Capstone Project: Medium Article Analysis

By Arik Levy

Objective: Investigative approach of the determinants of medium article success

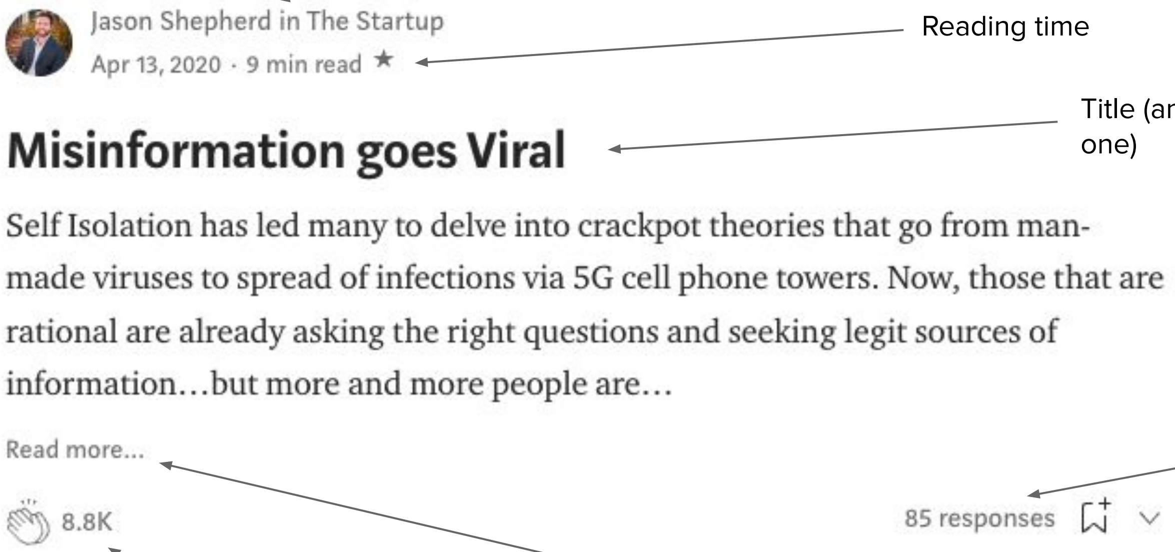
- Target: Claps (aka likes)
  - Classification task, achieved by splitting the amount of likes by median to create two classes.
  - Median Value for claps: 95 claps

- Questions:
  - What are the most important features of a good article?
  - Are there any particular topics/words that receive more claps?

WEB SCRAPE

Over 90 thousand articles scraped

The actual body of the text and the amount of followers the author had are scraped by extracting the article URL and using that as the URL for the embedded scraper.



Jason Shepherd in The Startup

Apr 13, 2020 · 9 min read ★

Misinformation goes Viral

Self Isolation has led many to delve into crackpot theories that go from man-made viruses to spread of infections via 5G cell phone towers. Now, those that are rational are already asking the right questions and seeking legit sources of information…but more and more people are…

Read more…

8.8K

85 responses

Author Handle

Reading time

Title (and subtitle if there is one)

Article URL

Number of Claps

Number of Comments

The date was also extracted through the for loop.

What information was added to this for the model?

- Publication Name
- Number of Words
- Subjectivity
  - How opinionated an article is.
- Polarity
  - Measures the positivity or negativity of the text.
- Day of the week - Extracted from the Date

And dropped:

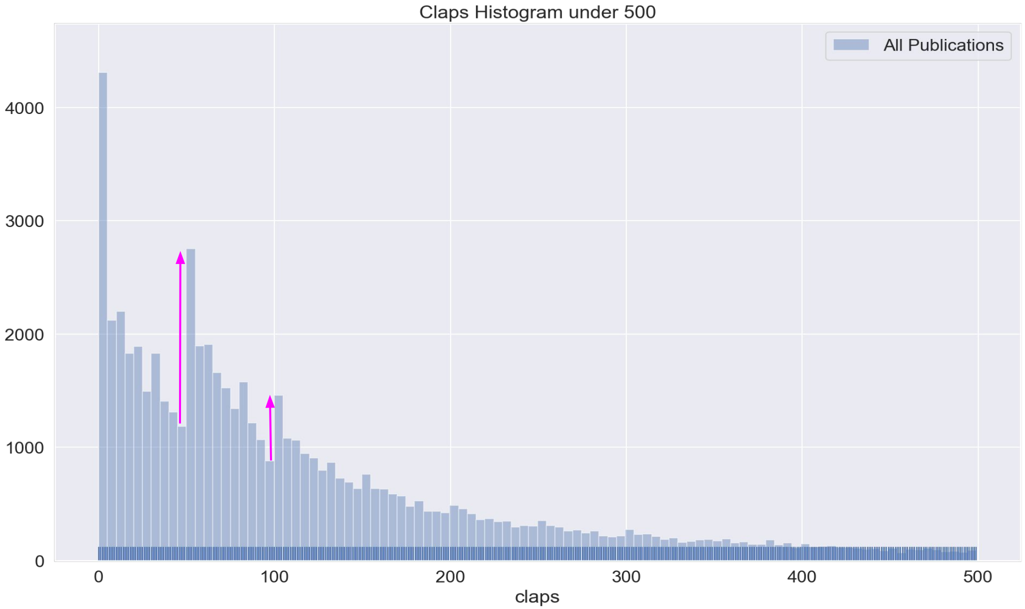
- Number of comments/responses
  - Multicollinearity being the main concern.

