Forecasting Popularity of News Articles

Kelley Zhao
MengMiao Wu

ABSTRACT

The data set is collected from bbc.com on 300 online news articles. The main objective of this paper is to determine different parameters that attract Internet user attention in order to predict the content popularity of online articles. Analysis presented in this paper uses linear regression with closed-form solution and gradient-descent solution. We found that the data generates a complex combination of weights and express a large error score.

I. INTRODUCTION

During the past two decade, the everyday use of digital technology increased drastically. People are not limited to the traditions of printed books or newspapers, Internet accessibility lead to an extremely easy access to information.

With a fast diffusion through various social media and networks, the appearing of E-newspapers impacts the news industry. Consequently, an increasing interest is focused on the "winning formula" for making successful online articles that will receive a great amount of user attention.

Our task is then to predict popularity of online news articles that will compute a display strategy and advertisement method that affect future products.

Many websites are able to provide desired information on news statistics. We chose BBC.com to generate an appropriate dataset.

In this paper, we focus on analyzing an existing dataset using regression and gradient descent to see the popularity of it. Then we compare it to running regression on a new dataset with different structure parameters to see the impact of those kinds of structure to the feature list.

II. REGRESSION ON EXISTING DATA

2.1 Methodology

The task is to predict the number of shares using the dataset provided. Both closed-form solution and gradient-descent solution are required. The header row and the URL column are removed from the dataset to yield a numeric matrix. The dataset X includes the first 59 features, and Y is the 60^{th} feature, the number of shares, in the original dataset.

Also, 31038^{th} example, the article titled "Civilians Fleeing Rebel-Held City in Ukraine Are Attacked", is removed. Few of its features are inconsistent with the purpose of the features, which lead to a suspected outlier.

Then, to avoid ambiguity of the dataset, normalization is applied on both X and Y matrices before perform gradient-descent.

Due to the large sample size, K-fold cross-validation is used in this case. We set k=5, the dataset is dividing into 5 equal subsets. For each round, four subsets are used as training sets and the remaining subset is used as test set. The true prediction error is computed by averaging of the errors for 5 rounds, and the model with the least test error over the 5 trials results the final model.

For gradient-descent solution, a difference vector is computed by subtracting w_{k+1} and $w_{k.}$ If the norm of that difference vector is smaller than epsilon, in this case, 0.0001, the algorithm has then reach the minimum. Therefore, the function returns the w_{k+1}. However, if one entry in the difference vector reaches positive or negative infinity, the algorithm rejects alpha. To determine the appropriate step size, K-fold cross-validation, where k=5, is applied to the dataset with alpha starting with $1/10^1$ to $1/10^{10}$, the exponent increments by one. As mentioned previously, the alpha with minimal test error over 5 trials result the proper model.

2.2 Results

The first analysis task is to find the best linear model with the linear regression. Generate the closed-form solution without normalization on the dataset results a weight vector with error score of 3.95E+9. We observed than the number of shares is largely influenced by the weekday on which the article is published. The 31^{st} 32^{nd}, 33^{rd}, 34^{th} and 35^{th} parameters, articles published on Monday, Tuesday, Wednesday, Thursday and Friday respectively, all have values of 3.19E+15. The 36^{th} and 37^{th} represent whether the article is published on Saturday and Sunday, these parameters have a negative weight, -4.33E+15 for both. This result the significant positive weight, 7,52E+15, on the 38^{th} feature, whether the article is published on the weekend. The feature w_0, intercept, is, however, negative with value of -2.19E+15. (See Figure 1)

