Motivation

What is it about?

- For online trend analysis, user twitter profiles are classified.
- The focus is on descriptive features, such as user-created micro-blogging content

Where ? Yahoo Labs

Why do we use it?

- This will help in automatic user profiling,
- in order to build better recommendation systems, and finding users ,expert on a topic,
- along with providing most-relevant tweets to information seeking users.

When ? 2010

How we do it?

- User **classification** is done using two mechanisms:
- Machine Learning model (Gradient Boost Model, 10 folds cross-validation and confusion matrix) : it uses user centric information and labeled data
- Graph based updating component : it uses social graph information (users friends on Twitter)

Dataset :

Gold standard followed, i.e. active users with 5 or more tweets
At least 50% of them in English

Pennacchiotti, M.,&Popescu, A. M. (2011, August).

Nidhi Agarwal nig42

Method

Data Modelling

The **ML model** has 4 predictors:

- **PROF (profile)** : username, location
- **Linguistic content:** extract proto words (using LDA technique)
- **Tweet behavior:** the average number of tweets and replies.
- **Social network features:** user replies, retweets and follows

The **graph based label update**

- it helps in finding the friends and followers of a user on Twitter.
- It corrects the error of ML components by inverting the classification label

- **Important**: Connection strength combined with network wide influence

$$w_{ij} = \frac{1}{2} \cdot \frac{|mentions_{ij}|}{\sum_{k \in F_i} |mentions_{ik}|} + \frac{1}{2} \cdot \frac{|ratioFolFriends_j|}{\sum_{k \in F_i} |ratioFolFriends_k|}$$

Experiment: Classify user using 3 characteristics

- **3 instances of architecture:**
- **I-** using only ML component
- **II-** using only graph-based updating component
- **III-**hybrid of the above two models

- **Political affiliation**: Democrats/Republicans
- Data from WeFollow, Twellow , 10338 users
- Best model : **Hybrid** with 80% accuracy level

- **Ethnicity**: Afro-American(AA)/ else
- Data: 3000 AA + 3000 non AA users
- Collect 909k friends for users
- Reduced to 508k after 5 tweets cut
- AA is the **most-represented** community
- Best model : **ML**

- **Starbucks** fans: 5000 +ve + 5000 −ve examples
- 1.9M friends → 981k after 5 tweets cut
- Best model : **Hybrid**

Pennacchiotti, M.,&Popescu, A. M. (2011, August).

Nidhi Agarwal nig42

Take-away

Loop-holes:

- Only class specific membership score is used. Follower data might provide more insight
- African-American is not a closed group, they connect to users of other ethnicities as well. Therefore, the classification results might not be accurate.

Inference from the paper:

- Boosting algorithms have over-fitting problem
- Topic based LDA are reliable
- Class specific topic models outperform generic topic models. Example- at Nov 2010 elections, tweets were party specific, there were get-out-the-vote messages and voting related discussions. Its easier to identify such tweets. They might even have hashtags.
- Its challenging to find user ethnicity. But still by using classification techniques, we can study opinions of a huge population. Such as African-American community
- Profile and linguistic information are most helpful features to find brand-related efficacy among users
- The profile feature is not good enough for classification purpose. This is because people upload celebrity pictures, and put address as "Wonderland"!

Learning:

- ML model is sufficient for user classification, and does not require graph-based updation.
- Each data-set requires different modelling methods. There is no one correct model.
- Automatic user classification in real time is feasible.