Online News Popularity

Dhivya Swaminathan, Hasika Mahtta

Abstract
Social Media consists of numerous applications and platforms for creating and sharing contents humans generate. Social Media has grown immensely over the last decade especially in the online social networking services. Social Media is easily accessible and it also brings together the opinions and experiences of diversified groups of users globally. Social Media also facilitates News Sharing - both for media organizations and individuals. So, if we extract and analyze the data on social media content properly, it can help us derive important predictions of human related events. Such predictions have great benefits in many realms, such as finance, product marketing and politics, health and mindfulness awareness, etc. The number of shares under a news article indicates how popular the news is. The ability to predict which online articles will be most popular will help us know about how media companies attract their clients, how important information is spread, how advertisements can be more effective and how public opinion is formed. The main goal of this project is to use features of online articles to predict online article popularity which can be accomplished using supervised machine learning algorithms. We also intend to predict the popularity before an article is published.

Keywords
Social Media — News Sharing — Machine Learning — NLP

Contents

Introduction

Social media [3] has become an important form of online news sharing. People gain and share knowledge from social media and it has become an important source of entertainment across the world. We intend to build a model that will be able to classify the popularity of a news article based on how many times an article was shared. We will make use of the dataset collected with over 39000 articles from Mashable website to select important feature and analyse them to compare its performance with machine learning algorithms.

1. Background

We have referred and studied the paper - ”Predicting and Evaluating the Popularity of Online News” [1] for our analysis. This paper has used supervised machine learning techniques (Regression and Classification) to predict the popularity of an article based on the labeled data where number of shares for each article is the label. They have also scrapped the titles and article contents from the urls of the articles to predict the popularity of the articles based on bag-of-words techniques. We have further researched upon regression and classification techniques for machine learning models and optimized them using GridSearchCV method to obtain stat-of-the-art results. We have also used NLP techniques such as Naive Bayes to predict the popularity.

The main goal of this paper [3] provides a review of articles that examine the relation between news sharing and social media in the period from 2004 to 2014.They have shown results from the reviews that were used to provide analysis of current research on the progress of news sharing research. This paper [5] has emphasized to present a comprehensive review of detecting social media, inclusive of psychology, and social theories, analyse existing algorithms from a data mining perspective, evaluation metrics and the data sets.They

have also discussed areas that are related to the problem of fake news detection. They aimed to identify the differences between these areas and fake news detection by explaining the task goals and highlighting some methods such as Rumor Classification, Truth Discovery, Click-bait Detection and Spam and Bot Detection.

2. Models and Methodology

2.1 Exploration Data Analysis

2.1.1 Dataset

For this project, [1] we will be using the Online News Popularity Dataset from the University of California Irvine Machine Learning Repository. The dataset [2] summarizes a heterogeneous set of features about articles published by Mashabale (www.mashable.com) in a period of two years (2013-2015) whose goal is to predict the number of shares in social networks (popularity). It has 61 features out of which 58 are predictive attributes like topic, day of publication, number of images, length of the words and sentiment analysis, etc, 2 non-predictive (URL of the article and days between the article publication and the dataset acquisition and one goal field (shares : Number of shares). The attributes are integer or real values and have no missing values. These features do not include information about the actual words found in the article, which limits how specific the predictions can be. However, the dataset also contains URL links to the articles themselves, which allows to gather the content of the articles.

2.1.2 Feature Selection and data exploration

We have used basic data exploration steps and found that there were no missing values or null in the data and the dataset is balanced. We have further done feature normalization to build the models. We have plotted a graph to visualize the correlation among various attributes in the dataset. We selected the features based on their importance and correlation to each other. We have dropped the features which are highly correlated to each other by setting up a threshold value of 0.5.

For classification technique,we considered a threshold value as the mean of the labels which came out to be 1400. We labelled the data whose Number of shares value was greater than 1400 as 1 and the 0 otherwise.

For both the regression and classification models, we split the data with test size of 0.33 and random state value as 42.

