Understanding the Effects of Narrative Trajectories on the Popularity of Online News Articles

Raiyan Abdul Baten
University of Rochester
Rochester, NY
rbaten@ur.rochester.edu

ABSTRACT

This project illustrates the effects of narrative trajectories on the popularity of online news articles. A collection of 39,644 articles was analyzed. The emotional tones of each of their sentences were extracted using the IBM Watson Tone Analyzer. The trajectories of the tones were clustered using the DBScan algorithm. Statistical tests revealed that clusters of articles that are confident, have seesaw shapes in their joyfulness and end with a positive note enjoy a significantly higher number of shares ($p < 0.5$) than articles that have flatter trajectories.

KEYWORDS

Data mining, tone analysis, narrative trajectory, clustering

ACM Reference Format:
Raiyan Abdul Baten. 2018. Understanding the Effects of Narrative Trajectories on the Popularity of Online News Articles. In *Proceedings of Data Mining (CSC 440).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In this project, I aimed to explore how the narrative or plot trajectories can affect the popularity of online news articles. The ability to predict such popularity can potentially add value to content creators, providers, advertisers and even politicians. Therefore, it is of no surprise that the topic has attracted widespread attention in literature [1, 2]. Researchers have previously employed various machine learning techniques to predict the popularity of news articles, using features extracted from the contents. It has been found that Random Forest models have a superior predictive ability compared to other classical approaches [4].

Contrary to previous approaches, this project exclusively focuses on 'narrative trajectories' of online news articles to investigate the popularity information that lie therein. Narrative trajectories have previously been studied in the domains of English stories and public speeches. For instance, computationally extracted plot trajectories have been shown to hold predictive information about a story's genre [11], and the audience responses to a TED Talk [14]. However, whether or not narrative trajectories hold any predictive information about online news articles' popularity has not been reported in literature—a gap this project aims to fulfill.

I analyzed a large dataset of 39644 news articles, and extracted their sentence-wise tone information using the IBM Watson Tone Analyzer API [6]. I processed the extracted tones to bring the information of variable-length articles to a uniform signal length. Density-based clustering of the processed trajectory signals revealed typical narrative shapes that the articles display. Statistical tests on the clusters showed that articles that have confident wordings, go through seesaw variations in their joyfulness emotion and finish on a joyful positive tone are received significantly more favorably by the readers.

In summary, the contributions made in this project are as follows:

(1) Demonstrating the effects of narrative trajectories on the popularity of online news articles
(2) Using a large dataset ($N = 39644$) in the analysis to ensure high confidence in the validity of the claims. The IBM Watson generated tone information are included as part of the augmented dataset in the project.

2 BACKGROUND

2.1 Popularity of Online News Articles

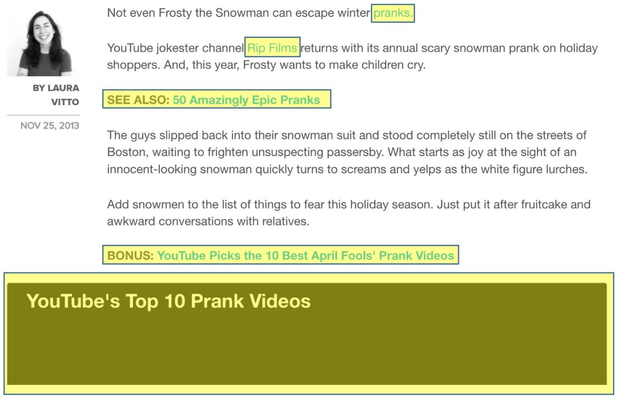
Predicting the popularity of online news articles has been an important topic for research, due to its vast array of high-impact applications. The popularity is often measured by social media markers: the number of likes, shares, sentiment of comments, speed of propagation through social network, so on and so forth [1, 2, 4]. To make the predictions, researchers previously employed features generated from the articles such as the number of words, the number of photos and videos, which day the article got published, etc. For instance, Petrovic et al. [8] predicted the number of re-tweets using features related with the tweet contents, such as the number of hashtags, mentions, URLs, length etc. Machine learning models such as Random forests have been found promising in predicting the popularity of online news articles [4, 10, 12]. However, to the best of my knowledge, none of the previous work in this domain investigated the connection between narrative trajectories and the articles' popularity, leaving a scope for contribution.

