Duke University
March 5th, 2018
Team 23: Jess Beering, Molly Carmody, Bella Hutchins, Ryan Nicholson, Catalina Sanchez-Carrion
CS 216 - Proposal write up
Title: Predicting Popularity in News Media

Why we chose this topic:

Social media today is one of the most prominent ways people find important information about our world regarding politics, the economy, social issues, and more. The conflation of social media and news reporting is an especially contentious issue since the 2016 U.S. election, given the impact that social media posts are reported to have had on the election results. News and media outlets aim for their posts to reach the largest number of people as possible, and thus have a high popularity among social media users. There are various factors that the popularity of an article depends on, including the day and time it is published, the subject matter, multimedia within the piece, and more. Therefore, we are interested in studying the impacts that factors like these, as well as other outside factors we will explore, have on the number of shares of an article in order to predict articles' popularity.

Questions We Have:

While working with this dataset and with ancillary data, several questions arise that will contribute to our project's overall direction. Namely, which attributes will best predict the popularity/number of shares an article receives? Do specific days of the week garner more shares than others? Which category of news, for instance, lifestyle, entertainment or tech, brings in the most shares? Do those with more images or videos do better than those with fewer? How does the rate of positive versus negative words in the content affect the number of shares? Are trending topics on Twitter correlated with which articles get the most shares? What about trending topics on Google? Does the total number of Facebook and Twitter users impact the popularity of articles?

Data Sets to be Used:

We are sourcing some of our data from UCI, using their Online News Popularity Data Set. The data is based off select articles published on Mashable. For each article, the data set presents 60 attributes, which are located at the bottom of the proposal report. Additionally, we will be using outside data to further analyze attributes that impact the popularity of articles that aren't included in the UCI dataset.

Within the UCI data set, we are given two files: OnlineNewsPopularity.csv and OnlineNewsPopularity.txt. OnlineNewsPopularity.csv is a csv file with each column corresponding to an attribute and each row corresponding to an article. The repository for this data set specifies that the data was used, with their specific model, to generate predictions about the number of shares an article would receive. Given that the archive and data information do not specify the actual predictive model used, just the attributes, we plan to create our own predictive model to approximate article shares. The dataset includes the actual number of shares that each article received, which we will be able to use to check the success of our model.

We also want to examine data outside of this extensive dataset to see how other factors can impact the popularity of articles. Twitter provides datasets that list trending topics from specific dates. We want to use this information to see if articles that use words present in trending topics during the time the articles were collected (November 2015-July 2016) receive more shares. Additionally, Google provides infomation regarding the most trending searches per month, so we can similarly look at this data and measure the strength of the correlation between an article's popularity and the number of times a trending term appears in an article. We plan to look for similar data from the other social media websites that you can share articles from through Mashable, such as Facebook and Pinterest. We also want to find and organize data regarding the number of users that these different social media platforms had over the specified time frame so that we can examine the correlation between the number of users on the platform/the date when the article was published and the article's popularity. For example, we have one source that depicts a graph showing the growth in Facebook users between 2008 – 2017, so we could look to see if the increase in users between 2015 and 2016 impacted the popularity between articles published in these two years.

Evaluating Success

In order to evaluate our success, we intend to compare how close our estimates are to the actual total number of shares that these articles received on Facebook, Twitter, and other social media platforms. These values are present in the csv data file. Additionally, we will compare our results to other people's predictions using either the same or different data sets, perhaps through a competition that Kaggle is hosting (see below) for predicting the number of shares with the UCI data. Our hope is that by bringing in the outside ancillary data we've begun to collect, regarding number of users, trending topics, and more, we will be able to more accurately predict the number of shares that articles have than we would had we only been looking at the UCI data alone. Finally, we will try to optimize the multivariate linear model in order to gain the maximum R^2 value between predicted values and actual values.

Competition for predicting popularity: https://www.kaggle.com/c/predicting-online-news-popularity

Steps

1. Analyze the content and results from other studies that have used this dataset to help broaden the scope of what we can do with the data.
2. Create SQL queries to clean the csv file we downloaded, filter out attributes that aren't important for our prediction, attributes which are not well documented, and the "shares" column.
3. Create organized tables of the ancillary data we found, as much of it is not yet organized into tables that will allow us to easily access information when we need it for training data, etc.
4. Plot attributes (both from UCI data and ancillary data) on shares to visually examine/approximate which attributes contribute the most toward number of shares
5. Examine different predictive models
6. Compare our results to the actual results of shares given to us from the csv file
7. Submit our results to the Kaggle competition

Attributes, as listed on https://archive.ics.uci.edu/ml/datasets/online+news+popularity#

0. url: URL of the article (non-predictive)

1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity

46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

Dataset Source

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.