What Makes a Good Title in Social Media? Using Language to Elicit Positive Response.

Joseph C. Somody, Martin Müller, Puspanantha Shanmuganathan, Amin Karbasi, Gábor Bartók, and Andreas Krause

Department of Computer Science, Swiss Federal Institute of Technology, Zurich, Switzerland

Introduction

- A title influences a reader's judgment of the work it describes, but what makes one title better than another?
- By analysing data from Reddit, where content can be rated either positively ("upvoted") or negatively ("downvoted"), the factors that influence a good title were studied.
- Repeatedly reposted content, after normalising over variables of non-interest, demonstrated the relative efficacy of different titles.
- Various language models were fit to the normalised data to determine the effect of particular title characteristics (for example, the use of community-specific keywords and the title's linguistic complexity) on the popularity of the resubmission.

Methodology

Data

- **Raw data were acquired from the Stanford Network Analysis Project (SNAP).**
- **The data were normalised by the average score of posts, depending on the particular posting time and subreddit.**

Community Model

- **The community model was built to account for popularity effects (post's prior exposure, time since previous submission, inherent popularity, etc.).**
- **The model's parameters were determined by training over all (re)submissions for each unique post.**
- **The targets for the language model were taken as the differences between the actual scores and the community model's predictions.**

Good/Bad Words

- **For each community of interest, two topic corpora of titles whose residuals were sufficiently above/below a given threshold were generated.**
- **A labelled latent Dirichlet allocation (LDA) model was trained and used to identify words that inflate/deflate a post's score: the good/bad words.**

Title Complexity

- **Each post's title was parsed into parts of speech.**
- **A linear model was fit to approximate the residual scores of titles based on the quantities of each present part of speech.**

Language Model

- **A language model was designed to combine the submodels used for good/bad words and title complexity with a further two title attributes: two metrics for title length.**
- **This model was fit to the training data to properly weight each submodel.**

Validation

- **The validation data (~1% of posts removed from the training data) were fed into the community and language models to calculate their predicted scores.**
- **The posts' actual scores being known, this project's overall performance was calculated.**

Results

$$\hat{A}_{h,n} = \beta_h + \phi_h \exp\left\{-\sum_{i=1}^{n-1} \frac{1}{\Delta_{i,n}^{h}} \cdot \left(\delta(c_{h,i} \neq c_{h,n})\lambda_{c_{h,i}} + \delta(c_{h,i} = c_{h,n})\lambda'_{c_{h,i}}\right) \cdot A_{h,i}\right\}$$

Equation 1 — The equation used for the community model, where A is the score, β is the inherent popularity, φ is the resubmission decay coefficient, Δ is the time difference, δ is the Kronecker delta, c is the subreddit, λ and λ' are balancing parameters, h stands for the unique content, and n stands for the resubmission number.

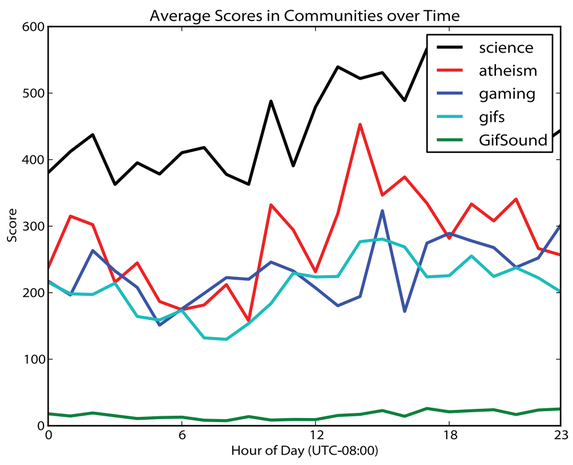

Figure 1 — Graph displaying the average score of submissions across various subreddits at different times of day, an example of an attribute that was normalised for.

