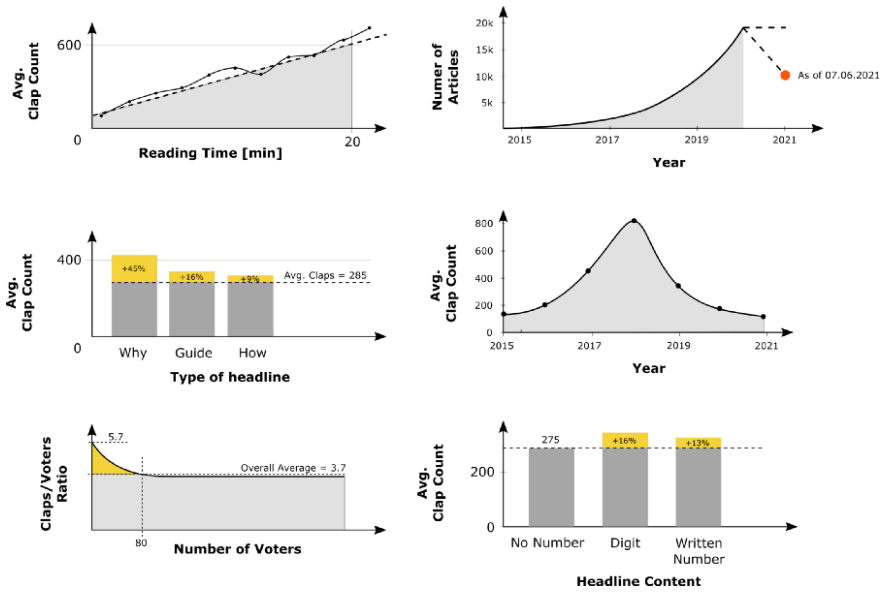
What makes an Article popular on Towards Data Science? An Analysis of 44k+ TDS Articles



Summary TDS Articles analysis — Image by the author

Table of Content

. . .

1. Data Collection

I don't want to go too deep here, as there are already a number of articles that deal with web scraping of Medium articles. A backup with the data of around 44 000 articles can be found on Kaggle.

What I would like to briefly discuss is what data I thought is important and gives a hint on how successful an article will perform. Beside giving an answer to some basic statistical questions, I want to make an attempt to find explanations for the performance of the articles.

Mike Sall`s conclusion in "When is the best time to publish? Wrong question." was:

> *What really matters is writing good stuff and, in turn, building a following.*

This sounds like a very plausible insight to me. But what characterizes the "good stuff". And more importantly, can we map these characteristics into features in a data set?

The challenges are probably not only the data collection, but also the way in which articles differ. A beginner tutorial would certainly be measured by other aspects than a scientific article or experience reports.

But let's start from the beginning, let's start with the basic statistics about TDS articles, like:

- **How many articles** have been published on TDS to date? And since when?
- **How many Claps** do they usually get?
- **How often does a user usually clap?** (Claps/Voters relation)

To identify possible performance drivers, I analysed the correlation between **performance (Claps and Voters)** and the following characteristics:

1. **Length (Minutes of read):** e.g. Do longer articles get more claps?

2. **Headlines:** e.g. What characteristics of a headline result in a higher performance: Length, Numbers, Type of header
3. **Visual Content:** e.g. Do Articles with self-made images perform better?

To be able to describe the above characteristics with data, I collected the following information from (all) TDS articles:

Table Articles

- Title, Headline
- URL
- Reading Time
- Publication Date
- Archive
- Author
- Clap_Count (Number of Claps)
- Voter_Count (Number of Users clapped)
- Response_Count

Table Figcaptions

- Caption
- (Article ID)

. . .

2. Analysis

The data set used for the analysis contains:

- Data from **44118 TDS articles** with **211521 figures** (from **38603 articles, 5515 articles** haven't used images)

In the sections below, I have addressed the following questions:

- How many Articles published in recent years? And how did they perform?
- Do longer articles get more claps?
- How often does a user clap?
- Do Articles with self-made images perform better?
- Does the type and length of the headline influence the performance of the article?