2.2 Predicting the number of shares of an article

Regression [6] is a technique which is used to predict the quantity values when the output quantity is continuous.

2.2.1 Baseline Models

We have used baseline models initially to predict the popularity with Linear Regression, Random Forest, Gradient Boosting, AdaBoost, Ridge Regression, and Lasso Regression wit default parameters. Out of these Lasso Regression performed better than other models with RMSE Value for test dataset as 12732.531.

2.2.2 Lasso Regression

Lasso Regression [7] Lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. (Wikipedia Definition)

The goal of lasso regression is to get the predictors that could reduce the prediction error for that variable. It can be used when we want simple models with fewer parameters.

2.2.3 Optimizing the Model using GridSearchCV

Lasso Regression Model has been tuned and optimized using **GridSearchCV** technique with five fold cross validation value. **GridSearchCV** [8] implements a ”fit” and a ”score” method. It also implements “predict”, “predictproba”, “decisionfunction”, “transform” and “inversetransform” if they are implemented in the estimator used.

The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

2.3 Classification of articles based on popularity

2.3.1 Baseline Models

We have used baseline models initially to predict the popularity with K-Nearest Neighbours, Logistic Regression, Support Vector Machine, AdaBoost, Random Forest and Gradient Boosting Classifier with default parameters. Out of these Gradient Boosting Classifier performed better than other models with test accuracy of 64 percentage. We have further optimized the model hyper parameters using GridSearchCV method.

2.3.2 Gradient Boosting Classifier

Gradient boosting [9] is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. (Wikipedia definition)

In Gradient Boosting Classifier, we first build baseline models and then analyze the data for any errors which indicate those data points that would be hard to fit. And so for further models, we will have to focus on those points to get improve the models. Finally, we will have to combine all the predictors by assigning weights to each predictors.

2.3.3 Optimizing the model using GridCV

Gradient Boosting Model has been tuned and optimized using **GridSearchCV** technique with five fold cross validation value. **GridSearchCV** [8] implements a ”fit” and a ”score” method. It also implements “predict”, “predictproba”, “decisionfunction”, “transform” and “inversetransform” if they are implemented in the estimator used.

The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

2.3.4 Polynomial Regression

Polynomial regression is special case of linear regression. It is based on how we select the features.

Polynomial regression [10] is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y. (Wikipedia Definition). It is also called multilinear regression.

We have used quadratic, three degree and four degree polynomial function to derive different set of features. Thereafter , we have applied gradient Boosting Classifier to these features and test accuracy. We did not get any improvement over other models with this polynomial regression.

Gradient boosting model without polynomial regression with parameters learning-rate = 0.01, n-estimators = 1750 seems to perform the best for determining popularity of an article.

2.4 Article Popularity classification using NLP Techniques

We intend to predict the popularity of articles directly from the content of the article by extracting the information of the article from the article URLs provided in the dataset.

2.4.1 Beautiful Soup

Beautiful Soup [11] is a python library that is very useful in web-scrapping to retrieve data from HTML and XML files/links.

We have used it to extract title and article contents and added them to the database and found that scrapping all the 39644 URLS is taking more time to process. So we decided to take 2000 article titles and contents for predicting the popularity.

2.4.2 Tfidf Vectorization

TFIDF (term frequency–inverse document frequency)[12] is an important statistical technique used in information retrieval. It tells us the importance of a work in a document or courpus. It's value is directly proportional to the number of times a word appears in the corpus.

TfidfVectorizer [12] transforms text into feature vectors that can be used as input to the estimator. We used this methodology on both titles and contents attributes of the dataset and selected best 500 words for further analysis to build the estimation model and concatenated them with the initial dataframe.

2.4.3 Feature Scaling

One important aspect of machine learning algorithms is that they consider only the measurements and not the units of those measurements. And that affects the prediction a lot because a feature that has a very high magnitude. Scaling helps to solve this problem. We performed Feature Scaling [13] using scaling algorithms in Scikit-Learn like Standard Scaler and MinMax Scaler

The Standard Scaler - This algorithm assumes Gaussian distribution(Normal Distribution) for the data. The concept behind Standard Scalar is that i transforms the data in way that the distribution will have a mean value of 0 and a standard deviation of 1. It is not very useful if the data is not normally distributed.