2.2 Narrative Trajectory Analysis

Narrative or plot trajectories of English stories have been widely explored in literature, and their implications are well documented. For example, it has been shown that the emotional arcs of English stories are dominated by six basic shapes [7, 9]. Samothrakis et al. [11] extracted the trajectories of 6 basic emotions [3] from

CSC 440, Fall 2018, Rochester, NY, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4



Not even Frosty the Snowman can escape winter pranks.

YouTube jokester channel Rip Films returns with its annual scary snowman prank on holiday shoppers. And, this year, Frosty wants to make children cry.

BY LAURA VITTO

NOV 25, 2013

SEE ALSO: 50 Amazingly Epic Pranks

The guys slipped back into their snowman suit and stood completely still on the streets of Boston, waiting to frighten unsuspecting passersby. What starts as joy at the sight of an innocent-looking snowman quickly turns to screams and yelps as the white figure lurches.

Add snowmen to the list of things to fear this holiday season. Just put it after fruitcake and awkward conversations with relatives.

BONUS: YouTube Picks the 10 Best April Fools' Prank Videos

YouTube's Top 10 Prank Videos

Figure 1: An example of pre-processing the original web articles. The yellow boxes mark elements that needed processing, including hyperlinks, ads, photographs, videos, among other elements.

the fictions of Project Gutenberg, using the WordNet Affect Lexicon [13]. They showed that such trajectories can help predict the stories' genre [11]. Recently, Iftekhar el al. [14] analyzed the transcripts of more than 2000 TED Talks, and showed that the narrative trajectories of the talks hold predictive information about the audience ratings. For instance, it has been shown that talks that have flat joyfulness tones are found to be 'longwinded' by the audience, while talks that have a hill shaped trajectory are found to be more 'beautiful'. However, effects of online news articles' narrative trajectories on their popularity have not been explored in literature, a gap this project addresses.

2.3 IBM Watson Tone Analyzer

I used the IBM Watson Tone Analyzer API [5, 6] for generating the narrative trajectories. The API takes a body of text, and returns 13 scores per sentence, each within a range of 0 to 1. The scores are:

(1) **Emotion scores:** anger, disgust, fear, joy, and sadness
(2) **Social scores:** openness, conscientiousness, extraversion, emotional range, and agreeableness
(3) **Language scores:** analytical, confidence, and tentative

IBM uses n-gram features, lexical features, person-based features, dialogue-specific features etc. to generate these scores [5].

3 DATASET CONSTRUCTION

The dataset I based this project on comes from UC Irvine Machine Learning Repository[1], and deals with articles published in `www.mashable.com`. The dataset comes with 58 predictive features generated from 39644 articles. The number of shares achieved by the articles are provided as the target attribute. The URLs to the original articles are also included in the dataset.

However, while the 58 given features capture various summary information of the articles (e.g., number of words in the content, number of images and videos), they do not provide or capture any narrative trajectory information. This made it necessary for me to scrape the original articles for analysis, as I explain in the sequel.

Algorithm 1 Constructing the Narrative Trajectories

```
Input: Array of sentence-wise scores, $S$ for an article
Output: Narrative Trajectory $T[n]$ for that article
  procedure BUILD_TRAJECTORY(S)
      Filter: Apply averaging filter on $S$ to get $S_{\text{smooth}}$
      Crop: Remove the boundary effects to form $S_{\text{crop}}$
      Interpolate: Interpolate to make length = 10, $S_{\text{intp}}$
      $T[n] = S_{\text{intp}}$
  return $T[n]$
```

3.1 Scraping and Processing Original Articles

I wrote a python script to scrape all of the 39644 articles using the source URLs, process them and save them to disk. An example article is shown in Figure 1. As shown in the figure, the article contents contain hyperlinks (<a></a> tags), word processing information (<em></em> and <b></b> tags), images (<img/>), videos (<object></object>), styling information (<i></i>, <span></span>) etc., which are irrelevant for narrative trajectory analysis. In order to get a clean text of the articles, I filtered the web contents for these tags using regular expressions. Finally, I saved all of the data to disk.

3.2 Extracting Sentence-wise Tones

I fed the articles which had more than 5 sentences to the IBM Watson Tone Analyzer API. The API generated and returned 13 scores for each of the sentences, as explained in Section 2.3. I saved all of the tone data to the disk as article-wise pickle files. This information enabled the construction of 13 tone or narrative trajectories for each of the articles.

