Do hot topics get more claps?
Evaluating trends in *Towards Data Science*

Miguel Conner, Joule Voelz, Dominik Wielath

March 2023

Abstract

In this project, we scraped the website *Towards Data Science* to construct a novel dataset of all 55,000 articles published since the website's founding in 2016. Using LDA, we modeled 10 general categories of topics discussed on the website. Using the TF-IDF score, we charted the popularity of specific topics over time and extracted words associated with the "hot" topics for each month. Using linear and random forest regression, we observed a positive relationship between article length and engagement as measured in claps. Coefficients on an article's top topic were significant but difficult to interpret.

1 Introduction

Towards Data Science is an online publication that publishes daily articles by independent authors on a wide range of topics in data science, from introductory Python tutorials to writing on cutting-edge AI research. As the field of data science has exploded in popularity in the past decade, so has the blog. While *Towards Data Science* published only 42 articles in 2016, in 2021 it published over 12,000 – an average of over 30 articles per day. Thus *Towards Data Science* is a valuable resource not only for working and aspiring data scientists but for those seeking to understand the fast-paced development of the data science field in the last five years.

In this project, we compiled a novel dataset of text and metadata for all articles published on *Towards Data Science* in the years 2016-2022. We analyzed the dataset using tools including Latent Dirichlet Allocation (LDA) and TF-IDF to understand the changing share of topics over time. LDA was particularly successful at capturing the share of the publication devoted to big-picture topics, while the changing TF-IDF measures of certain terms charted the popularity of specific topics over time. Hoping to understand the popularity of individual articles as a function of the topics treated, we ran a fixed effects regression on several variables and found that article length and article readability are positively associated with engagement. Moreover, our project contributes a large, rich text dataset that can be further analyzed to gain insight into trends in data science.

2 Data

For this project, we scraped the archives of *Towards Data Science* to compile a novel data set of all articles published from the beginning of publication through 2022, over 55,000 articles in total. Each observation includes the article's URL, title, subtitle, number of sections, number of paragraphs, section titles, and full text. We also include measures of the article's engagement, including number of claps (likes) and the number of responses (comments) received. Before applying ML methods, we preprocess the text data by lemmatizing and removing stopwords.

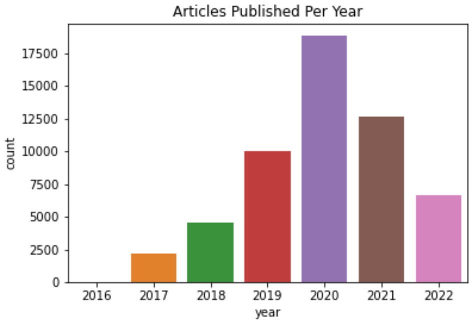


Figure 1: Number of articles published per year on *Towards Data Science*.

In addition, we calculate a measure of the article's reading complexity by using the Flesch Reading Ease score, which measures a text's readability on a scale from 1 to 100, with 100 being the most readable. A score of 30-50 corresponds to a difficult, college-graduate reading level, while a score of 60-70 indicates material that should be readable by high school students. As shown in Figure 2, the middle 50% of articles in *Towards Data Science* are between 51 and 64, meaning they are fairly difficult to read.

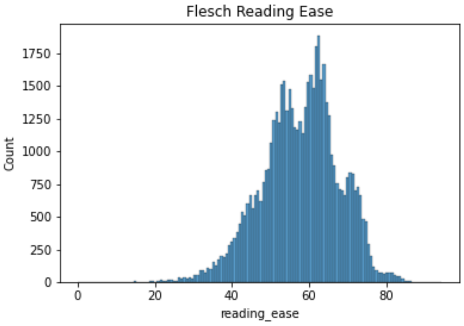


Figure 2: Number of articles of given difficulty levels, as measured by the Flesch Reading Ease score.

3 Methods

"Data science" is a catch-all term that encompasses many disciplines, from data engineering and business intelligence to state-of-the-art machine learning algorithms. Therefore we are interested in learning the different general and specific topics treated by *Towards Data Science* in order to assess changes in their relative importance over time. We experiment with different methods to do so.

