Statistical Learning Project

1st Milestone

Group 7 : Anil Keshwani - Eleftheria Tetoula Tsonga - Iason Tsardanidis - Mathieu Grasland - Félicie Bizeul

Research Title

Twitter Sentiment Analysis and Classification of Political Persuasion

---

Abstract

We scrape tweets related to politics targeting (1) accounts of individuals with a known political affiliation and (2) a set of hashtags identifying a political theme. We label those tweets according to political affiliation of the account holder (1) or via a manual labelling process (2) with a binary partisanship label. After preprocessing the tweets, we represent in a *word* feature space and construct predictive classification models using one, all or an ensemble of (penalised) regression, support vector machines and tree-based approaches.

NB Our outcome of *partisanship* is yet to be decided concretely. We could target party affiliation using tweets around the upcoming US election and current primary races. We would like however, to focus on environmentalist issues or sentiment analysis implementing on various social movements (veganism,feminism e.t.c) and analyze the public opinion.

Main research aim & framework

Our main research aim is to train a model which can predict binary political partisanship labels.

Our framework will be as follows:

1. Identify data sources from Twitter
   - Accounts from which to pull tweets
   - Hashtags, keywords or phrases for which we pull *all* tweets within a selected time window (possibly retrospective)
2. Write Twitter pipeline - possibly in a combination of R, Python and/or others (e.g. SQL)
   - Use Twitter API
   - Use NLTK; RVest in R
3. Provide labels for data
   - Provide binary labels for accounts which are assumed to have partisan opinions
   - Manually label tweets which are pulled according to hastags or key phrases
4. Data Cleaning and Preprocessing
   - Clean initial data

- Build representations amenable to being fed into model e.g. bag-of-words document representations

5. Statistical modelling for Classification
   - Logistic Regression (GLM); Regression Trees; Ensemble Approaches; SVM; maybe Neural Networks?
   - Build separate models using classified accounts versus classified tweets
   - If time: experiment with multi-level models to take account of clustering of tweets within people

---

Data source(s)

Twitter using a "twitterbot" leveraging Tweepy which allows scraping.

Keep in mind the limitations of Twitter API!

Once the idea for the project has been approved, we will create a list of accounts and hashtags on the basis of which to scrape tweets.

Data collection

- We will collect data via a twitterbot.
- Manual tweet labelling for tweets collected by targetting hashtags or phrases will require human input.
- Given a groupsize of five, we hope to be able to label ~1250 tweets manually
- Combining these with the tweets collected by targeting accounts with known affiliations - which can be expediently labelled - results in an arbitrarily large sample size (e.g. 10^6). Tweets collected in the latter manner have the caveats that we
  - we have to identify individuals' political persuasions (relatively low manual time commitment)
  - we are limited by the non-independence of tweets "within" accounts

Potential difficulties: - Converting tweets to numerical representations (word/text embeddings) well suited to use as input for modelling requires additional preprocessing/programming time commitment - *Data Labeling* of data since we cannot be sure and objective about the corrent identification of tweets. It is something that will be done probably manually with some package assistance maybe (e.g TextBlob)

How "heavy" will the dataset be?

The dataset is likely to be a CSV containing at least columns for username, time, location, tweet text and a label. Additionally, there will be the numerical representations of these tweets which will be word vectors. We are not yet sure on how large these data structures are likely to be (in terms of bytes).

Model & Methods

After creating our database containing various information of tweets and their content, we will establish a classification of those. To do this, we will use multiple statistical methods that we already know and experiment with new ones. We will try to explain a binary variable.

We will be able to use **logistic regression** models with and without penalties, which allow us to explain our binary Y variable, using explanatory variables (i.e. the word vector representations of our tweets). We can also use **SVM** techniques which are a set of supervised learning techniques intended to solve problems of discrimination and regression. Also, we will use tree classification methods like the **cart and chaid algorithms**.

We may also use other classification methods such as **KNN** and **Discriminant Factorial Analysis** and planning to try some ensemble classification methods: **RandomForest**,**GradientBoosting**,**XGBoosting**. Finally, we can compare all those methods according to the sensitivity and specificity rates, by constructing the ROC curves of these methods and by comparing the AUC. Last but not least, we are planning to use probably a combination of the best of them according to their performance using K-fold Cross Validation and a Grid-Search or Random-Search searching for the optimal hyper-parameters of these models.

Software/Hardware Toolkit

- **R**
- **Python**
- Possibly AWS depending on computational load...

Project Timeline

1) Project Orientation (politics, environmental stuff e.t.c)
2) Data gathering
3) Data Pre-processing (transform text to vectors)
4) Model Training/Evaluation
5) Analysis of the Results
6) Presentation

References

- KLEINBAUM, D., KLEIN, M., *Logistic regression : a self-learning text (3rd ed.)*, Springer, 2010
- Agresti A., *Categorical Data Analysis (3rd ed.)*, WILEY & Sons, 2012
- Hosmer D.W., Lemeshow S., *Applied Logistic Regression (3rd ed.)*, WILEY & Sons, 2013
- https://xgboost.readthedocs.io/en/latest/
- James et al. (2013) An Introduction to Statistical Learning with Applications in R. 8th Edition. Springer; "ISLR"
- Murphy K. (2012) Machine Learning A Probabilistic Perspective
- Hastie T., Tibshirani R. and Friedman J. (2009) The Elements of Statistical Learning. Springer; "ESL"
- Steven Bird, Ewan Klein and Edward Loper (2009) Natural Language Processing with Python; "The NLTK Book"