. . .

How many articles published in recent years?

First of all, I dealt with the most basic questions: How many articles had been published via TDS and how did they perform.

According to available data set, the "oldest" article in TDS was published on 21 November 2010. Nevertheless, only a handful of articles can be attributed to the years from 2010 to 2014. It is also not possible to tell whether these have already been published previously and subsequently added to TDS.

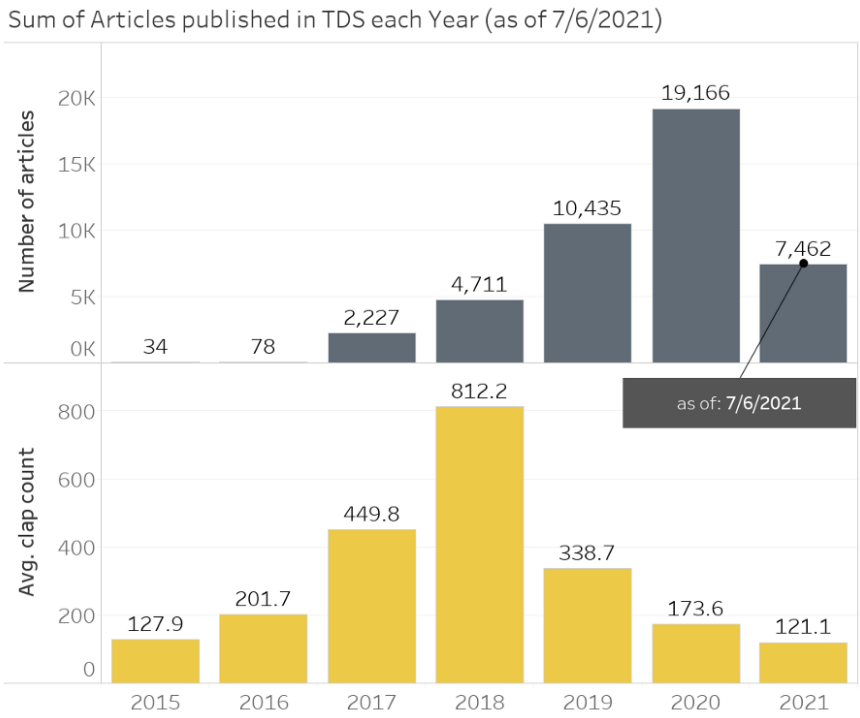
> *The number of Articles published has more than doubled each year*

After 2016 the number of articles published per year increased sharply. There is roughly an exponential increase over the years 2015 to 2020. A doubling in 2021 is no longer to be expected, since "only" 7000 articles were published in the first 5 months of 2021.

> *Articles published in 2018 performed extraordinary well*

It gets interesting looking at the average clap count over the years. **Articles published on TDS in 2018 were rated significantly higher** than in previous and subsequent years—more precisely, there was a significant increase in the years from 2015 to 2018 and a sharp decrease in the following years. The 4700 articles published in 2018 reached an average ***Clap Count*** (Number of Claps) of **more than 800 per article.**

In general, it can be said that it has **become relatively hard to reach several hundred Claps per article,** while it is usual for articles published in 2018.



History of articles published in TDS – Image by the author

. . .

Do longer articles get more claps?