3.1 Latent Dirichlet Allocation

We first employ the baseline topic modeling method of Latent Dirichlet Allocation (LDA), explored in Blei, Ng, and Jordan 2003 . LDA assumes that every document in the corpus (in this case, each article in the dataset) is assigned a proportion of topics, where these proportions are drawn from a Dirichlet distribution over K topics. Then, each word in each document is randomly assigned to a topic based on a categorical distribution. Once assigned a topic, the word is randomly drawn from the vocabulary of that certain topic, which has its own categorical distribution. More formally, for D documents with N_d words in each document and K possible topics over vocabulary V, we have:

$$\begin{aligned}
\Phi_{k=1\ldots K} &\sim Dirichlet_V(\beta) \\
\theta_{d=1\ldots K} &\sim Dirichlet_K(\alpha) \\
z_{d=1\ldots D,w=1\ldots N_d} &\sim Categorical_K(\theta_d) \\
w_{d=1\ldots D,w=1\ldots N_d} &\sim Categorical_V(\Phi_{z_{dw}})
\end{aligned}$$

Python packages for LDA learn the parameters of these distributions by using variational Bayes to approximate the posterior distribution of the parameters given the observed data. In our project, we implement LDA for ten topics over the corpus D of all articles produced 2016-2022. The most prevalent topic in each article is also recorded, to be used later on.

3.2 TF-IDF

We complement the big-picture topic modeling of LDA with a granular exploration of TF-IDF (term frequency-inverse document frequency) for specific topics within data science. TF-IDF quantifies the importance of a term in a certain document, relative to the importance of the term to the corpus as a whole.

$$tf(t,d) = count(t \in d) \tag{1}$$

$$tf(t,d) = \frac{count(t \in d)}{N_d} \tag{2}$$

$$idf(t,D) = \ln\left(\frac{M}{count(d \in D : t \in d)}\right) + 1 \tag{3}$$

The term frequency $tf(t, d)$ measures how many times the term t appears in document d. It can also be normalized by the length of document d, N_d. The inverse document frequency $idf(t, D)$ is the log of the inverse of the fraction of documents in corpus D that contain the term t. If the *idf* is large, then the word is very infrequent in the corpus.

Hence, the $tf(t, d) \times idf(t, D)$ increases as a term t is relatively more frequent in document d and less frequent in the corpus as a whole. It decreases if the term t is more frequent in the corpus as a whole. This means that if a term appears many times in one document but is found in almost all of the other documents in the corpus as a whole, it will not have a high $tf(t, d) \times idf(t, D)$. In this way, the $tf(t, d) \times idf(t, D)$ captures the specific terms t in document d that are "unique" to document d and hence more likely to be the focus of the document.

In our project, we are interested in the most important specific topics on *Towards Data Science* each month. Thus instead of treating each article as a document, we treat the titles of all articles published in one month as the document d, and the collection of all months as our corpus D. We calculate the $tf(t, d) \times idf(t, D)$ over all terms in each document of the corpus in order to capture the "hottest" topics for each month in years 2016-2022. Note that the inverse document frequency log equals zero for terms appearing in titles each month of our analysis. Adding one, as in the definition of the $idf(t, D)$ above, assures that even if a term appears in all documents, its $idf(t, D)$ score does not equal zero but rather one. This definition hinders us from identifying the most relevant terms for each month as terms such as "data" with a very high $tf(t, d)$ score that appear in titles every month are still getting high $tf(t, d) \times idf(t, D)$ scores and appear to be the most relevant in all periods. We exclude terms occurring in over 50 percent of all documents to capture only words significant for a specific month from calculating the $tf(t, d) \times idf(t, D)$ scores.“

3.3 Regression

We are interested in exploring the relationship between the topics explored in an article and the article's level of engagement as measured in claps. Are there certain topics that tend to get more engagement, conditional on other factors like article length? To explore this question, we run the following regression.