Exploratory Data Analysis

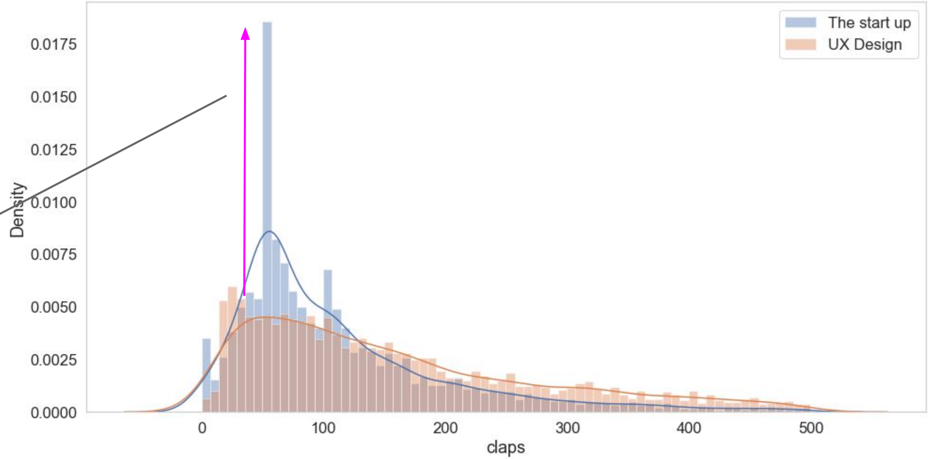
Interesting Trends - Looking at our target

- We can see the histogram has a consistent downward trend
- However there are two big spikes in the distribution that we can see at the 50 and 100 claps mark

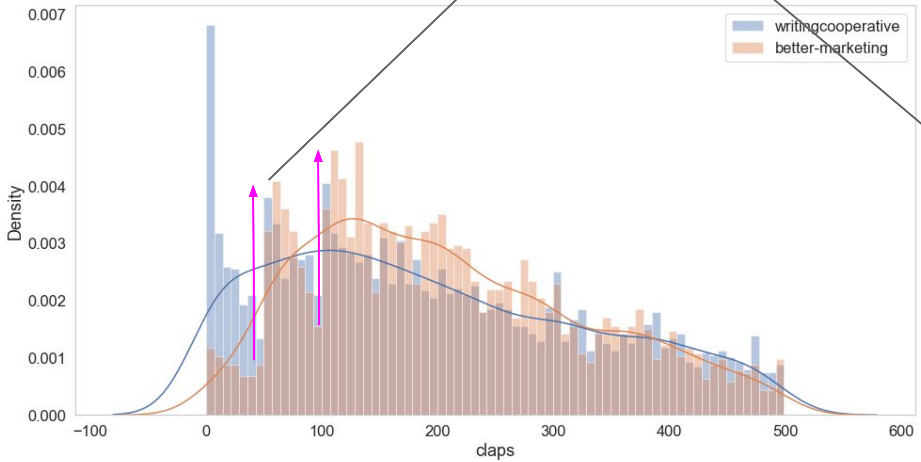
- Indicative that Medium publicizes and pushes out articles that reach a certain threshold

