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:

- Gender classification (Male / Female)
- Age classification (Young / Old)

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

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)

Data sets

Stage 1: Zamal_ICWSM_2012 dataset

Stage 2: My own dataset on Indian users (targeting 200-300 twitter users)

Algorithm

1. Machine Learning models
   - Profile features(twitter bio, name, twitter handle, website) [DONE]
   - Tweeting behaviour features(e.g. tweets per day) [Work in progress]
   - Linguistic features of users [Work in progress]
     - Currently I am using frequency of LIWC2007 classification of words as features
   - Network Features
     - I am planning to use verified users by classifying them into genres such as sports, news, entertainment, technology etc. and using them as feature.
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 [Done]

28 February - Code to extract profile features ready [Done]

7 March - Report profile feature results [Done]

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

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

Results

Preliminary results for **Gender classification** based on only profile based features: Name and screen name

Using: SVM, 10-fold validation on 299 user data

Code: https://github.com/puneetsl/tlassify-gender

Results of various runs: http://goo.gl/QaIKX8

accuracy: 90.31% +/- 5.65% (mikro: 90.30%)

	true female	true male	class precision
pred. female	151	23	86.78%
pred. male	6	119	95.20%
class recall	96.18%	83.80%	

Thank You