$$\begin{aligned} claps_{it(i)} &= \alpha + \beta length_i + \delta complexity_i + \mu dayssincerelease_i + \gamma_{t(i)} \\ &+ \sum_{k=1}^{K} \phi_k \mathbb{1}_{maxtopic_i=k} + \phi_5 count(w_{TOP5_{month(t(i))}} \in title_i) + \phi_{10} count(w_{TOP10_{month(t(i))}} \in title_i) + \varepsilon_{it(i)} \end{aligned}$$

The subscript i indicates article i, published on day $t(i)$. The first coefficient β captures the effect of the article length on engagement. In fact, we test several variables that capture article length, including the number of sections and paragraphs. δ captures the effect of textual complexity as measured by the Flesch Reading Ease score. Next, μ captures the linear effect of days since release on engagement. The thinking here is that in general, older articles may have more claps simply because they have been on the site longer. Or they may have fewer claps, as the site is more popular now than it was in 2017. The time fixed effect $\gamma_{t(i)}$ captures the fixed effects associated with each particular day of publication. This is important to include since there may be certain periods of more or less readership (for example, the beginning of COVID-19 when people had more times to read articles on the Internet) that are not captured by a linear time trend.

Moving on to the next row, ϕ_k captures the effect on claps when topic k is the topic with the highest share of words in article i. ϕ_5 captures the effect of having one of the month's top 5 words (by TF-IDF score) in the article title. ϕ_{10} captures the effect of having one of top words 6-10 in the title. $\varepsilon_{it(i)}$ is a noise term.

4 Results and Discussion

4.1 LDA

Topics were nicely categorized by LDA into many of the different styles of articles found on *Towards Data Science*. The top words for each topic are shown in Table 1. Topics like ”Deep Learning,“ ”Time Series,“ and ”NLP“ seem clearly tied to specific ML fields, while topics like ”My Journey“ and ”Business“ speak to project management and the business intelligence side of data science in a more narrative style. Figure 3 shows how the proportions of articles on each topic have changed over time.

My Journey:	science me re thing project don question scientist good ve
Business:	ai system business product user team company customer process ml
Neural Nets:	network layer training neural image function input deep neural_network loss
ML:	feature dataset prediction class training algorithm machine_learning tree test performance
Applied:	were year country game map day analysis player over show
Python:	file python run create command library package project user following
Prob/Stats:	distribution probability sample test random variable estimate equation hypothesis effect
Time Series:	plot variable point series dataset linear regression time_series cluster line
Preprocessing:	function column method list python row table type object graph
NLP:	image word text language task sentence vector transformer topic document