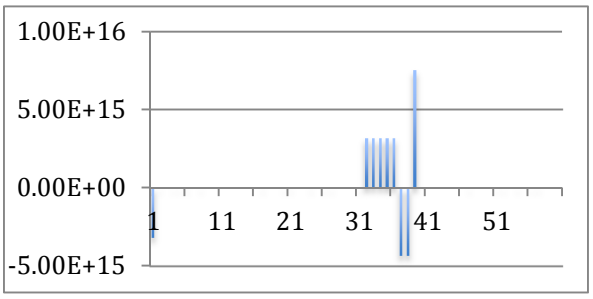

FIGURE 1: Weights of the features using linear regression

The results generated from closed-form with k-fold cross-validation show different conclusion. While the intercept and 31st 32nd, 33rd, 34th and 35th parameters stay mostly the same, 36th and 37th features have important positive weights, and 38th feature has a huge negative number, in contrast with data previously mentioned. (See Figure 2)

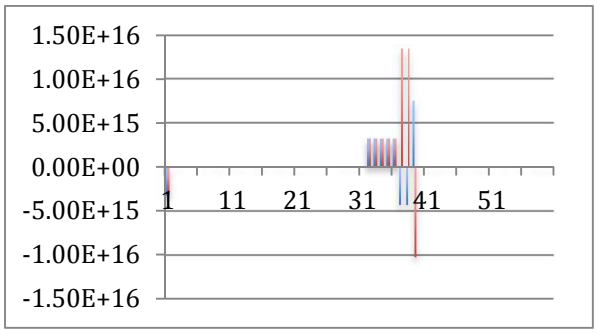

FIGURE 2: Comparison of the features weights between the result of linear regression with and without cross-validation

For gradient descent, alpha is determined with k-fold cross-validation, where k=5. (See Table 1) It is observed that the function has the lowest score with alpha=0.000001. Also, the dataset is normalized before performing gradient descent to avoid redundant value.

Figure 3 is generated with alpha =6. The parameters display a huge difference with both linear regression results. High weighted features are: the number of images, 0.76; Entertainment channel, 0.77; average keyword, 0.84; minimal shares of referenced articles, 0.97; whether it's Thursday, 0.91; numbers of positive words, 0.99, the highest; number of negative words, 0.81; and title subjectivity, 0.96. Low weighted features are: Social Media channel, 0.01; Tech channel, 0.02; and LDA related topic, 0.00.

ALPHA	TESTING ERROR
$1/10^6$	3.40E+04
$1/10^7$	6.75E+04
$1/10^8$	8.95E+04
$1/10^9$	7.91E+04
$1/10^{10}$	6.36E+04

TABLE 1: Test error score of gradient descents with alpha starting with $1/10^6$ to $1/10^{10}$

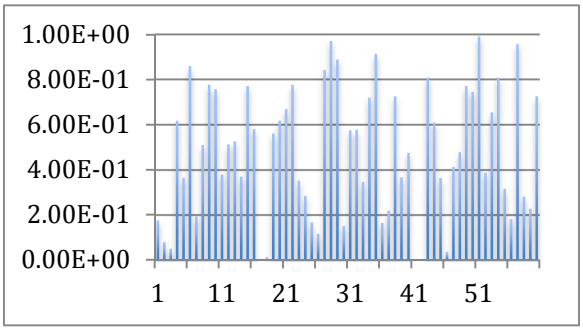

FIGURE 3: Weights of the features using gradient descent with cross-validation

3. REGRESSION ON NEW DATASET

3.1 Previous Works

Many researchers have focused on predicting the popularity of news sites in order to find the perfect formula for news release. The vast majority of these studies have explored how features that would normally be linked to popularity actually effected popularity. For example, Bandari et al. (2012) examined the category, subjectivity, named entities, and source influence to predict the popularity measured by how many times the article was tweeted. Arapakis, Cambazoglu, and Lalmas (2014) were influenced by the above, using the same metrics, except they looked specifically at cold-start news popularity. However these papers focused on typical attributes of articles, but have not analyzed the structure of the article at all. This experiment aims to use the normal attributes of articles along with the structure of the article to see if prediction accuracy increases.

3.2 Methodology

The primary purpose of this task is to generate a new dataset based off of BBC news articles and apply linear regression to it to predict the number of shares of an unknown article. However, the focus of this experiment is to see how detrimental or effective adding the structure of the article to the feature set is

to the prediction accuracy. A web crawler was built using Beautiful Soup, to grab the webpages, Selenium, to generate the needed javascript details, and NLTK to gather features about the structure of the article. 1700 random articles were scraped using an iterative deepening method.

