Statistical Learning Project

2nd Milestone

Group 7: Mathieu Grasland, Felicie Bizeul, Eleftheria Tetoula Tsonga, Iason Tsardanidis and Anil Keshwani

Research Title

Classifying Climate Change Attitudes with Twitter

Abstract update

We scrape tweets related to climate change selecting only those authored by individuals with a known position on climate action and containing one or more keyword strings. We labelled the tweets according to the known positions of the account holders using a binary label reflecting "believer" or "denier" status or pro-/anti-climate action. After data cleaning, we represent tweets in a word vector space and construct predictive classification models using one, all or an ensemble of (penalised) regression, support vector machines, naive Bayes, tree-based approaches and neural networks.

Main research aim & framework update

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

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)
    - We used two scrapers for data collection: twint and the Old Tweets Scraper Using Python
3. Provide labels for data
    - We label tweets according to the known position of their author (username)
4. Data Cleaning and Preprocessing
    - Clean initial data

- Tokenisation, Stemming and optionally other pre-processing (e.g. stop word removal)
- Word embedding through neural networks or simpler techniques (e.g. dimensionality reduction)

5. Statistical modelling for Classification
    - Logistic Regression; Naive Bayes; Regression Trees; Ensemble Approaches; SVM; Neural Networks
    - If time: experiment with multi-level models to take account of clustering of tweets within people

---

Data collection & source(s)

We created a list of user accounts for which we could also identify clear attitudes: believers or deniers of climate change. We scraped tweets from these accounts filtering only for the relevant ones by requiring tweets to contain one of the substrings "climate" or "global warming". A balanced dataset of ~10k tweets was compiled and is submitted as a CSV with the core fields:

- Twitter Username
- Timestamp
- Tweet
- Label (outcome; indicates attitude)

We intend to pre-process and model these tweets (see Models and Methods section) to build a classifer with high attitude prediction accuracy on a subset of our collection (i.e. a test set).

Models & Methods update

Our modelling will consist of two substantive stages after basic data cleaning: pre-processing and modelling.

Pre-processing

We will represent embed tweets in a word vector space using either neural network approaches which do this implicitly through optimisation of hidden layer weights (e.g. word2vec) or through other means as listed in the references (e.g. dimensionality reduction). We would also like to attempt to use more elaborate approaches which account for polysemy or use fragments of words (e.g. BERT) time permitting.

Depending on the ability of an approach in this initial step to handle these issues, we may or may not perform basic pre-processing (tokenisation, stemming and/or stop word removal) as required to *normalise* the natural language data.

Modelling

We will experiment with a variety of modelling techniques including:

- **Logistic regression** models with and without penalties
- **Support Vector Machines**

- Tree classification methods such as the **CART and CHAID Algorithms**

We may also use other classification methods such as **KNN** and **Discriminant Factorial Analysis** and plan to try some ensemble classification methods including:

- **RandomForest**
- **GradientBoosting**
- **XGBoosting**

Model Evaluation

We will compare methods according to their sensitivity and specificity, in practice by constructing the ROC curves for them and comparing their AUCs. We are planning to use the best method (or a combination of the best methods) with hyperparameters tuned via K-fold Cross Validation and a Grid or Random-Search.

Software/Hardware Toolkit update

- **R**
- **Python**
- Possibly cloud computing (e.g. AWS) depending on computational load

---

Problems so far...

We were unable to collect data quickly enough or far enough into the past using the API. We circumvented the rate limits and limited timespan of the archive when using Twitter's Standard (free) API by using twint and the Old Tweets Scraper Using Python.

Project Timeline update

1) Project Orientation
2) Data gathering (performed up to a sample size of ~10k as of this milestone)
3) Data Pre-processing (transform text to vectors) - Until May 24th
4) Model Training/Evaluation - Until June 14th (allows time to attempt different pre-processing, i.e. word embedding techniques)
5) Analysis of the Results - Until June 21st
6) Presentation - TBC

References update

Word Embedding

- Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. "Efficient Non-Parametric Estimation of Multiple Embeddings Per Word in Vector Space." http://arxiv.org/abs/1504.06654.
- Terry Ruas, William Grosky, Akiko Aizawa, Multi-sense embeddings through a word sense disambiguation process, Expert Systems with Applications, Volume 136, 2019, Pages 288-303, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.06.026.
- Li, Jiwei, and Dan Jurafsky. 2015. "Do Multi-Sense Embeddings Improve Natural Language Understanding?" http://arxiv.org/abs/1506.01070.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." http://arxiv.org/abs/1310.4546.
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013) Efficient Estimation of Word Representations in Vector Space http://arxiv.org/abs/1301.3781 - This is the original word2vec paper
- Lebret, Remi, and Ronan Collobert. 2013. "Word Emdeddings Through Hellinger Pca." http://arxiv.org/abs/1312.5542. - This paper describes an alternative means of constructing embeddings using dimensionality reduction
- Globerson et al. (2007) Euclidean Embedding of Co-occurrence Data. Journal of Machine Learning Research 8 2265-2295

Modelling

- 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
- James et al. (2013) An Introduction to Statistical Learning with Applications in R. 8th Edition. Springer
- Murphy K. (2012) Machine Learning A Probabilistic Perspective
- Hastie T., Tibshirani R. and Friedman J. (2009) The Elements of Statistical Learning. Springer

Software

- https://xgboost.readthedocs.io/en/latest/
- Steven Bird, Ewan Klein and Edward Loper (2009) Natural Language Processing with Python

---