Table 1: Top 10 words for the 10 topics from the LDA analysis

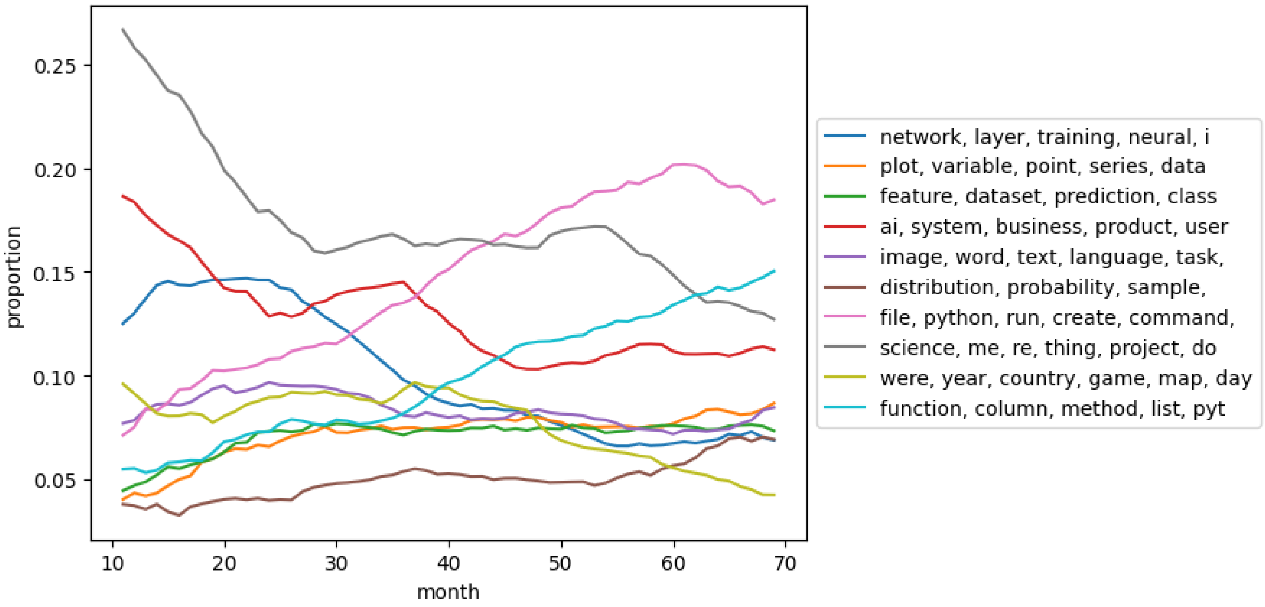
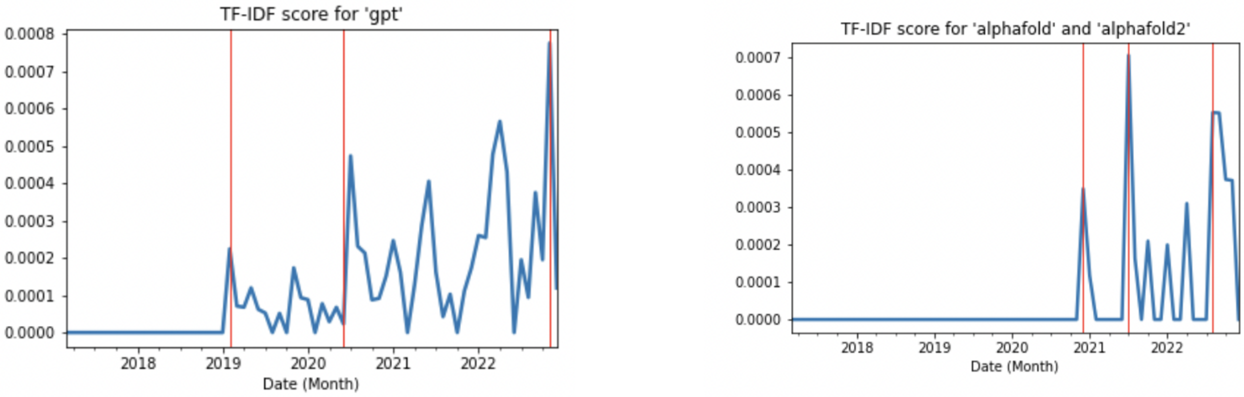


Figure 3: The proportion of articles written on a certain topic from 2017 to 2022, with a 12-month moving average to smooth out noise.

4.2 TF-IDF

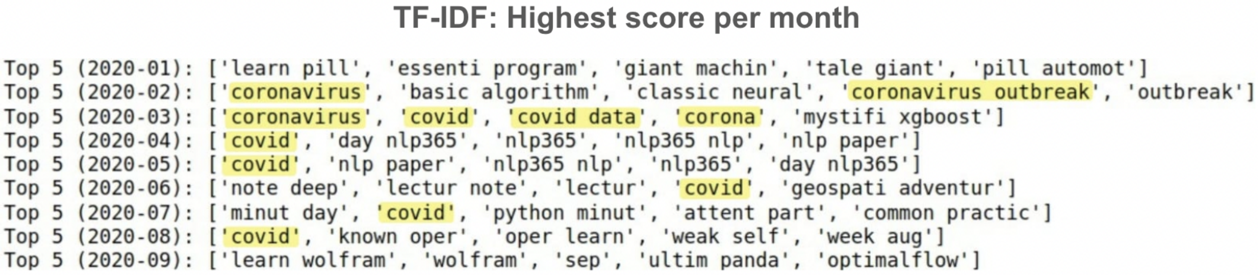
The TF-IDF approach was successful in determining popular topics for each month of publication. After some experimentation, we noticed it was necessary to normalize the TF-IDF by document length 2, which in our case was the length of all titles per month. Doing this allowed us use TF-IDF for two purposes. The first is to track the evolution of topic popularity over time. For example, plots of the TF-IDF score of ‘GPT’ and ‘AlphaFold’ (Figure 4) capture spikes that correspond to major breakthroughs regarding the NLP model and protein fold predictor, respectively (Deepmind 2022). The second use for TF-IDF is to identify ”hot topics“ for each month of publication. We do this by selecting the bigrams with top TF-IDF scores each month. This approach is validated by Figure 5, which shows the top 5 words by TF-IDF in most months of 2020 included ‘COVID’ or ’coronavirus.’ We use the method of choosing the top words in order to construct variables for our regression.



