Classification of Tweets

Problem Statement

To automatically classify Tweets from Twitter into various genres based on predefined Wikipedia Categories.

Motivation

In microblogging services such as Twitter, the users may become overwhelmed by the raw data. In the past few years, Twitter has become a major social networking service with over 200 million tweets made every day. Although Twitter provides a list of most popular topics people tweet about known as Trending Topics in real time, it is often hard to understand what these trending topics are about. Therefore, it is important and necessary to classify these topics into general categories with high accuracy for better information retrieval. To address this problem, we classify Twitter Trending Topics into several categories such as sports, politics, technology, etc.

Classification Description

The user tweets has to be classified into various categories like News, Personal Message (1-1 Communication). These are further classified into subcategories. For e.g. News can be further classified into Sports, Politics, lifestyle etc. The user tweet further is checked to see whether it's a grievance or not.

The user tweet is classified into 8 main categories.

1. News : The news can be in various form that can be in form of :
    - Headlines: Coverage of breaking news and personal eyewitness accounts of news events.
    - Sport: Identifiable results of sporting events
    - Politics: Identifiable actions of politics events
    - Events: Live coverage
    - Weather: tweet reporting temperature

2. Personal Messages:
    These are user to user communication.
    These can be a comment on the other user tweet.

3. Status Updates:
    These are the messages of the users on the status updates.
    The status updates can be related to
    - Work: Reference to work related activity.
    - Location: Geographic references and location statements, including statements of

traveling, location change.

- Personal / Activity: What the user is doing at present or what he/she is thinking.
- Temporal , Mechanical , Physical , Automated etc.

4. Phatic:

   These are the views or opinion of the people. These can be:

   - Greetings: Statements of greetings to the broader Twitter community.
   - Broadcast: Textual soliloquy, monologue and undirected statements of opinion.
   - Fourth Wall , Unclassifiable etc.

5. Information Sharing:

   Twitter authors have a level of trust, credibility and perceived expertise in their role as a content provider for their followers' Twitter streams, and the inclusion of another user's message, URL.

   These can be random thoughts of the user.URL of any news for sharing information. The URL can contain video, pictures or any other links. The information can be about the user itself (only ME).

6. Pass Along:

   - RT: Retweeting other user comments
   - UGC: Any full–length or shortened URL which can be identified as the user's blog
   - Endorsement etc.

7. Controversial:

   - Query , Referral , Action ,Response etc.

8. Spam

Approach

Our task is to split the tweets into sentences, tag them, gather some feature information from each tweet, and use this information in machine learning classifiers that estimate the Twitter account (i.e., the twit) that produced a given tweet.

Classification can be done based on:

1) Naive Bayesian Model

The Naive Bayes classifier is closely related to the Binary Independence Model. Naive-Bayes classifier assumes that the features (an attributes which helps to identify that object) are not associated with each other.

2) Support Vector Machines

Based on a labeled dataset of tweets, a parser extracts present features in the text to create a vector. Once a collection of vectors is compiled, data is trained and tested to create a working model, which can then be evaluated to determine the effectiveness of the classifier.

Datasets:

Input Data is the static / real-time data consisting of the user tweets.

Final Deliverable:

It is the graph / chart categorizing the user tweets on various genres.

REFERENCES

http://firstmonday.org/ojs/index.php/fm/article/view/2745/2681
http://www.statsoft.com/Textbook/Naive-Bayes-Classifier
http://cucis.ece.northwestern.edu/publications/pdf/LeePal11.pdf
https://dev.twitter.com/rest/reference/get/search/tweets
http://iag.me/socialmedia/how-to-create-a-twitter-app-in-8-easy-steps/
http://www.research.rutgers.edu/~sofmac/paper/and2010/michelson-and2010.pdf
https://en.wikipedia.org/wiki/Category:Main_topic_classifications