# Project Proposal – Online News Popularity

# Sonya Tahir

We have entered into the Information Age which has brought with it the Digital Revolution. People are becoming more and more comfortable with using computers and mobile phones in their everyday life for more than their traditional uses. They are being used to watch videos, listen to music, read books, do shopping, make financial transactions, plan and make travel arrangements, connect with friends and family, learn new skills, etc. This shift has come about due to accessibility of such technology to everybody as well as the convenience that they provide to people.

I want to focus on one particular use of the internet and the digital age and that is online news. Traditional news channels now maintain their websites and update the latest news online. Many other websites have also sprung up containing news related to a wide variety of areas and topics. The news websites have taken on the feel of a magazine that is always up-to-date and that can bring to the reader the news from his area of interest. “Past studies in this field have shown that there are 9 socio-technical advantages that have determined the adoption and use of online news: no costs, multitasking, more news choices, in-depth and background information, 24/7 updates, customization, ability to discuss the news with peers, the existence of different viewpoints, and the opportunity to ‘talk back to the media’.

Out of the 9 attributes, immediacy seemed to be the main reason for online news adoption. 70% of online news users had visited news sites a few times a day, while 47% of them would go to the internet first if they found out something interesting had happened.” (Online Journalism Blog, n.d.)

The availability of the latest news whenever the reader wishes is the most attractive reason for the adoption of online news. The reader can customize the topics of his interest to be shown related articles only. Websites also allow their articles to be shared on Facebook and Twitter due to which the reader can share articles with friends and family who may be interested in reading the same article. The sharing of articles results in other people reading the news even if they haven’t subscribed to the news website. More readers might visit the webpage due to the sharing of articles and this generates traffic to the website. The websites with high traffic are attractive to advertisers. All websites are interested in increasing webpage traffic as to attract advertisers and increase their revenue. For news websites, it is important to know which articles are the most popular so that they may sell ad space on the same page. They may even price it highly on an article that is expected to be popular. The number of page views and the number of page shares are important indicators of popularity of an article.

For the purpose of this project, I want to analyze what attributes of an article result in making it more popular. The research question is: Can we predict in advance which article will be more popular based on some characteristics of the article? In other words, what are the number of shares expected for each new article? To analyze this, I will be looking at the data from Mashable which is “a leading global media company that informs, inspires and entertains the digital generation. It has an audience of 45 million monthly unique visitors and 27 million social followers.” (Mashable, n.d.) The data was collected on May 31, 2015 and summarizes a set of characteristics of approximately 40,000 articles published by Mashable in a period of two years. The data has 58 predictive attributes and 1 target field which is the number of times the article was shared on social media. (UCI, n.d.)

I will use R for the analysis of this dataset and to build a model to predict the number of shares based on the characteristics of the article. Multiple linear regression will be a good choice for this problem since the target variable is numeric. Binning the target variable into a categorical variable is also an option after which I can try models such as logistic regression or decision trees. T-tests and F-test will be used to test for the significance of the variables and the model. The normality assumptions for the regression model will be tested using graphical methods as well as Shapiro Wilk’s test. Constant variance of errors will be tested using ncvTest. Cross validation techniques such as LOOCV and K-fold cross validation will be used to test the performance of the model.

The R packages required for this analysis include leaps, boot, car, mgcv, gamclass, glmnet. I will also use ggplot2 for visualizations. Other libraries may be used as required while the analysis is underway.

# References

“Why do people read online news?” Online Journalism Blog <http://onlinejournalismblog.com/2010/04/27/why-do-people-read-online-news-research-summary/> April 27, 2010

Mashable <http://mashable.com/about/>

# Dataset

“Online News Popularity Dataset” UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>