MinMax Scaler - The Min-Max Scaler uses the following formula for calculating each feature:

(xi–min(x))/ (max(x)–min(x))

It transforms the data such that we can pass a specific range as required. It can be done by passing in a tuple to the feature-range parameter. By default, the range is between 0 and 1 (-1 and 1 if there are negative values). Min-Max Scaler can be used instead of Standard Scaler when the data is not normally distributed.

2.4.4 Model Implementation

We scaled the features using Standard Scaler and used Logistic regression and random forest algorithms from Scikit-Learn to estimate the accuracy.

Logistic regression [14] is used when the dependent variable is binary. We started with Logistic regression model and optimized it using GridSearchCV and obtained 71.0 percentage accuracy.

Random Forest Classifier - Random forests [15] Classifier are an ensemble learning method which classifiers the data by creating decision trees at the training time and outputs the classification or mean prediction (regression) of the individual trees.(Wikipedia Definition)

We obtained an accuracy of 57 percentage using random forest classifier with Standard Scaler which was not an improvement over Logistic regression with Standard Scaler.

Naive Bayes Model - Naive Bayes [16] classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with naive independence assumptions between the features.The naive bayes model generalizes strongly that each attribute is distributed independently of any other attributes.

We scaled our features using MinMax Scaler and Naive Bayes algorithm which gave a better accuracy of 79 percentage which is better than other alogrithms.

3. Experiments and Results

3.1 Predicting number of shares of an article

For predicting the number of shares for an article, various regressors like Linear regression, Random forest, Gradient Boosting, Adaboost, Ridge Regression and Lasso Regression were employed as baseline models. The performance of these models were gauged using Root Mean Squared Error (RMSE) and are as listed in Fig 1.

From this, it can be seen that Lasso regressor performs the best and thus it was selected for parameter tuning. Upon using GridSearchCV with a 5-fold cross validation, the ideal value for alpha was found to be 9. The RMSE for the Train data set was found to be 10839.55 shares and the RMSE for the Test

Baseline Model	Test RMSE
Linear Regression	12756.85
Random Forest	13366.35
Gradient Boosting	12823.87
Adaboost	26309.23
Ridge Regression	12733.21
Lasso Regression	12732.53

Figure 1. RMSE of different baseline regressor models

data set was 12731.51 shares. The paper that we referred to [4] reported a test RMSE of 12696 shares.

3.2 Classification of Article based on popularity

Initially, five baseline models were established that includes - K-Nearest neighbours, Logistic regression, Adaboost Classifier, Random Forest Classifier and a Gradient Boosting classifier. Out of these models, Gradient boosting classifier gave the best performance with a Test accuracy of 64.71% and and AUC score of 0.705. Therefore, this model was analysed further, its parameters optimised using GridSearchCV with a 5 fold cross validation to achieve the best results.

Baseline Model	Test Accuracy	Test AUC
K-Nearest Neighbours	57.59%	0.590394
Logistic Regression	56.16%	0.605481
Adaboost	64.37%	0.694879
Random forest	60.35%	0.646661
Gradient boosting	64.71%	0.704359

Figure 2. Performance of different classifiers for predicting article popularity

Using a 5 fold cross validation in GridSearchCV model, various parameters like learning_rate, n_estimators, max_depth, min_samples_split, min_samples_leaf and subsample were optimised. The parameters and their optimised values are as shown in the table below:

Parameters	Optimized values
Learning rate	0.01
n_estimators	1750
max_depth	7
min_samples_split	2
min_samples_leaf	1
subsample	1

Figure 3. Gradient Boosting model and its optimized parameter values

Further to this, it was found after experimentation that a three and four degree polynomial features generated and added as part of the feature set does not aid in improving model performance.

