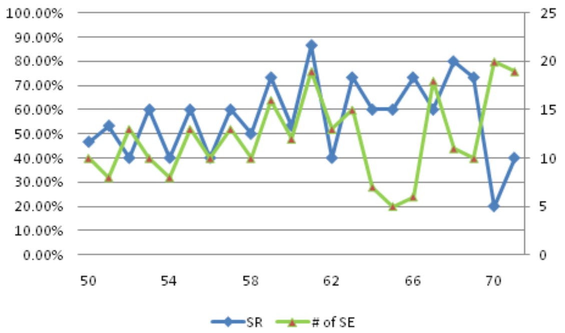

Fig. 2(b) SR and # of SE for Brazil

X-axis: time(minute), Left-Y-axis: Support rate, Right-Y-axis:# of occurrences of supporting entities

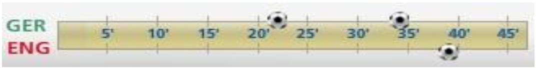

Fig. 1(C) The play by play records provided by FIFA.

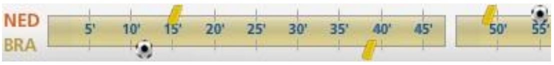

Fig. 2(C) The play by play records provided by FIFA.

and manually label them certain sentiment values. 93 out of 200 tweets are assigned to be positive.

4.2 Support Rate Analysis

In the experiments we conduct, we choose to show the results of two dramatic matches: England (1) vs. Germany (4) [2] and Brazil (1) vs. Netherlands (2) [3]. Fig.1 and Fig. 2 show the support rate calculated according to classification results (SR), the real support rate we manually checked(RSR) and the number of occurrences of supporting entities(#ofSE) of England(ENG Vs. GER) and Brazil(NED Vs. BRA) respectively:

The curves of SR in both of Fig.1(a) and Fig.2(a) are consistent with the curves of RSR, which means the results generated by the classifier trained with SVM are reasonable. After checking with the play by play record, we can confirm the rationality of our SR. In Fig.1, the support rate of England was relatively lower than Germany when Germany took one goal lead at the beginning. It experienced fluctuation for a few minutes which however, was followed by a sudden drop when Podolski made Germany's second goal at 32'. Yet, a significant rise of England's support rate occurred when they dramatically scored two goals in just one minute. But the high support rate of almost 70% did not last long as the second goal was disallowed and the following tweets could not be classified into a positive category since most of them were discussing the mistake made by the referees. In Fig.2, as the five championships owner with an early goal in the game, Brazil received a higher support rate form beginning. After the half time break (45'-60'), the average support rate was near 65%, and it kept at a high level with little fluctuation. However, the turning point of this game came at 70' (55', if excludes the half time break), Melo headed the goal in the direction of their own, Own goal! All the following tweets concentrated on this dramatic own goal, therefore, the support rate dropped to the bottom.

Regarding supporting entities (# of SEs), both of Fig.1(b) and Fig.2(b) which contain the distribution of

[2] http://www.fifa.com/worldcup/matches/round=249717/match=300061501/index.html

[3] http://www.fifa.com/worldcup/matches/round=249718/match=300061507/index.html

Table 4. Accuracy of the classification

Team	# of tweets	#of subset	#of tweets properly classified	Accuracy
England	1575	390	267	68.4%
Brazil	1125	330	232	70.3%

Table 5. Average/max/min classification time

Team	#of tweets	time cost (ms)	Average(ms)	Max	Min
England	390	3264	8	11	4
Brazil	2730	14753	5	7	3

supporting entities of England and Brazil over time show that basically the number of supporting entities has a positive correlation with the SR. As the most important feature in our approach, # of SEs largely determine the trends of SR, only one exception in the own goal point of the game of Between Brazil and Netherlands, although the occurrences of supporting entities is high ("Melo" is mentioned frequently), the support rate still drops to the bottom, we believe this is mainly results in the feature of numerical indicator played a negative role.

4.3 Accuracy Analysis

The reason that we did not show the support rate of the whole game is that manual classification of the tweets for the entire game is costly. We use a subset, which contain some goals during its duration, of all the tweets of a game instead.

Table 4 shows the accuracy of the experiments that we manually checked, by calculating the percentage of the tweets which classified into the correct category. We obtained an acceptable result of accuracy around 70%.

4.4 Verifying Realtimeness

In this part, we show a breakdown of execution time to check realtime requirements. Our cost model consists of two parts: time t_1 for feature extraction and time t_2 for classification. With the prepared classifier, time cost for classification is so tiny that can be ignored. The major time cost is in the process of feature extraction.

Table 5 shows the average, maximum and minimum time t_1 in our experiments. In the case of Brazil, we use all the tweets collected through the game to check the performance in the batch job with a large amount of tweets. In both of these two cases, the time cost for each tweet is less tan 0.01s, a satisfactory result as a realtime system.

5. Conclusion and Future Work

In this paper, we have shown how the supporting entities can be used in sentiment classification, specifically, in the classification of tweets with obvious sentiment tendency but lacking sentiment lexicon. Drawing on a soccer game, we collected one million tweets and classified these tweets into two categories to compute the support rate for the two competitors using SVM. Compared with the manually classified results, curves drawn according to the results of classifications based on supporting entities are reasonable. A satisfactory result returned in the process of verifying the feasibility for realtimeness.

This supporting entity based method can be extended to large panoply of topics with minor modification according to different subjects and their supporting entities. At a deeper level, this work of estimating support rate can yield an extremely powerful and accurate indicator based on social media.

References

[1] Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. "Earthquake Shakes Twitter Users: Realtime Event Detection by Social Sensors", WWW2010, April 26-30, 2010, Raleigh, North Carolina (2010).

[2] K. Dave, S. Lawrence and D. Pennock. "Mining the peanut gallery: opinion extraction and semantic classification of product reviews". (WWW2003).

[3] B. Pang, L. Lee and S. Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques". (EMNLP'2002)

[4] Kay-Yut Chen, Lesilie R. Fine and Bernardo A. Huberman. "Predicting the future". Information Systems Frontiers, 5(1):47-61, 2003.

[5] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. "Support Vector Learning for Interdependent and Structured Output Spaces", ICML, 2004

[6] Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs", Proc. Int.Conf. Weblogs and Social Media (ICWSM), 2007

[7] T. Joachims. "Text categorization with support vector machines". In Proc. ECML'98, pages 137–142, 1998.

[8] M. Sahami and S. Dumais and D. Heckerman and E. Horvitz. "A Bayesian Approach to Filtering Junk {E}-Mail". AAAI Technical Report WS-98-05, 1998

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[10] http://apiwiki.twitter.com/Twitter-API-Documentation