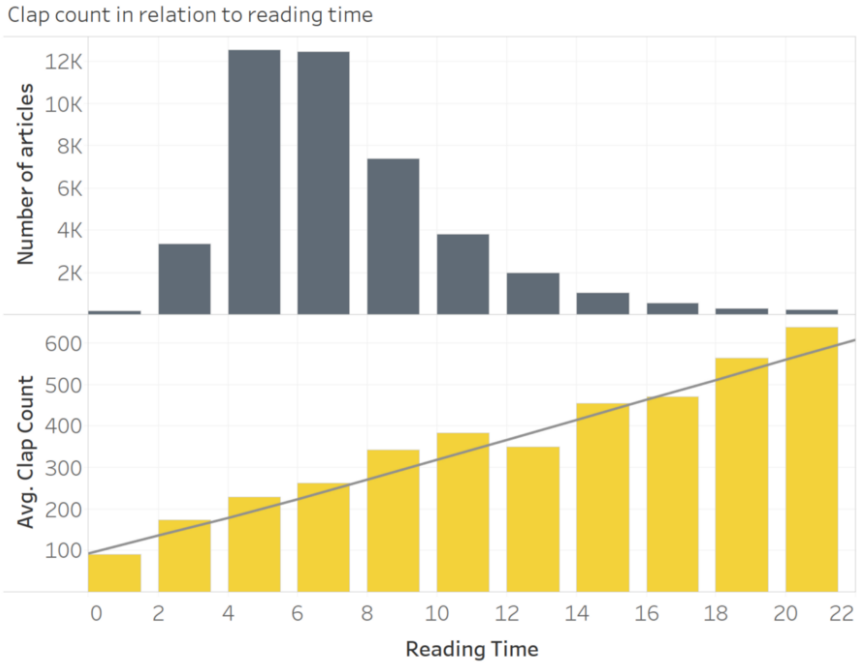
As shown in the graph below, most of the articles have a length of about 4–8 minutes read. But is it worth investing a little more time in writing longer articles?

> *The average number of claps increases with reading time*

It actually looks like longer articles tend to get more claps. In general, the clap count increases with the reading time, but a local maximum can be identified for articles with a length of 8–12 minutes. After that, the Clap Count initially drops a bit, before it goes up again.

That longer article perform better is probably not necessarily related to the words written, but more to the fact that the author of a longer

article has in most cases dealt with the topic more intensively.



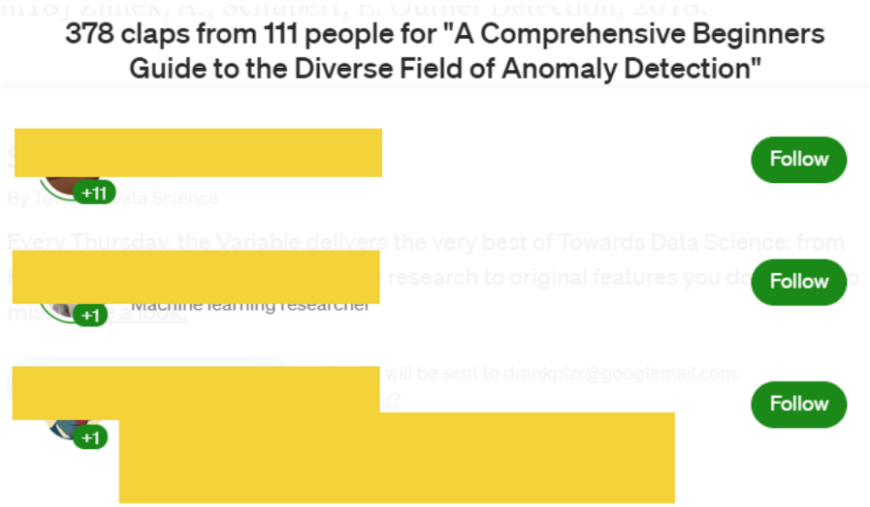
Avg. Clap Count over Reading Time—Image by the author

. . .

How often does a user clap?

What always confused me was the fact, that one user can clap up to 50 times on Medium. So what does a clap count of 50 or 3k actually mean? How many users clapped for the story?

To find an answer to this, I collected not only the number of claps from each article, but also the number of users behind these claps (the number of ***voters***).