The features that were selected were picked to try and accurately show how popular the article should be. The main features that were selected were the number of likes, number of comments, and weekday that it was published. The number of likes and number of comments were represented by just the number itself into the matrix. These features were motivated because they are clear representations of how popular an article is. The weekday was represented by seven fields where the value in that column would be 1 if the article was published on that day and 0 otherwise. This feature was motivated because readers might read the article more if it was on certain days, as seen from the above experiment. Although these were the primary indicators of how popular an article could be, four structure parameters were included as well: number of words, categories, entity, and part of speech. The number of words in the article was motivated because how long an article is can affect the popularity of it. This was represented by just the number itself in the matrix. The category parameter was selected to because different categories will be more popular and certain categories have different structures (more or less words). This data was represented with X fields in the matrix for each category. The entity parameter tags each word with one of the following categories: Person, Location, Organization, Money, Date, Time, and Percent. The part of speech feature was similarly done, except with each word's part of speech. These two parameters were represented in the matrix by how many words were part of that category. These two features were selected because they mainly show the structure of the article. An appendix is included showing each feature and what they mean.

Finally, when the features for the part of speech were picked, many parameters were underrepresented (1 in 1000 articles for instance) so these parameters were tuned out of the feature set as they would just generate unnecessary noise.

3.3 Results

The features in the linear regression graph (Figure 4) that were useful were mostly the categories, the days of the week, and the word count. The categories that were impactful were: help, disability, entertainment, magazine, UK, and elections with the highest at 1.44E+3 for help. For the days of the week, Monday was the most impactful with a value of -1.35E+02. The word count had a value of 9.28E+01. The features that were most useful for the cross validation (Figure 5) are also the categories, days of the week, word count, and the number of likes. The helpful categories were all the same with the most at help with a value of 1.45E+03. For the days of the week, not only was Monday large, but every day except Thursday was impactful with the most at Saturday with a value of 1.47E+02. Finally, the values for the word count and likes are 1.52E+02 and -1.13E+02 respectively. The error of the linear regression and cross validation were 2.6E+06 and 1.8E+08.

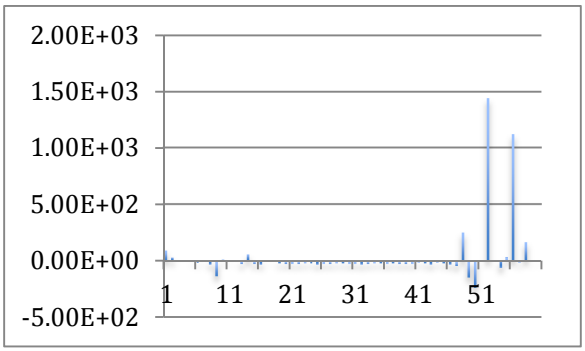

Figure 4: Weights of the features using Linear Regression

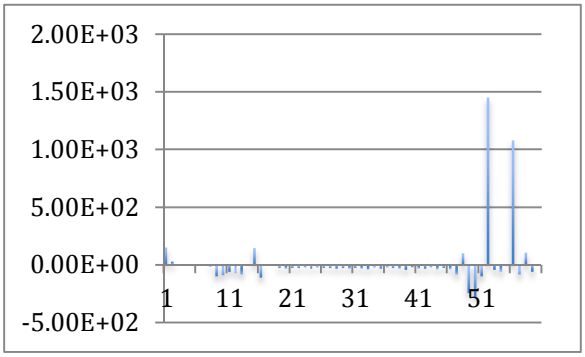

Figure 5: Weights of the features using Cross Validation

From these results, we can conclude that most if not all of the part of speech and entity identification was not as impactful as the other features. However, the category and word count of the articles were the most impactful and could be parameters for the structure of the article. Thus we can conclude that the intricate details of the structure of the article like the entity tagger and part of speech tagger were not as significant, but the bigger structure of the article like the category and word count were.

