Latent Attribute Inference of tweeple

Puneet Singh Ludu

Version: 0.1 Date: 19, Feb, 2014

Introduction

A few statistics:

- Social media is expanding exponentially
- 645,750,000 approx active users on twitter
- 135,000 new twitter signups everyday
- 58 million tweets per day
- 43 % of total users use phone for tweeting

This overwhelming data and information makes a compelling case for getting deep into analysis of user behaviours, trends, user profiling, interests of users, attribute inferences (for e.g. age, gender, ethnicity, brand loyalty, affinity for a particular business etc. of a user)

Source: http://www.statisticbrain.com/twitter-statistics/

Problem Statement

Many web technology players(Google, Bing, twitter etc.) made significant efforts to use social information with existing search and retrieval models for developing new applications such as user and post recommendation services.

In this context, a problem of significant interest is that of automatic user classification and profiling, i.e. mining values of various user attributes such as demographic characteristics (e.g., age,gender, ethnicity, origin), coarse and fine-grained interests (e.g. politics, soccer, Starbucks, Game of thrones [TV series]), stances on various issues (e.g. liberal, pro-choice), etc.

In this semester, I would target particularly following problems:

- Age classification (Young / Old)
- Gender classification (Male / Female)
- Political affiliations (Democrats / Republicans)
- If time permits, exercising political affiliations of Indian users (three classes AAP, BJP, INC)

[1] Pennacchiotti, Marco, and Ana-Maria Popescu. "Democrats, republicans and starbucks afficionados: user classification in twitter." ACM SIGKDD 2011.

Motivation (or demotivation)

> “Numerous papers have reported great success at inferring the political orientation of Twitter users. This paper has some unfortunate news to deliver: while past work has been sound and often methodologically novel, we have discovered that reported accuracies have been systemically overoptimistic due to the way in which validation datasets have been collected, reporting accuracy levels nearly 30% higher than can be expected in populations of general Twitter users.”

~ Cohen, Raviv, and Derek Ruths. "Classifying Political Orientation on Twitter: It’s Not Easy!." ICWSM **2013**.

This means:

- It is not an easy problem to classify a simple political scenario(2 party system), forget about complex multi party system scenarios like India.
- Current state-of-the-art (even those who reported 90% accuracies), may even perform as bad as 60-65% with normal users
- This problem extends beyond political affiliations.
- No one classifier to rule them all, one kind of features extracted for one group of users might not work on another group of users.
- While this news is certainly not good, at the same time this gives us opportunity to target issues in current state-of-the-art and improve results for normal twitter users.
- There has been hardly any attempt made on multi party system political affiliations, such as Indian polity.
- Even 80% accuracy on normal users can be significant contribution in this direction.

Previous works (State of the Art)

- Pennacchiotti et al. 2011, published a hybrid approach using ML and Graph based update.
- Faiyaz Al Zamalm et al. 2012, published an approach which extended previous works using homophily, Homophily is the tendency of individuals to associate and bond with similar others. They implemented this by categorizing twitter friends in following classes
    - most popular (number of followers)
    - least popular (number of followers)
    - closest (more mentions, more interaction)
- Raviv Cohen et al. 2013, published that all the previous approaches have over estimated accuracies. While there results are having astonishingly 90%+ accuracies, but actually on normal users the same approach would produce nearly 65% accuracy.

Data sets

Stage 1: Zamal_ICWSM_2012 dataset

Stage 2: Try to improve results on Raviv Cohen and Derek Ruths 2013 dataset(yet to obtain)

Stage 3: My own dataset on Indian users (targeting 200 twitter users)

Algorithm

1. Machine Learning models

- Profile features(twitter bio, name, twitter handle, website)
- Tweeting behaviour features(e.g. tweets per day)
- Linguistic features of users
  - If the user use word like “dude”, “lmao”, “yo!”, “mate”, “bhai” etc.
  - What kind of hashtag user uses (and sentiments towards that hashtag)
  - User-Topic modeling (state of the art used LDA)
    e.g. Democrats may have, on average, a higher probability of talking about social reforms, while Republicans may mention oil drilling more often
  - Sentiment Analysis (e.g. “Ronald Reagan” is generally viewed positively by Republicans and negatively by Democrats)
- Network Features (Homophily)
  - Friends (I follow they, they follow me), followers, and people I follow(e.g. friends to follower ratio user’s tendency towards producing vs consuming information on Twitter (Rao et al. 2010))
  - What kind of tweets, and by whom I Retweet

2. Later use “Graph based label update” algorithm to improve results obtained from machine learning.

Tools and Infrastructure

- Java,eclipse (to write code)
- Sentiment Analysis Library (lexicon and ML based)
- Twitter Garden Hose (using twitter4j)
- Rapidminer (Automating Machine Learning)
- If my laptop would not suffice for graph processing, I would use Apache Giraph.
- Rest of the stuff I would figure out as I move forward

Evaluation Metrics

- Precision, Recall, F-measure, Accuracies and kappa in various configurations
- Final 2 class confusion matrix, with accuracy
- 10 fold cross validation to compute statistical significance

Project Plan (things to do)

- Getting the data
- Keep on extracting features and running them through 10-fold SVM, recording results
- Finding ways to improve the same feature results on other datasets.
- Meanwhile, get my twitter friends(or random people) to fill up a survey, so that I can collect their data as well.
- Using graph based label update, finally to improve the results obtained from machine learning

Milestones

28 February - Dataset ready

28 February - Code to extract profile features ready

7 March - Report profile feature results

7 March - Code to extract tweeple behaviour feature and two linguistic features ready

20 March - Report results of individual and combined features

20 March - Code to extract Topics, Sentiments and Network features Ready

31 March - Report results from machine learning based classification

31 March - Code for graph analysis is ready

31 March - Data of Indian tweeples is ready

10 April - Buffer time for any kind of mismanagement, hurdles etc.

25 April - Time devoted to improve the results, using graph mining and other tweakings

26 April - Running my algorithm on Indian dataset

30 April - Reporting final results

Rest of the days till presentation is for buffer and improving the results

Thank You