Project Overview – Information Quality in Social Networks

# Initial Stage – Information Procurement

The training data for the future classifier is being collected from social media platforms. For the purpose of this study Twitter is being used, due to its ease of access and extensive API. The which was collected is being centered around the topic of E-Commerce and the organizations active in this field. Among those are **Amazon, Alibaba** and **Groupon**.

Using a wrapper written in Python around the Twitter API, allows for tapping to the live Twitter stream (reduced to 1% of its overall volume). The tweets are being collected according to a given search parameter. The names of the above-mentioned Companies were used to collect relevant Tweets.

# Second Stage – Observing and Analyzing the Data

The Program I wrote to collect tweets allows for filtering them according to language of the content and omitting tweets containing certain **Stop-Words**. The desired information should provide insights and novel data. Trial and error showed certain Stop-Words common among Non-News tweets. The terms are: **gift, giftcard, giveaway** etc. Further types of tweets observed:

* **Clickbait** – contains catchy titles or phrases and references to current topic without containing any news but rather a link, which leads the user usually to a commercial.
* **Hashtag Spam** – non-news tweets, usually of online advertising. Hashtags of **trending topics** are added to the message to get a wider exposure, despite having nothing to do with the trending topic. Example: adding **#Trump** to a commercial
* **Complaints** – users express their discontent with the platform. Bad customer service, late deliveries or general disappointment with the product or service provided

In the future, these could be defined as separate subclasses, but for the purpose of this initial analysis they are all being grouped into **Not-News.**

# Third Stage – Building Feature Sets

Features to be used later as input data for training the different classifiers. The features could be divided into two categories:

* **Message related** – describes different characteristics of the message. Such as words and symbols contained. Ratios of letter types and other unique elements such as hyperlinks, hashtags and reference to other users
* **User related** – Details about the users account such as age, homepage, number or friends and followers

# Fourth Stage – Labeling and Training

All the collected tweets have to be labeled according to the two classes: News and Not-News. Afterwards a classifier could be trained and tested. I plan on using several types of classifiers and tallying their classifications into a vote-classifier, which would return a classification and a confidence level (how many of the classifiers voted for the elected class).

The Algorithms which I will try to implement (not final list):

* Clustering
  + K – Nearest Neighbors
  + K – Means
* Support Vector Machines
  + hard margin
  + soft margins - cost function for outliers
  + different kernels
* Neural Networks
  + 1 Layer
  + Convolutional