4 METHODS

4.1 Constructing the Narrative Trajectories

For each of the 13 tone scores, I constructed a narrative trajectory using the procedure outlined in Algorithm 1. This algorithm is based on that proposed by Tanveer et al. [14], with modifications made to the lengths of the smoothing kernel and canonical signals to suit the nature of the dataset at hand. First, I used a 2-point averaging kernel to filter the original tone trajectory to a smoothed signal. Then, I cropped the signal in order to remove the boundary effects caused by the filtering. Since the articles contain a variable number of sentences, I interpolated the cropped signals using piecewise linear technique to have a canonical length of $N = 10$ samples. This final signal, $T[n]$, is referred to as the narrative trajectory for the particular tone of the particular article under consideration.

4.2 Clustering

I used density based clustering (DBSCAN) for finding major patterns in the narrative trajectories. The idea is to group similar trajectories together, and test whether the groups correspond to any difference in their member articles' popularities. Rather than enforcing model-based constraints over the shapes of clusters, as done in algorithms

[1] https://archive.ics.uci.edu/ml/datasets/online+news+popularity

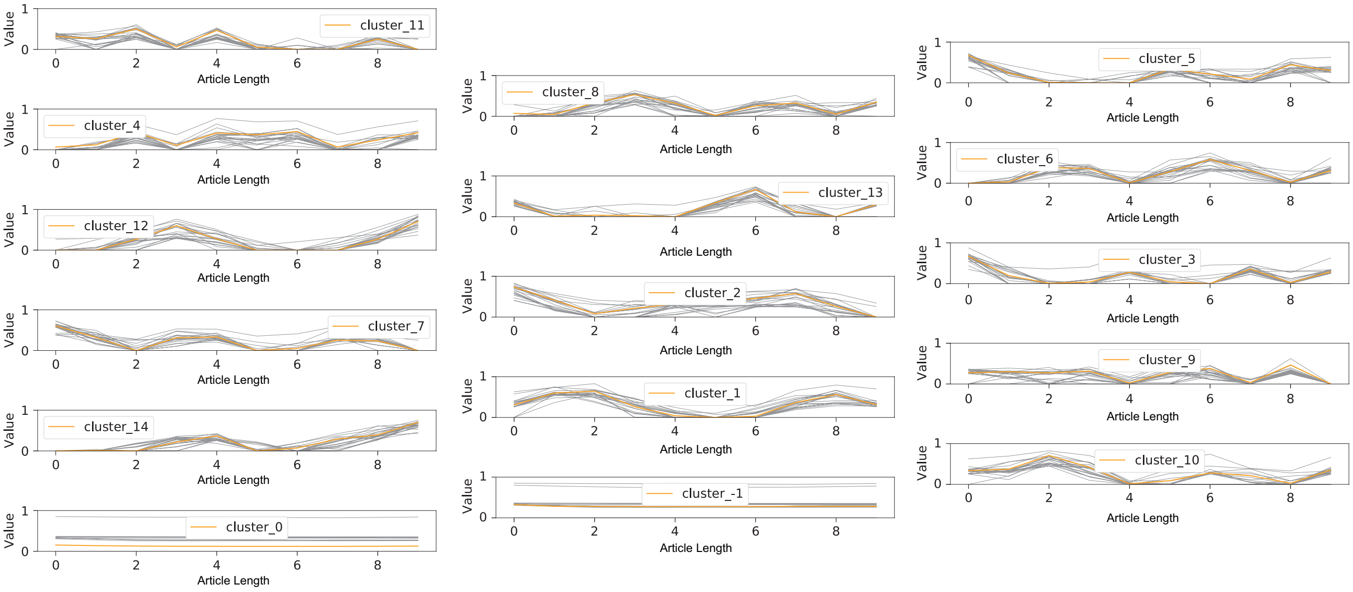


Figure 2: Clusters of joyfulness trajectories of the news articles. Among these clusters, clusters 6 and 10 both had significantly higher number of shares than cluster -1. Both of clusters 6 and 10 have a double-hill seesaw shape in their joyfulness, and end on a positive note.

like k-means, DBSCAN works by grouping data points based on density. Therefore, I found it to be a better choice of clustering algorithm for this project's purposes over non-density-based ones. I used the Scikit-learn package in python for the implementation.

4.3 Hypothesis testing

Once the clusters were obtained, I analyzed whether any of the clusters enjoyed significantly higher number of shares than others. For this purpose, I performed pair-wise t-tests for each of the cluster-pairs. I performed Bonferroni Correction to account for the statistical likelihood of obtaining significant results by chance. Using a Bonferroni-corrected significance level of 0.05, I then reported the cluster pairs showing a difference in popularity. I also reported the effect sizes using Cohen's d.