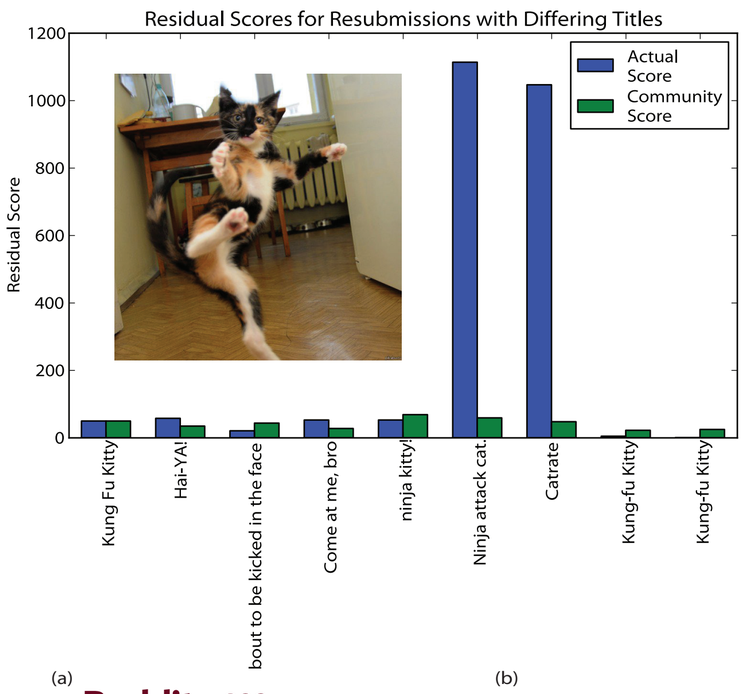

Figure 2 — Graph displaying both the actual attained score and predicted community score of an example image, which was submitted eight times with different titles. Note the sixth and seventh submissions were exceptionally well-received compared to the predictions made by the community model.

Figure 3 — Visualisations of the words most highly associated with (a) an inflated score in the whole of Reddit, (b) a deflated score in the r/atheism subreddit, and (c) an inflated score in the r/gaming subreddit. The size of the word is representative of the strength of the association.

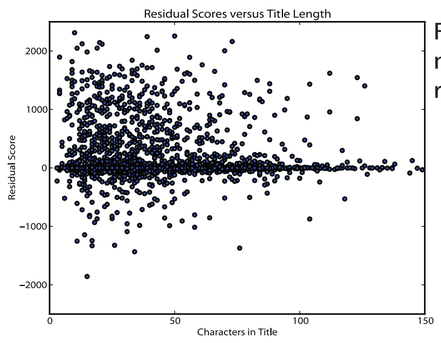

Figure 4 — Scatter plot displaying the relationship between the number of characters in the title of a post submitted to the r/gaming subreddit and the residual score of the submission.

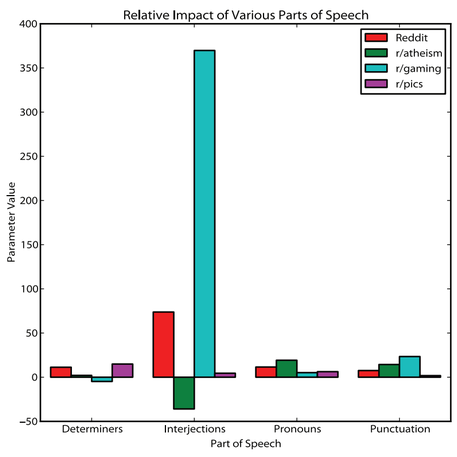

Figure 5 — Graph displaying select parameters for the title complexity model from four communities. Different subreddits have fundamentally different part-of-speech propensities, which helps to predict the success of content, given a title.

This project's performance was evaluated using a 1000-element subset of the raw dataset, the validation data, which was entirely disjoint from the training data. After the removal of outliers, it was found that 70% of the validation set data had an absolute error of less than 250. The median absolute error was 155.

References

Lakkaraju, H., McAuley, J. & Leskovec, J. (2013). *What's in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media.* Paper presented at the International AAAI Conference on Weblogs and Social Media, Boston, USA.