(a) GPT

(b) AlphaFold

Figure 4: TF-IDF captures monthly trends



```
TF-IDF: Highest score per month

Top 5 (2020-01): ['learn pill', 'essenti program', 'giant machin', 'tale giant', 'pill automot']
Top 5 (2020-02): ['coronavirus', 'basic algorithm', 'classic neural', 'coronavirus outbreak', 'outbreak']
Top 5 (2020-03): ['coronavirus', 'covid', 'covid data', 'corona', 'mystifi xgboost']
Top 5 (2020-04): ['covid', 'day nlp365', 'nlp365', 'nlp365 nlp', 'nlp paper']
Top 5 (2020-05): ['covid', 'nlp paper', 'nlp365 nlp', 'nlp365', 'day nlp365']
Top 5 (2020-06): ['note deep', 'lectur note', 'lectur', 'covid', 'geospati adventur']
Top 5 (2020-07): ['minut day', 'covid', 'python minut', 'attent part', 'common practic']
Top 5 (2020-08): ['covid', 'known oper', 'oper learn', 'weak self', 'week aug']
Top 5 (2020-09): ['learn wolfram', 'wolfram', 'sep', 'ultim panda', 'optimalflow']
```

Figure 5: TF-IDF captures hot topics

4.3 Regression

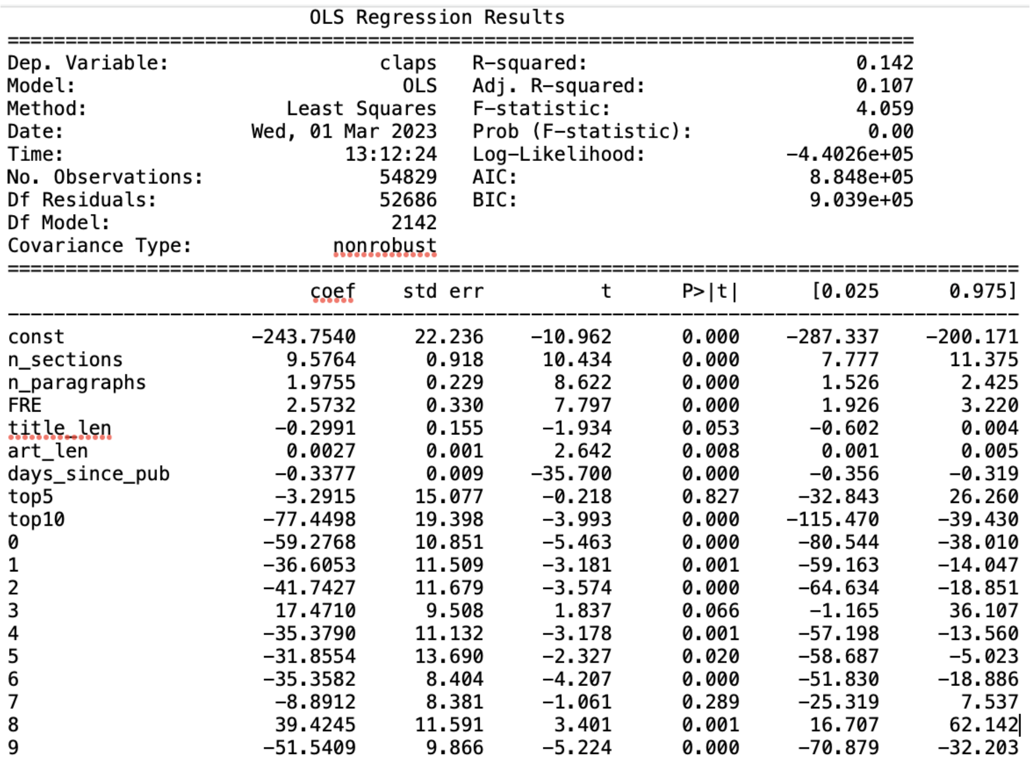
The results from the regression are shown in Figure 6. The R^2 value of 0.142 is not high, which is as we would expect it to be given there are many unaccounted factors that explain engagement.[1] Interestingly, most of the coefficients are significant at the 10% level or below.