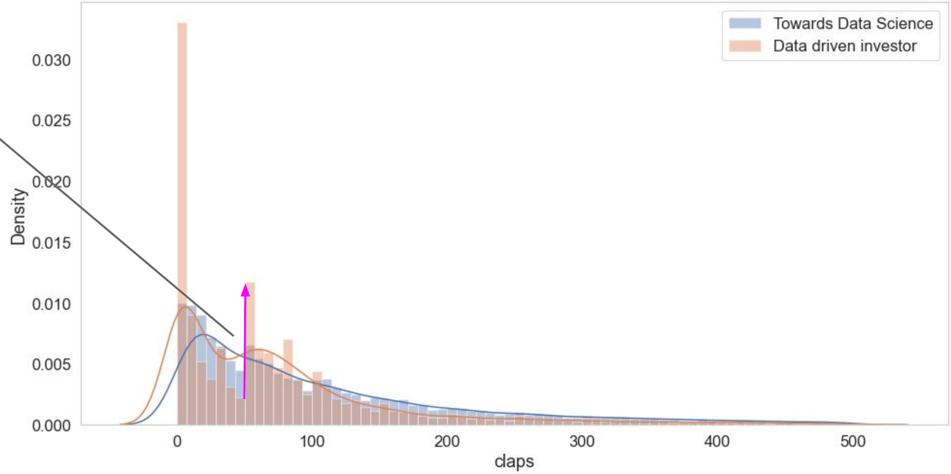


Same effect as seen above for each publication



Biggest density hikes



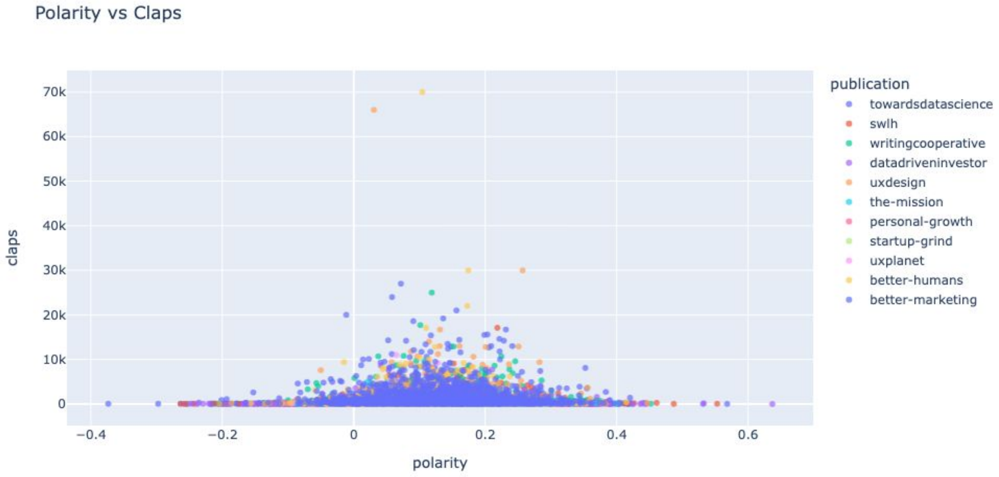


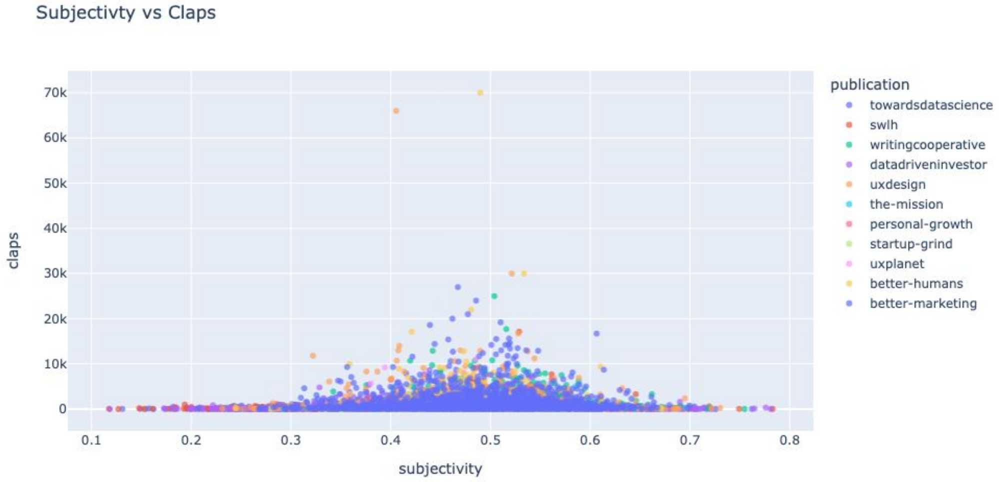
Subjectivity and polarity

It is very interesting that the best articles have moderate levels of both polarity and subjectivity with readers definitely penalizing extremes in both cases.

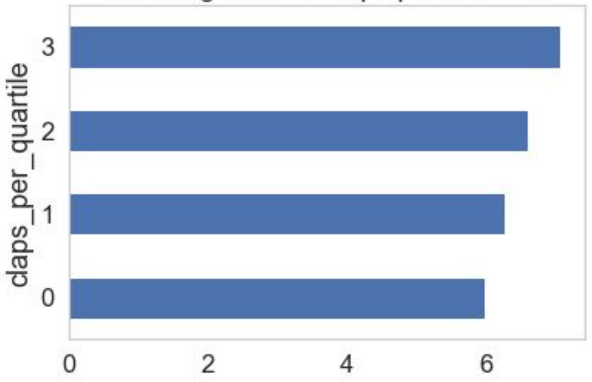
Polarizing articles are penalized with the optimal amount being from between 0 and 0.2.

The optimal level for subjectivity seems to be between 0.35 and 0.6.