Therefore, after implementing the gradient boosting model with optimised parameters, we were able to achieve an accuracy of 65%. The paper we are referring to [4] reported an accuracy of 66% using a Random forest classifier.

3.3 Classification of article popularity using NLP techniques

In order to improve the accuracy of classifying articles based on popularity, in addition to the existing feature set, we decided to utilise the title and content of the article to see if that influenced its popularity. Using Beautiful soup, we scraped the content using the url available as part of the dataset. Post extraction, using concepts of TFIDF model for text scoring, we extracted the top 500 best words based on their TFIDF score and added them to the dataset as new features. This TFIDF tokenization and selection of top 500 best words was carried out for both title and article content.

Experimentation with two different scalers- Standard and Min-Max was carried out. A logistic regression model and Random forest model was built using the features scaled by Standard scaler and a Naive Bayes model with MinMax scaler was built which ultimately resulted in the best accuracy of 79.55%. The paper that we referred to [4] reported an accuracy of 67%. Since they took only the top 300 words, the results might be poor as compared to our model with 500 best words.

Model	Test Accuracy
Logistic Regression with Standard Scaler with top 500 best words	71.10%
Random Forest with Standard Scaler with top 500 best words	57.61%
Naïve Bayes with MinMax Scaler with top 500 best words	79.55%

Figure 4. Gradient Boosting model and its optimized parameter values

4. Summary and Conclusions

In this project, we have attempted to predict the number of shares an article could garner using Lasso regression with an RMSE of 12732.5 shares, classify the articles based on popularity using Gradient boosting model with an accuracy of 65% and using Natural Language Processing techniques of TFIDF model and Naive Bayes and arrived at an accuracy of 79.55%.

Future works in this area can be done by implementing various other techniques of NLP. Models like BM25, Vector Space models, etc can be employed. Also, there might be some articles with catchy titles to attract views, but wouldn't be having the promised contents. Models to identify such articles and penalize accordingly could be carried out as part of future work.

References

[1] Ren, He and Yang, Quan - *Predicting and Evaluating the Popularity of Online News*

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

[3] Kumpel, A. S., Karnowski, V., and Keyling, T.(2015) - *News Sharing in Social Media:A Review of Current Research on News Sharing Users, Content, and Networks. Social Media + Society.*

[4] Joe Johnson, Noam Weinberger- *Predicting News Sharing on Social Media*

[5] *Fake news detection on social media: A data mining perspective*, author=Shu, Kai and Sliva, Amy and Wang, Suhang and Tang, Jiliang and Liu, Huan,journal=ACM SIGKDD Explorations Newsletter,volume=19,number=1,pages=22–36,year=2017,publisher=ACM

[6] *https://www.quora.com/What-is-regression-in-machine-learning*

[7] *https://www.statisticshowto.datasciencecentral.com/lasso-regression/*
https://www.coursera.org/lecture/machine-learning-data-analysis/what-is-lasso-regression-0KIy7
https://en.wikipedia.org/wiki/Lasso$_{(statistics)}$

[8] *https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html*

[9] *https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d*

[10] *https://acadgild.com/blog/polynomial-regression-understand-power-of-polynomials*
https://en.wikipedia.org/wiki/Polynomial$_{regression}$

[11] *https://www.crummy.com/software/BeautifulSoup/bs4/doc/*

[12] *https://en.wikipedia.org/wiki/Tf-idf*
https://stackoverflow.com/questions/25902119/scikit-learn-tfidfvectorizer-meaning

[13] *https://medium.com/@ian.dzindo01/feature-scaling-in-python-a59cc72147c1*

[14] *https://www.statisticssolutions.com/what-is-logistic-regression/*

[15] *https://en.wikipedia.org/wiki/Random*$_{forest}$

[16] *https://en.wikipedia.org/wiki/Naive_Bayes_classifier*
https://stackoverflow.com/questions/10614754/what-is-naive-in-a-naive-bayes-classifier