First of all, coefficients on article length, number of sections, and number of paragraphs are all positive. This suggests that the more meaty or informative an article, the more likely it is that a reader will contribute a ”clap“ after reading an article. (It is important to note here that claps are not the same thing as views, which Medium does not publish. It is only possible to give a ”clap“ to an article by scrolling down to the end of the article. And therefore claps measure whether readers liked an article, not just read it.)

Second of all, the coefficient on reading ease is positive, which suggests that the clearer and easier to read an article is, the more readers will like it. The coefficient on days since publication is negative, which suggests that older articles have fewer claps. This might be because *Towards Data Science* was much less popular in 2016 and 2017, meaning that older articles received less engagement. They may even be about topics that are not as relevant anymore and thus appear less in search results.

Interestingly, coefficients on the presence of top words in the article title are negative. We interpret this result to mean that perhaps writing about a popular topic does not necessarily have a large impact on how much praise the article receives. After all, if an article is well written, this fact alone should be much more important than what the article is about. There may also be less reader interest in topics that are ”hot“ and thus already saturated. Coefficients on all LDA topics are negative except for the topics ”ML“ and ”Preprocessing.“ Recall that the variable for topic k is an indicator that indicates topic k is the top topic by proportion in that article. However, being a ”top topic“ is not necessarily synonymous with the whole article being about that topic, or even having the title include that topic. The top topic may only be 20% of the article, if the article has small shares of many topics.