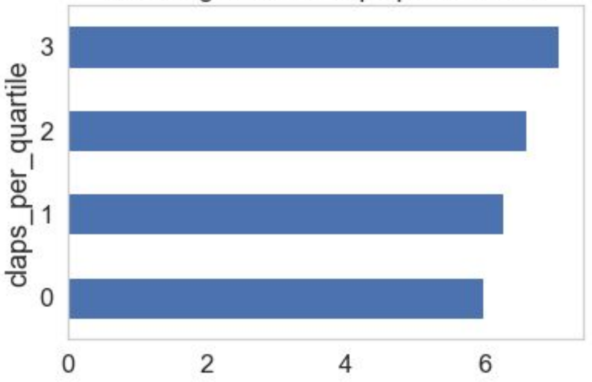




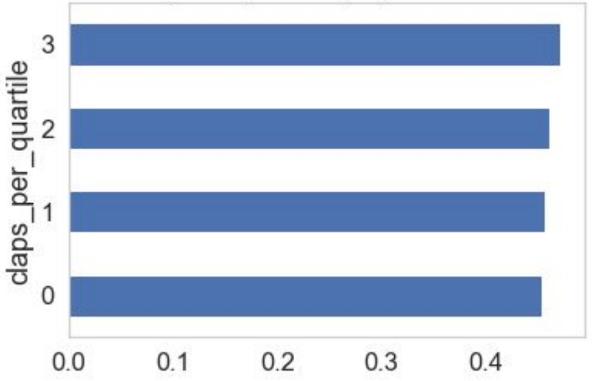
Claps by quartile analysis



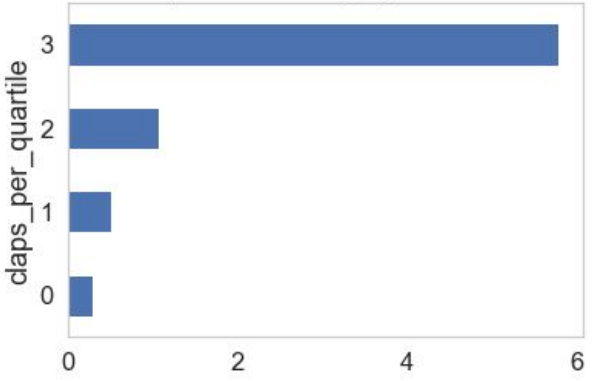
Reading time vs Claps per Quartile



Reading time vs Claps per Quartile

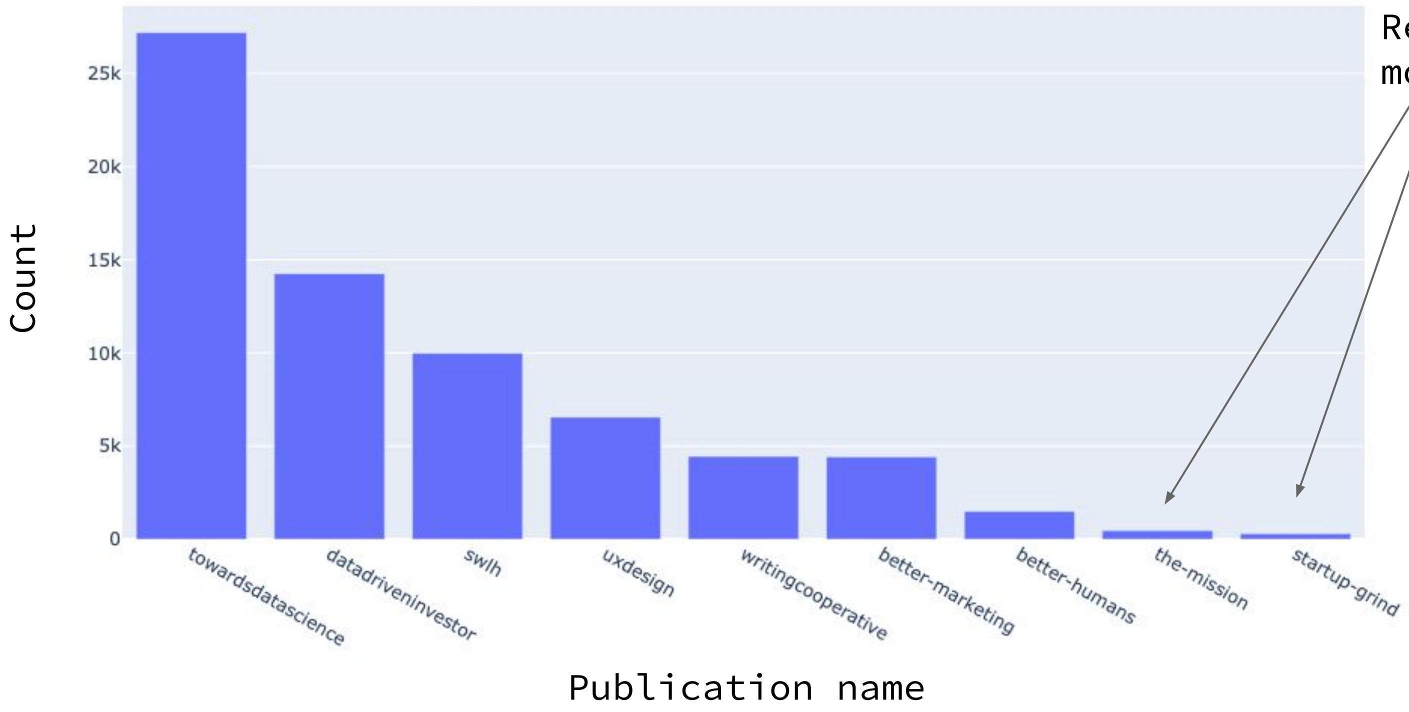


Subjectivty vs Claps per Quartile

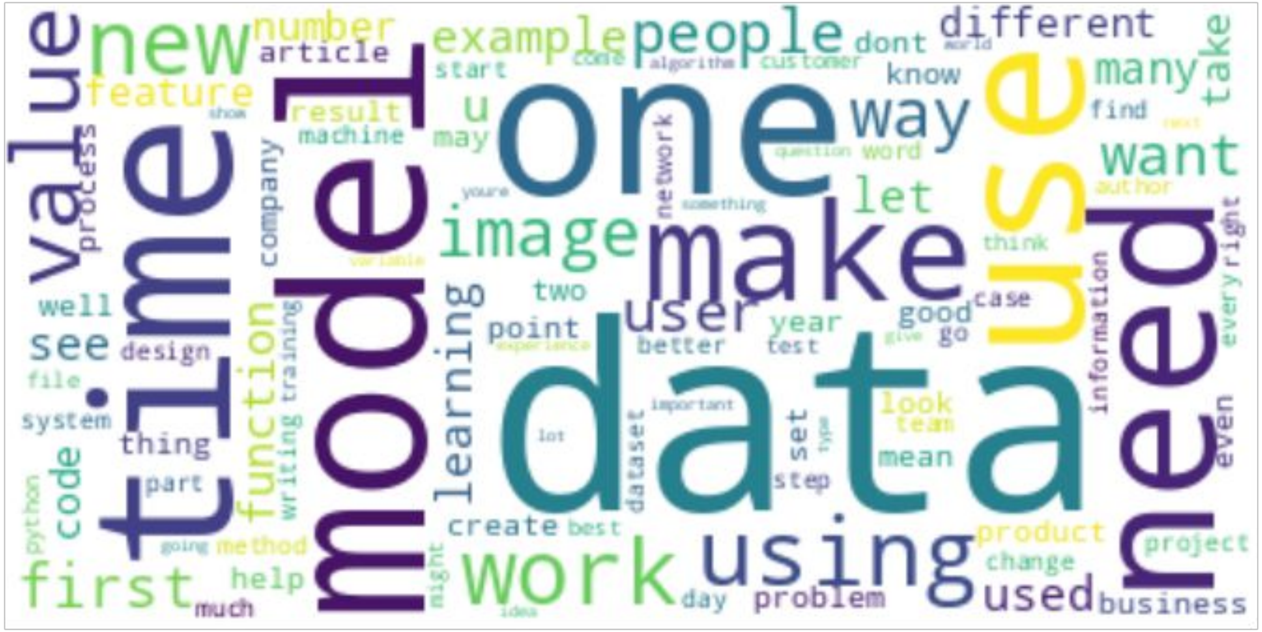


Responses vs Claps per Quartile

Small Caveat: amount of articles per publication



Removed from model



data
model
one
use
need
time
make
value
new
work
using
image
people
example
way
want
first
learning
function
user
different
number
feature
article
many
take
let
see
code
result
machine
company
process
dont
know
start
come
algorithm
customer
world
find
next
question
author
word
may
u
network
something
youre
show
think
variable
information
every
right
two
well
point
year
good
case
go
give
better
test
design
experience
file
writing
training
look
team
system
thing
important
lot
set
type
mean
dataset
even
part
step
create
best
product
project
python
going
method
change
help
might
day
problem
used
business
much
idea

Model Results

Models and pre processing

- I carried out countvectorizer on the title and the text features using n_grams 1 and 2
- Extracted other features from main dataset and created dummy variables where needed

