Methods II Term Paper

Tawa Suleman

31 May 2021

I. *Introduction*

My selected task was classifying whether articles are clickbait based on their titles. In a time when many articles are being shared on social networking sites from FaceBook to Twitter, it can be helpful to know whether the content being shared has potentially important information or if it is just content made for more ‘clicks’.

II. *The Data*

The data I used was downloaded from “Clickbait Dataset” by Aman Anand on Kaggle. The data contains two columns, “headlines” and “clickbait”. The “headlines” column contains the title for each article, and the “clickbait” column consists of 0’s and 1’s, where 0 represents non-clickbait articles and 1 represents clickbait articles. There are a total of 32,000 articles, taken from various sites such as BuzzFeed and New York Times. The data was split into *train* and *test* sets, with 80% of the data used for training and the remaining 20% for testing.

III. *The Process*

After getting familiar with the data, I worked on a list of features which I felt tended to distinguish clickbait headlines from non-clickbait ones:

i. *Contains Adverb-Adjective phrase*
A trait I noticed in a lot of the clickbait headlines was that they had adverb-adjective phrases. For example, “27 Insanely Clever Halloween Costumes For Your Dog.”

ii. *Contains superlative adjectives*
Another common trait I noticed was the use of superlative like “best” or “funniest”, as in the articles “The 19 Best Nonfiction Books Of 2015” and “This Video Of A Man Pushing A Bin Up An Icy Path Is The Funniest Thing You'll See Today”.

iii. *Contains wh-words*
Though there are non-clickbait headlines that contain wh-words, I found a lot of headlines that contained the words such as “why”, “who”, and “which, and felt it might help in classifying the

articles. (e.g. “26 People Who Got Shittier Christmas Gifts Than You”, “11 Reasons Why Kanye West Shouldn't Be So Sad About His Height”)

iv. *Contains first-person pronouns*

It is very rare for any headlines containing first-person pronouns to be non-clickbait, so I felt this would be a good feature to add.

v. *Contains second-person pronouns*

Similarly, most times when a headline included a second-person pronoun, it was clickbait, such as “Which K-Pop Song Should You Listen To Based On Your Birth Month” or “Which New Adele Song Are You”.

vi. *Contains “should”*

Another word that nearly exclusively appeared in clickbait titles was “should”. (e.g. “24 Potato Recipes That Should Be Illegal”)

vii. *Contains “?”*

viii. *Contains “!”*

ix. *Contains tweets*

Headlines discussing tweets were also seemingly exclusively clickbait. (ex. “43 Tweets About Grapes, The Fruit”, “23 Tweets From 2006 That Will Make You Feel Ancient”)

x. *About politics*

Headlines that discussed politics were most likely to be non-clickbait. For this feature I tried to target a handful of buzzwords that are usually only used in political contexts, such as “congress” and “governor”. (ex. “Utah Governor Chosen as Ambassador to China”, “US Congress may re-establish the Luxury Tax”)

I also considered the length of titles. I initially had an inclination that clickbait headlines tended to be shorter than those of non-clickbait ones. As I looked through the data, though, there were plenty of clickbait headlines which were longer than non-clickbait headlines, and some around the same lengths. Surprisingly, when I removed this feature the accuracy of my classification lowered. So, I decided to keep it. Other features I considered included proper names, verbs, and numbers. However, these were all features that were found in both clickbait and non-clickbait headlines alike.

Once I had all my features coded, I used sklearn to build a vectorizer and model based on my training data.

IV. *The Results*

After testing my model against the test data, I achieved an accuracy of about 68.91%. Looking at other models that were created to classify this data, I believe such features as the ones I used are not enough to determine if a headline is clickbait. There may be other features which I overlooked that may have helped, but the type of model has an impact on the results of the classification.

References

Anand, Aman. “Clickbait Dataset” Kaggle. https://www.kaggle.com/amananandrai/clickbait-dataset.