Online News Popularity Prediction

Abhijit Kannepalli, Bhavesh Chatnani, Hemil Kothari, Saurabh Saoji, Shrey Shah

School of Information Sciences, University of Illinois at Urbana-Champaign

IS 507: Data, Statistical Models, and Information

December 10, 2022

Executive Summary

Introduction and Dataset

With the advent of the internet and social media, online news has become integral to everyone's life. The most significant increase in attention paid to a specific news piece is known as news popularity. For reading various news stories, people use social networking and news websites. It is a medium through which everyone keeps themselves updated about the recent happenings worldwide. A lot of variables, including the quantity of social media shares, the number of visitor comments, the quantity of likes, etc., influence the popularity of online news. Building an automated decision support system to forecast news popularity is therefore vital because it will aid in business intelligence. An area that we are typically trying to focus on is the factors influencing the sharing of news articles.
The dataset that we are using is from the UCI Machine Learning Repository, which contains the data of 39797 news articles published by Mashable over a period of two years. There are 61 features, including the number of shares for each article. It contains details such as the number of images, videos, and keywords in an article. Positive and negative polarity ratings, the day the article was published, and the channel through which the article was published are also included. Through this research, we are trying to predict the number of times an article might get shared and categorize it as either a popular or a common article, segregate articles based on their polarity, and identify relationships of factors with the number of shares.

Methods

Firstly, we identify which features are essential and should be considered to predict the number of shares for the news articles. Since our dataset contains around 60 features, we tried to reduce the complexity. It is necessary to compress the maximum amount of information with fewer features. We can achieve this using Principal Component Analysis. We group the news articles into different groups such as positive, negative, neutral based on the sentiments of the content using K-means clustering. Whether the type of news article - lifestyle, technology, or business may affect its popularity is something that we intend to identify. The number of shares may be influenced by when one reads an article - weekday or weekend. Initially, we checked if we could attempt to reduce the features and then proceeded to use PCA. After finalizing the feature set, we proceed to categorize the different polarities - positive, negative, and neutral, using a K-Means clustering to group the articles in clusters based on their sentiments. In order to predict the number of article shares that categorizes the article to be popular or unpopular, we used XGBoost, a boosting algorithm. We were able to categorize them properly about 71% of the time.

Results

After applying the PCA technique, we reduced the number of features from 60 to 5 major components, namely, number of tokens, shares based on reference to other Mashable articles, keywords-based shares, the positive influence of news, and stop words and token length. These form new features that can be used to predict the popularity of news articles. The resulting model can explain 68% of the information in the original data. We successfully categorized articles

based on their polarity into three categories i.e., positive, negative, and neutral. It was found that negative articles and positive articles were almost the same in number. This tells us that authors don't just focus on writing negative articles, however we can say that authors tend to stick to a more polar article i.e positive or negative rather than being neutral. It was also found that the titles that were highly polarizing were more in number compared to neutral titles. An accuracy of 71% was achieved in predicting online news articles. This tells us that out of every 100 articles, 71 were categorized correctly in terms of popularity. Additionally, we used statistical tests to understand if the day of the article being published or the data channel of the news article had any impact in determining the popularity of the news article. Our subsequent results showed that these factors had little to no contribution in deciding the number of shares an article has, suggesting that factors like sentiment, keywords, and overall content have more impact on the popularity of an article regardless of where or when it is published.

Limitations

For identifying the reduced feature set, checking whether the data is suitable for performing the reduction process was a challenge, and we had to satisfy a few assumptions to proceed. Another limitation is that it may not always be possible to interpret the new features obtained. Our study assumes a threshold value. If the predicted number of shares is higher than the assumed threshold value, we categorize it as a popular article and vice versa. However, the threshold value is set based on domain knowledge and some calculations. A significant shift in the number of articles shared in the future may compel us to change the threshold, leading to a change in prediction accuracy. Even though we can classify articles based on their polarity into positive, negative, and neutral categories, we need to consider the intensity of the polarization. For example, one article might be popular due to being hyper negative, but another might be less popular due to it being moderately negative. Even though the dataset is quite large, it is not large enough to make a generalized statement about the author's writing style. It is essential to distinguish between the two, and such insights are not drawn from this study. Another limitation we faced was that we did not factor in the news article's content. The current factors available to us are more objective, and in spite of giving a good amount of information, the overall semantics of the article cannot be deciphered, which would be a critical factor in predicting the overall popularity of a news article.

Conclusion

The research centered on locating characteristics that may play a significant role in determining which news stories are widely read. The principal component analysis technique was used to condense the number of characteristics that were considered to more accurately represent the initial data. According to the findings of the experiments, applying XGBoost as a solution for predicting the level of popularity of online news articles is an efficient way to solve the problem. It was also found that the percentage of articles that were either positive or negative in polarity were higher compared to neutral ones, indicating that authors preferred writing more polar articles.