- To be able to work with such big volumes of data I had to use sparse matrices

- Main Classification Models Used:
    - Logistic Regression
    - XGBoost

Results of the Classification - XGBoost

Predicted Value

True Value

	Class 1	Class 0
Class 1	7584	3853
Class 0	2956	8422

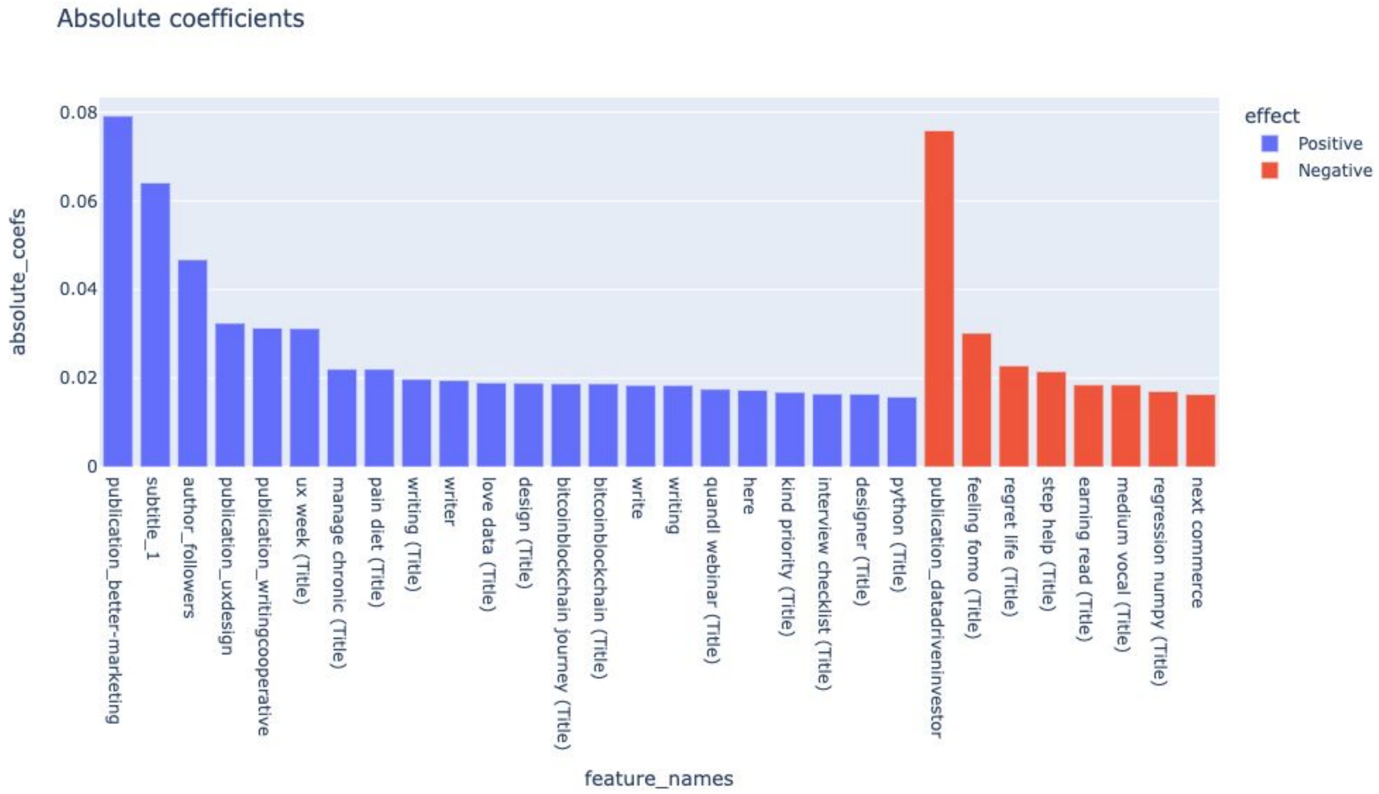
True positives and True Negatives i.e articles that have been correctly classified.