[1] Including the number of responses (comments) an article received increased the R^2 value to 0.640. However, we chose not to include responses as responses and claps capture nearly the same thing.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  claps   R-squared:                       0.142
Model:                            OLS   Adj. R-squared:                  0.107
Method:                 Least Squares   F-statistic:                     4.059
Date:                Wed, 01 Mar 2023   Prob (F-statistic):               0.00
Time:                        13:12:24   Log-Likelihood:            -4.4026e+05
No. Observations:               54829   AIC:                         8.848e+05
Df Residuals:                   52686   BIC:                         9.039e+05
Df Model:                        2142
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                -243.7540     22.236    -10.962      0.000    -287.337    -200.171
n_sections              9.5764      0.918     10.434      0.000       7.777      11.375
n_paragraphs            1.9755      0.229      8.622      0.000       1.526       2.425
FRE                     2.5732      0.330      7.797      0.000       1.926       3.220
title_len              -0.2991      0.155     -1.934      0.053      -0.602       0.004
art_len                 0.0027      0.001      2.642      0.008       0.001       0.005
days_since_pub         -0.3377      0.009    -35.700      0.000      -0.356      -0.319
top5                   -3.2915     15.077     -0.218      0.827     -32.843      26.260
top10                 -77.4498     19.398     -3.993      0.000    -115.470     -39.430
0                     -59.2768     10.851     -5.463      0.000     -80.544     -38.010
1                     -36.6053     11.509     -3.181      0.001     -59.163     -14.047
2                     -41.7427     11.679     -3.574      0.000     -64.634     -18.851
3                      17.4710      9.508      1.837      0.066      -1.165      36.107
4                     -35.3790     11.132     -3.178      0.001     -57.198     -13.560
5                     -31.8554     13.690     -2.327      0.020     -58.687      -5.023
6                     -35.3582      8.404     -4.207      0.000     -51.830     -18.886
7                      -8.8912      8.381     -1.061      0.289     -25.319       7.537
8                      39.4245     11.591      3.401      0.001      16.707      62.142
9                     -51.5409      9.866     -5.224      0.000     -70.879     -32.203
```

Figure 6: Regression results

4.4 Random Forest

In addition to linear regression, we ran a random forest regression using cross-validation and grid search. Rather than include a fixed effect at the day level (which would take a really long time to run with random forest), we one hot encoded by year. We then split our data 80/20 for a test-train split and achieved a R^2 value of 0.10 on the test set with an MSE of 586, 580.75. The most important features according to our random forest model are shown in Figure 7. Notice that the variable ”days since publication“ has a very high importance, whereas it did not in the linear regression. This is probably due to the fact that we cannot include time fixed effects in a random forest. Hence, days since publication is capturing much of the variation that was captured in the fixed effects.

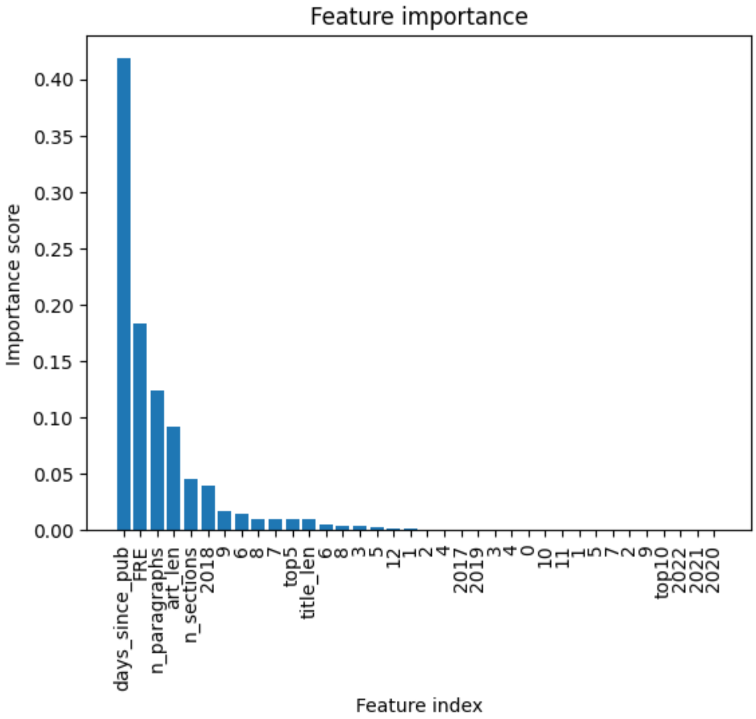


Figure 7: Feature importance from random forest model.

5 Conclusion

In this project, we scraped the website *Towards Data Science* to construct a novel dataset of all 55,000 articles published since the website's founding in 2016. Using LDA, we modeled 10 general categories of topics discussed on the website, which corresponded nicely to recognizable subfields of machine learning, data science, and business intelligence. We demonstrated that TF-IDF can be used in order to chart the popularity of specific topics over time, as well as to discover the ”hot“ topics for each month. Using linear regression and random forest regression, we observed a positive relationship between article length and engagement as measured in claps. Easier readability was also associated with more claps. Coefficients on ”hot“ topic words were negative. Coefficients on an article's top topic were significant but difficult to interpret. The random forest regression did not improve upon linear regression, perhaps due to the absence of time fixed effects.

Further work with this data could improve upon the LDA results to create a more specific, interpretable measure of an article's main topic. Language complexity analysis could also be improved by creating a complexity score tailored to the technical language prevalent in *Towards Data Science* articles. Additional scraping could add more metadata on the article's author and explore the relationship of the author to the article's popularity. In short, there is much more to discover in this large text dataset.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). ‘Latent dirichlet allocation’. In: *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435. DOI: `http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993`. URL: `http://portal.acm.org/citation.cfm?id=944937`.

Deepmind (2022). *Deepmind AlphaFold - Timeline of a breakthrough.* `https://www.deepmind.com/research/highlighted-research/alphafold/timeline-of-a-breakthrough`. Accessed: 2023-02-20.