5 RESULTS

The tone *confident* lead to 2 clusters, which had a significant difference in the number of shares achieved (Bonferroni corrected $p = 0.01$, Cohen's d=0.05). Manual examination of the generated clusters showed that articles that maintained confident wordings throughout its body and ended in a highly confident note achieved significantly more shares than articles that did not have these characteristics.

The tone *joy* generated 16 clusters of plot trajectories. Among those, two cluster pairs showed statistically significant differences in the number of shares achieved. These clusters are shown in Figure 2. Cluster 6 had a significantly higher number of shares than cluster -1 (Bonferroni-corrected $p = 0.02$, Cohen's d=0.13). Similarly, cluster 10 also had a significantly higher popularity than cluster -1 (Bonferroni-corrected $p = 4.39E - 9$, Cohen's d=0.16). Manual examination of the results showed that articles that had a double-hill seesaw shaped variation in their joyfulness and ended on a joyful note achieved significantly number of shares than articles that had a flat joyfulness trajectory. The only difference between clusters 6 and 10 is the initial tone: one started with a more joyful note than the other. However, this did not hurt the success in their popularities.

The tone *analytical* showed statistically significant differences in 6 of its cluster pairs. However, manual examination of the cluster pairs revealed no actionable insights.

The scores corresponding to the tones *disgust, openness, conscientiousness, extraversion, agreeableness* and *emotional range* were almost always very close to zero (in a range of 0 to 1), with a mean of 0.15 and standard deviation of 0.08. This made it impossible to mine for insights from the trajectories of these tones.

For the tones *anger* and *fear*, 2 clusters were generated for each. However, these clusters did not show any significant difference in the number of shares after the Bonferroni correction was made to the p-values. Similarly, the tones *sadness* and tone *tentative* generated 4 and 8 clusters respectively, with none of the cluster-pairs surviving the Bonferroni-corrected significance in the pairwise t-tests.

6 DISCUSSION AND CONCLUSION

This project shows evidence that the success of online news articles can indeed be influenced by how its narration begins, builds up and concludes. The results indicate that a more confident approach in writing helps the likelihood that people will share the article. Again, having a seesaw shape in the joyfulness of the story narration also helps, as long as one ends with a positive note. The analyses are not without limitations, however, as this project does not consider any of the 58 features that came with the original dataset. Analyzing how the narrative trajectory information can add value on top of the original features remain as a future work. Also, since the number of articles in the dataset is large, it is possible to deploy deep learning tools such as LSTM to make the popularity predictions, which also remains as a future work. The insights obtained in this project are reliable and robust, due to the safeguards taken against chance results. More importantly, the results are intuitive and actionable, and therefore have the potential to add value to content creators and marketers in the online news industry.

REFERENCES

[1] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*. ACM, 607–616.

[2] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity. *ICWSM* 12 (2012), 26–33.

[3] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[4] Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. 2015. A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*. Springer, 535–546.

[5] IBM. 2017. The science behind the service. https://console.bluemix.net/docs/services/tone-analyzer/science.html#the-science-behind-the-service.

[6] IBM. 2018. Tone Analyzer, Understand emotions and communication style in text. https://www.ibm.com/watson/services/tone-analyzer/.

[7] Matthew L. Jockers. 2015. The Rest of the Story. http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/.

[8] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2011. Rt to win! predicting message propagation in twitter. *ICWSM* 11 (2011), 586–589.

[9] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The Emotional Arcs of Stories are Dominated by Six Basic Shapes. *EPJ Data Science* 5, 1 (2016), 31.

[10] He Ren and Quan Yang. 2017. Predicting and Evaluating the Popularity of Online News. (2017).

[11] Spyridon Samothrakis and Maria Fasli. 2015. Emotional Sentence Annotation Helps Predict Fiction Genre. *PloS One* 10, 11 (2015).

[12] R Shreyas, DM Akshata, BS Mahanand, B Shagun, and CM Abhishek. 2016. Predicting Popularity of Online Articles using Random Forest Regression. In *Cognitive Computing and Information Processing (CCIP), 2016 Second International Conference on*. IEEE, 1–5.

[13] Carlo Strapparava, Alessandro Valitutti, et al. 2004. WordNet Affect: an Affective Extension of WordNet.. In *LREC*, Vol. 4. 1083–1086.

[14] M Iftekhar Tanveer, Samiha Samrose, Raiyan Abdul Baten, and M Ehsan Hoque. 2018. Awe the Audience: How the Narrative Trajectories Affect Audience Perception in Public Speaking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 24.