These have been incorrectly classified

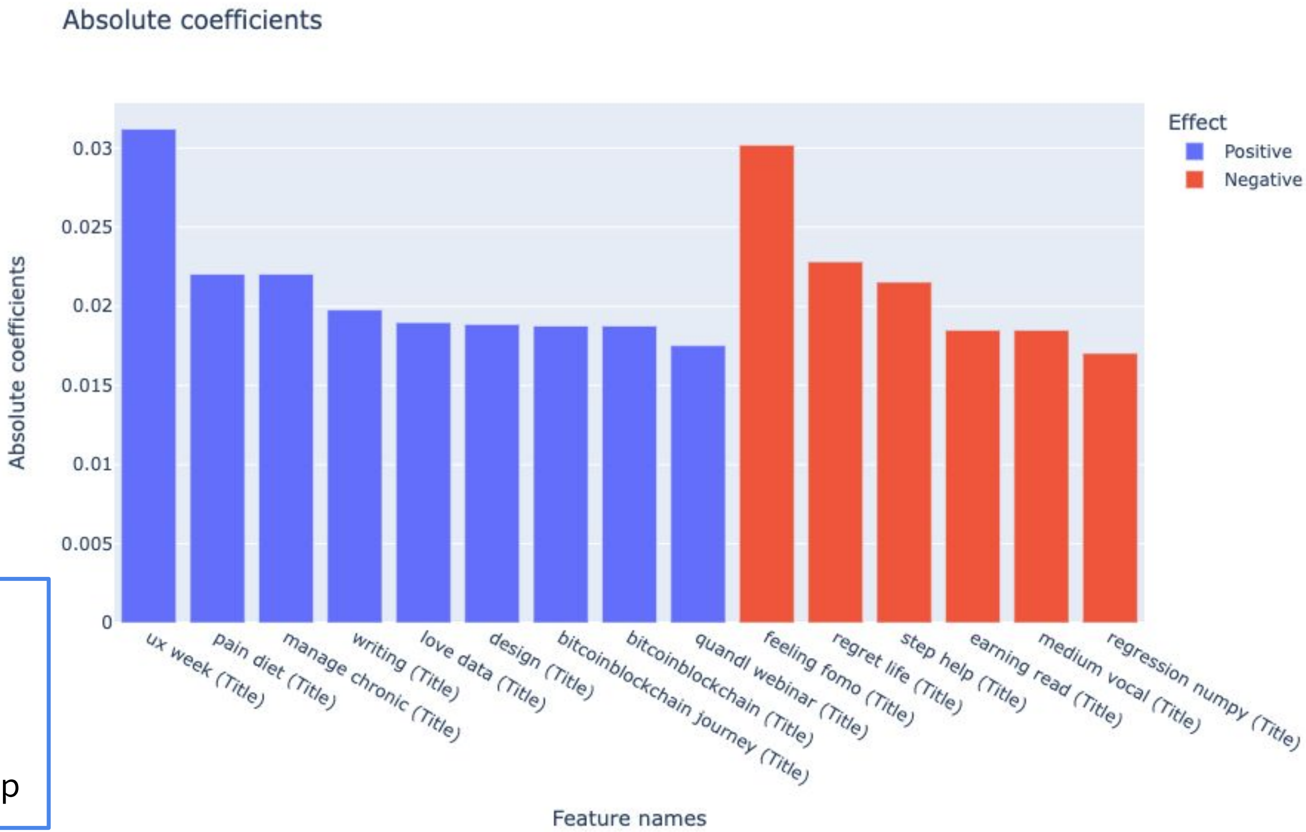
50% baseline score

70% accuracy score in best model

Coefficients: i.e What is causing the amount of claps

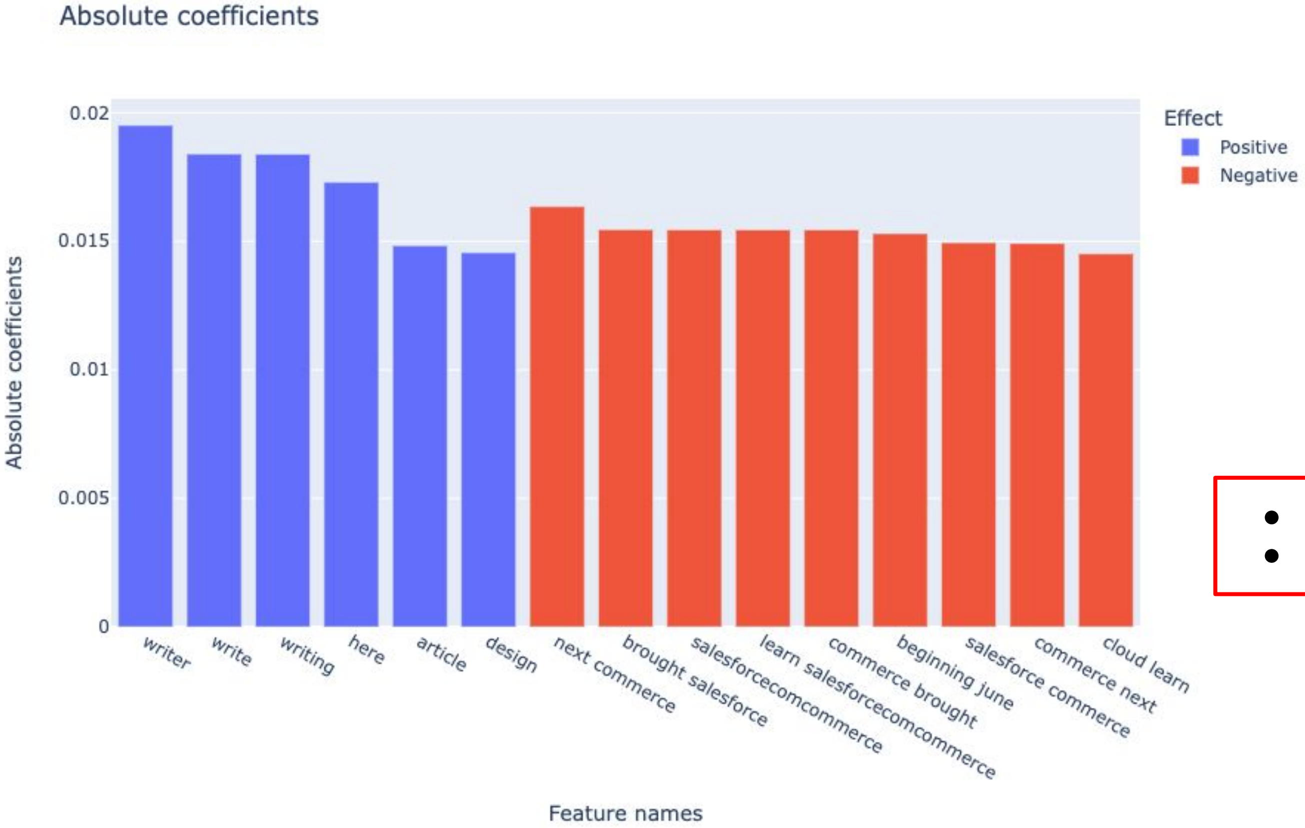


Coefficients for title



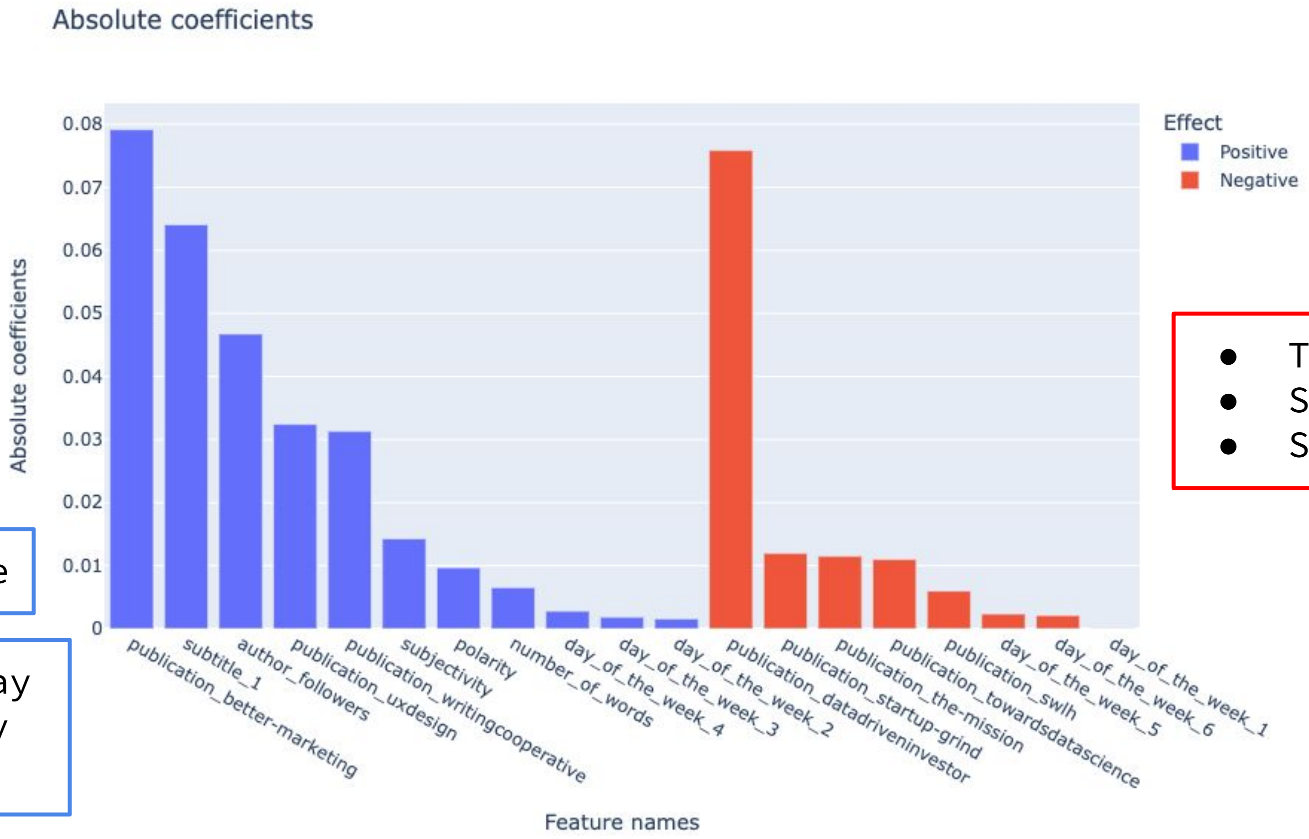
- Data
- Design
- Crypto
- Dieting
- Self-help