4. FINAL DISCUSSION

4.1 Usefulness and Advantages/Disadvantages

This dataset was limited in that there were only 1700 articles scraped which is small for this classification since it is difficult and very unpredictable. However, they should be enough to predict results for simpler classification such as genre of the article. Thus the data gathered can be used, except it will not be as accurate as a much larger dataset. Furthermore, no data was acquired about the subjectivity, negativity, and positivity of words, which were weighted highly in the gradient descent experiment above. The advantages however are that the dataset contains information about an article that were not heavily tested in the past for the popularity of an article.

4.2 Future Tests and Open Questions

In the future, the dataset could be gathered with the extra features above, which are shown to be important in this classification. Also, a different library should be used to gather a dataset as the library performed very slowly and was the primary reason the dataset was not that big. With regards to different kinds of features, the scraper could follow feeds from social media to gather real-time information about news articles that could generate information about popularity such as the time it takes for an article to obtain 100 likes. The biggest open question about the second experiment is if the second dataset was much bigger and had more features about the structure of the article, would it have performed better?

REFERENCES

[1] Alexandru Tatar, J´er´emie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida, “Predicting the popularity of online articles based on user comments,” in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, New York, NY, USA, 2011, WIMS ’11, pp. 67:1–67:8, ACM

[2] Arapakis, Ioannis, B. Barla Cambazoglu, and Mounia Lalmas. "On the Feasibility of Predicting News Popularity at Cold Start." *Lecture Notes in Computer Science Social Informatics* (2014): 290-99.

[3] Bokesoy, Deniz. "E-newspapers: Revolution or Evolution? E newspapers: Revolution or Evolution? | Bokesoy | *Scroll.* University of Toronto, 2008. Web. 28 Sept. 2015.

[4] R. Bandari, A. Sitaram, and B. A. Huberman. “The pulse of news in social media: Forecasting popularity”. *In Proc. 6th Int’l Conf. Weblogs and Social Media*, 2012.

APPENDIX

Features are in the order in the graphs above

Word Count: the number of words in the article
Comments: the number of comments for the article
Shares: the number of shares for the article
Likes: the number of likes for the article

Monday: if the article was written on Monday
Tuesday: if the article was written on Tuesday
Wednesday: if the article was written on Wednesday
Thursday: if the article was written on Thursday
Friday: if the article was written on Friday
Saturday: if the article was written on Saturday

World: news about the world
Blogs: news about blogs
Disability: news about disabilities
Entertainment: news about entertainment
Magazine: news about magazines
Technology: news about technology
Help: news about help
Business: news about business
Education: news about education
UK: news about the United Kingdom
Rugby: news about rugby
Election: news about election
Health: news about health
Science: news about science

DT: number of determiners (the)
JJ: number of adjectives (big)
NN: number of singular nouns (door)
VBZ: number of verbs, present, 3rd person (takes)
VBN: number of verbs, past participle (taken)
TO: number of to’s (to go)
VB: number of verbs, base form (take)
CD: number of cardinal numbers (1, third)
IN: number of prepositions (in, of)
JJS: number of superlative adjectives (biggest)
VBG: number of verbs, present participle (taking)
NNS: number of plural nouns (doors)
NNP: number of proper nouns (John)
VBD: number of verbs, past tense (took)
PRP$: number of possessive pronouns (my, his)
CC: number of coordinating conjunctions (and)
WP: number of wh-pronouns (who)
RB: number of adverbs (however)
PRP: number of personal pronouns (I, he)
VBP: number of verbs, present, non-3rd person (take)
RBR: number of comparative adverbs (better)
RP: number of particles (up)
MD: number of modals (could, will)
WDT: number of wh-determiners (which)
WRB: number of wh-adverbs (where)
JJR: number of comparative adjectives (bigger)
EX: number of existential theres (there is)
NNPS: number of plural proper nouns (Johns)