**Foundations of Data Science**

**Capstone Ideas v.2**

**Mike Chiem**

**1. What is the problem you want to solve?**

Online digital news publishing has gone through many cycles to improve readership and brand awareness of the publisher’s website. As more competition and more offerings become available online, viewership is much more fragmented amongst the various content providers. News providers primarily focus on increasing viewership by pursuing a breaking news cycle for their content creation and distribution strategy - which means quickly publishing urgent content (e.g. police chase) is prized over aspects such as content quality. However, these attentions are fleeting and do not generate return visits consistently. Furthermore, as more and more news publishers jump on this trend, users are also being inundated with constant breaking news alerts; the net effect could be a diminishing return on the overall breaking news cycle.

At the same time, these news publishers are not paying as much attention to other categories of news. For example, investigative journalism and lifestyle-based content (arts, theatres) require more time to write and research. These articles are available on the site, but are less focused and less prioritized due to the breaking news cycle strategy. These contents are an untapped source of revenue creations, but do not get the appropriate content strategy priorities. If we can find insights into these untapped contents and understand what drives its popularity, then we could potentially recommend new ways of content creation that aligns to users’ tastes.

My analysis will try to predict the popularity of an article by taking a broad approach to analyzing aspects of its content, primarily focused on segmenting individual paragraphs and looking at their isolated sentiment analysis rating. For the purpose of this project, I will define a content’s popularity by "the amount of shares" it gets on social media platforms such as Facebook, Twitter, etc. I believe this analysis could potentially impact the way editors write the news, giving them insights on how to craft a news article that increases its chance of being popular and shared amongst social platforms.

**2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

My clients will be managers in my workplace that are trying to improve ad revenue sales by focusing on viewership and recirculation. Page views as a measure of viewership and content popularity is an important determinant of advertising dollars and advertisement rates charged by websites such as ours. This analysis for feature significance, whether it is the title, the paragraphs or other combination of features, would give editors insight on how to best create content that is consumed by our audience at a greater rate. Once an article is shared on social media platforms, there are follow-on effects that generate even more traffic through referrals[[1]](#footnote-0).

Furthermore, this process can be iterative and transformative. From a strategic perspective, it will help the product team create a better overall news product. Instead of throwing resources at JUST the breaking news category, we can take a macro view of all the contents and see which features drive the content’s popularity. Using this approach in an iterative product release cycle can generate audience lift across the spectrum of content types, and thus diversify our product portfolio.

**3. What data are you going to use for this? How will you acquire this data?**

A sample dataset from Mashable.com will be used for the analysis. This is a good dataset because it does not have a breaking news product feature and so the full dataset can be used as a testbed for the idea. Although it does not fully share the news paradigm, it does have content written for different purposes (lifestyles, news, etc.) and represents a broad portfolio of content types that is similar to my workplace. Mashable is made available via UCI dataset repository: <https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

4. **In brief, outline your approach to solving this problem (knowing that this might change later).**

The Mashable.com dataset was used in a previous academic research to predict the popularity of an article before it gets published. In the research the authors used, amongst other features, a combination of sentiment analysis as well as topic-modeling (LDA closeness to 5 topics) to assess the predictability of the dataset. What they found was that given a threshold of 1400 shares to indicate popularity (vs. not popular), they were able to get an accuracy of 67% accuracy and .73 of AUC using Random Forest.

For my purpose, I will primarily focused on the following approach:

1. Feature engineering technique using parsing and tokenizing the content will be performed on each article’s paragraphs. Sentiment scoring for positive/neutral/negative will be captured on new feature/column that represents their respective paragraph. A separate sentiment analysis score will also be captured for the entire article.
2. For sentiment analysis, I will be using the Stanford’s Core NLP package: <https://cran.r-project.org/web/packages/coreNLP/coreNLP.pdf>

Once I have done feature engineering to capture sentiment rating of individual components of content (such as title vs. 1st nth paragraphs), I will:

1. Perform an iterative step-wise feature selection process to get at an optimal linear regression model. The start of the analysis will be based on this regression equation:

***Popularity (# of shares) = TitleSentiment + Paragraph1Sentiment .. + Paragraph2Sentiment + ParagraphTop50%Sentiment + ArticleCategory + ArticleDayOfWeekPublished***

1. If the model has a high degree of predictability, I will use it to predict against a test data set.
2. Using similar step-wise feature selection techniques, I will apply this regression analysis using the CART model.
3. I will take the result and compare the linear regression to the CART to see if a more predictive model can be generated.
4. Finally, I will repeat steps 3 and 4 with the Random Forest model and compare which type of analysis predicted better.

The business applicability of any insights generated will also depend on ease of implementation. In my workplace, the goal is to generate uplift in popularity, however small that may be. Therefore, It will be helpful to get focused insights by finding a small set of significant features that contributes to the predictability of a content's popularity. To that end, I will employ a logistic regression approach to predict the dependent variable (social shares) based on a low sharing threshold, say anything above the 1st Quantile (roughly 946 shares) to represent an article's popularity as 1, or 0 when below 946 shares. I will use the ROC AUC comparison, precision, recall, and F-score measure technique to fine-tune the model based on this small popularity threshold with the intention of maximizing the AUC while minimizing the amount of features used.

5. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

Primary deliverables will be an executive summary deck, the code, along with a written document highlighting steps taken to analyze the dataset. Furthermore, there will be a section on next steps to make it more applicable to my work, along with areas to improve upon and any additional refinements I deemed to obtain a better outcome.

1. http://www.forbes.com/sites/jaysondemers/2015/02/03/social-media-now-drives-31-of-all-referral-traffic/#33e4e2a21aee [↑](#footnote-ref-0)