Coefficients for Text



- Salesforce
- Commerce

Coefficients for Rest of features



- Tuesday
- Saturday
- Sunday

- Subtitle

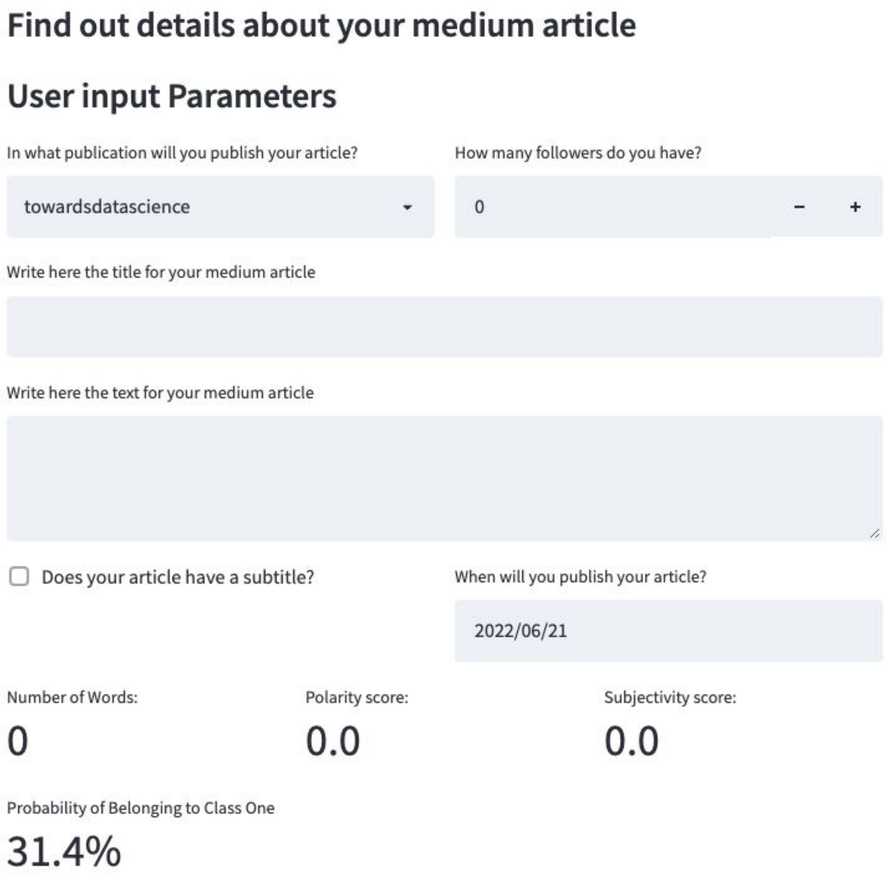
- Wednesday
- Thursday
- Friday

Model Deployment

Model Deployment

Here you can test out your own Medium articles and find out some cool information:

- What is your polarity score?
- How about subjectivity?
- Most importantly - What’s the probability of your article scoring higher than the Median (95 claps)?



Find out details about your medium article

User input Parameters

In what publication will you publish your article?

towardsdatascience

How many followers do you have?

0

Write here the title for your medium article

Write here the text for your medium article

□ Does your article have a subtitle?

When will you publish your article?

2022/06/21

Number of Words:

0

Polarity score:

0.0

Subjectivity score:

0.0

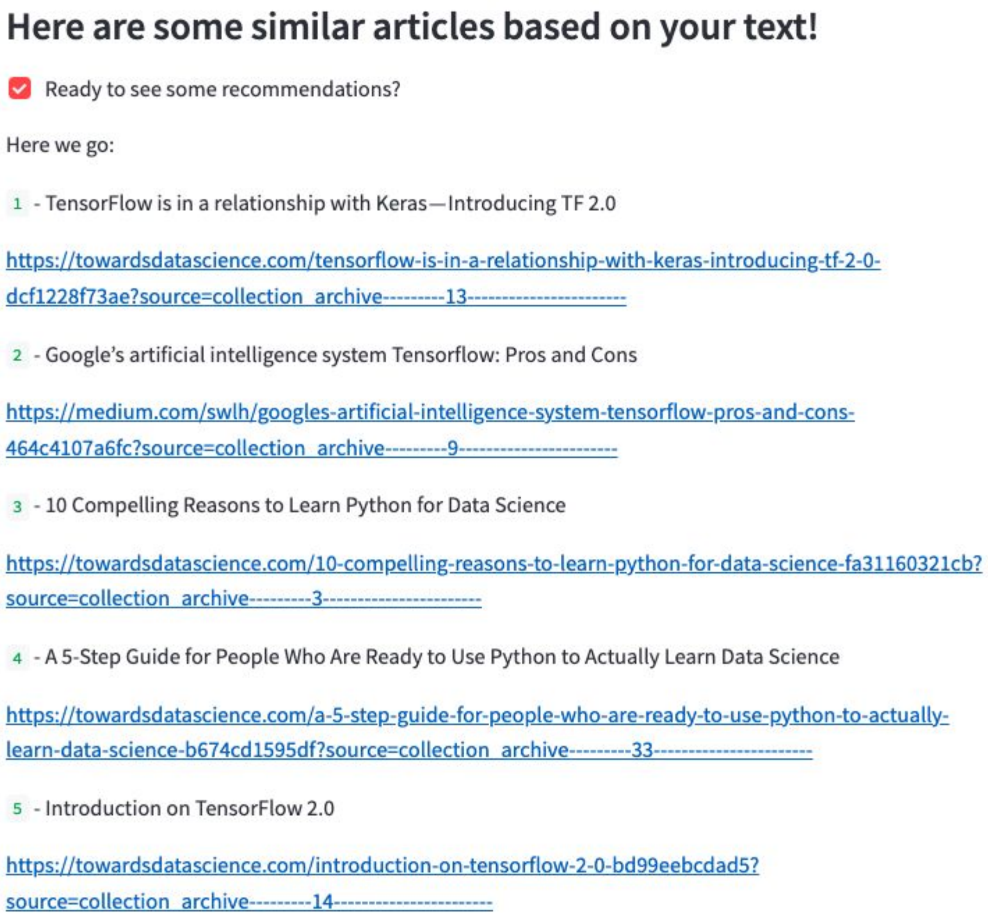
Probability of Belonging to Class One

31.4%

Article Recommendation

- Uses TF-IDF and cosine similarity of the article text to find similar articles!

- However it is limited to the articles I scraped.



Here are some similar articles based on your text!

☑ Ready to see some recommendations?

Here we go:

1 - TensorFlow is in a relationship with Keras—Introducing TF 2.0

https://towardsdatascience.com/tensorflow-is-in-a-relationship-with-keras-introducing-tf-2-0-dcf1228f73ae?source=collection_archive---------13-----------------------

2 - Google’s artificial intelligence system Tensorflow: Pros and Cons

https://medium.com/swlh/googles-artificial-intelligence-system-tensorflow-pros-and-cons-464c4107a6fc?source=collection_archive---------9-----------------------

3 - 10 Compelling Reasons to Learn Python for Data Science

https://towardsdatascience.com/10-compelling-reasons-to-learn-python-for-data-science-fa31160321cb?source=collection_archive---------3-----------------------

4 - A 5-Step Guide for People Who Are Ready to Use Python to Actually Learn Data Science

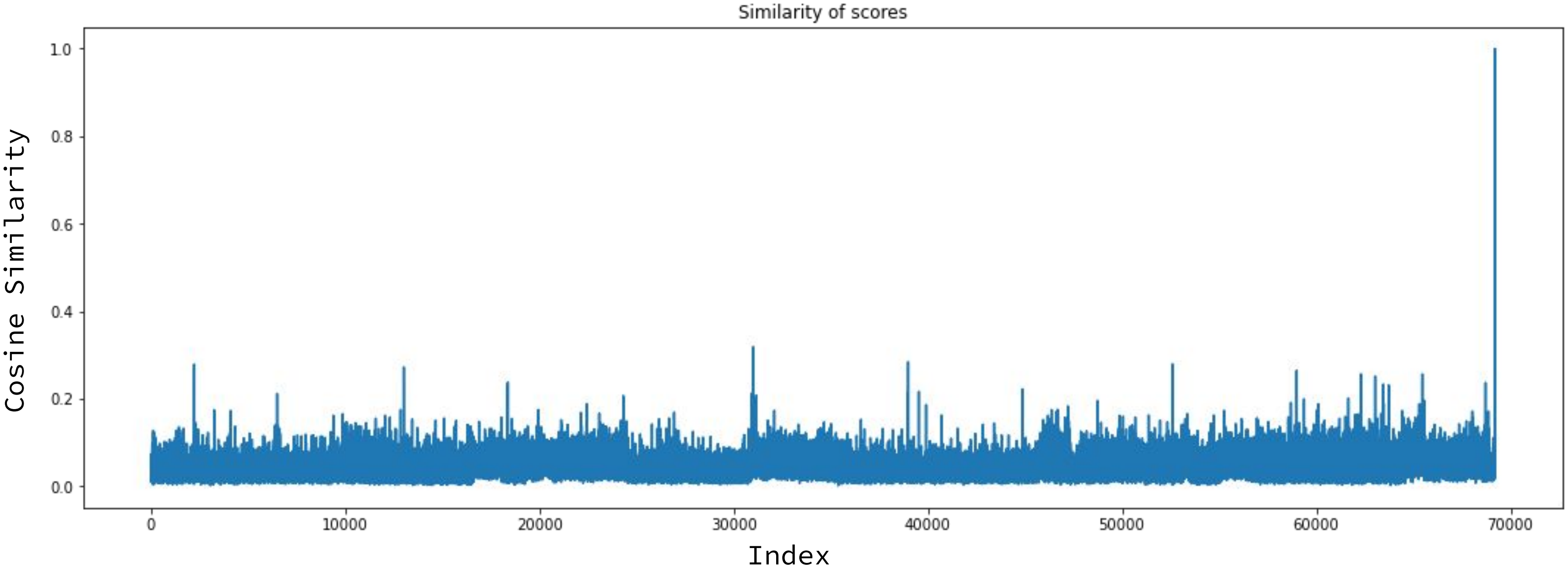
https://towardsdatascience.com/a-5-step-guide-for-people-who-are-ready-to-use-python-to-actually-learn-data-science-b674cd1595df?source=collection_archive---------33-----------------------

5 - Introduction on TensorFlow 2.0

https://towardsdatascience.com/introduction-on-tensorflow-2-0-bd99eebcdad5?source=collection_archive---------14-----------------------

How the recommender works

```
Cosine Similarity Scores:  array([0.05378748, 0.02409477, 0.03498535, ..., 1.        , 0.11315619,
                           0.05615479])
```



What Next?

Ideas for model improvement

- Amount of data
  - Scrape more data for the publications with less representation
  - Focus on a single publication or topic (using LDA)
- Elastic Net for Logistic Regression
- TF-IDF
  - For whole model & more specifically for single topic approach
- Better data cleaning
  - Part of speech tagging (POS) to further reduce the amount of features
  - Add amount of words in title as feature

Ideas to expand the project further

- Network Analysis by scraping users who actually liked each article
- Apply deep learning to the model to try to increase the predictive capability

Thank you!