Difference between Voters and Claps (here: 111 Voters and 378 Claps) — Screenshot by author

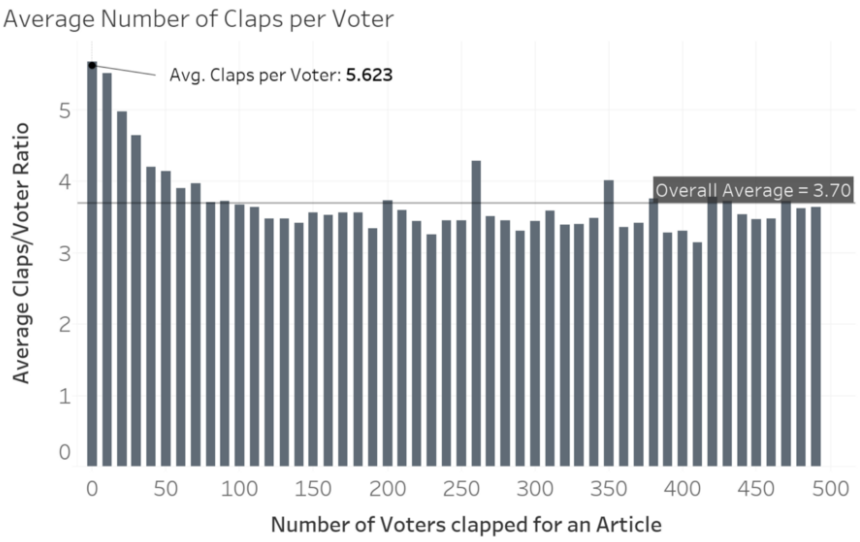
What I am interested in exactly is the ***Claps to Voters Ratio***, for each article. The following figure shows the distribution of the ***Claps/Voters Ratio*** for all TDS articles. The **maximum limit of claps per user is 50.** 31 Articles actually reach this maximum ***Claps/Voters Ratio***.

> *The average ratio is 3.7—so on average each voter claps around 3–4 times*

Whether this ratio changes depending on the number of claps, I wanted to investigate with the following graph. What can be seen, is that articles with a small number of voters, have a **higher *Claps/Voters Ratio***.

For example an article with **50 claps** needs on average **about 9 voters** to reach it (on average, each voter claps **more than 5 times**).

To reach a clap count of around **500**, you usually need around **130 voters.** (**ca. 3–4 claps per voter**)



Average Number of Claps per Voter—Image by the author

. . .

Do Articles with self-made images perform better?

Creating graphs, images and diagrams yourself can be time-consuming. But does the effort pay off?

Therefore, I collected all image captions from each article. In total the **44118 TDS articles** contain **211521 images**. So in average, each article is using around 5 images.

You can see a more detailed breakdown in the chart below. According to the graph below, about **53% of all articles contain less than 4 images, 21% 4–6 figures and only 25% of all articles using more than 6 figures** to visualize the content.

I also tried to assign the pictures to a source on the basis of the caption, by searching for a series of keywords. I was able to assign about 57% of the images to a specific source type (for the other 43% the source is unclear).

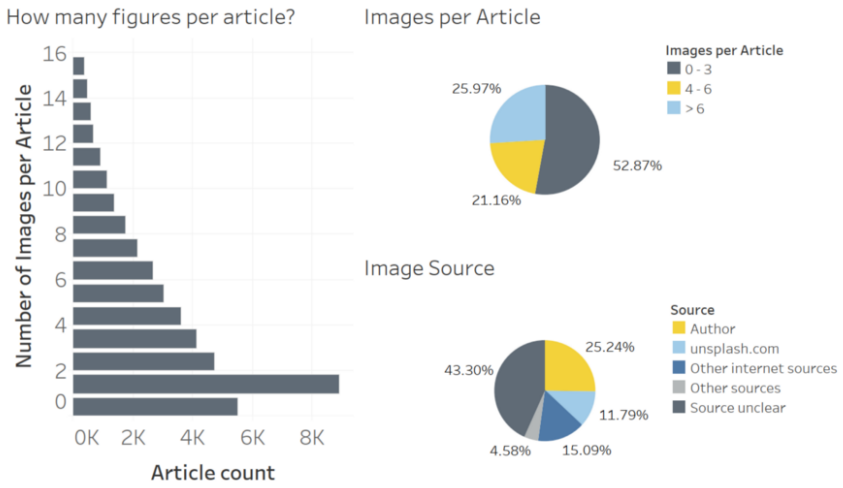


Image count and sources used in TDS articles—Image by the author

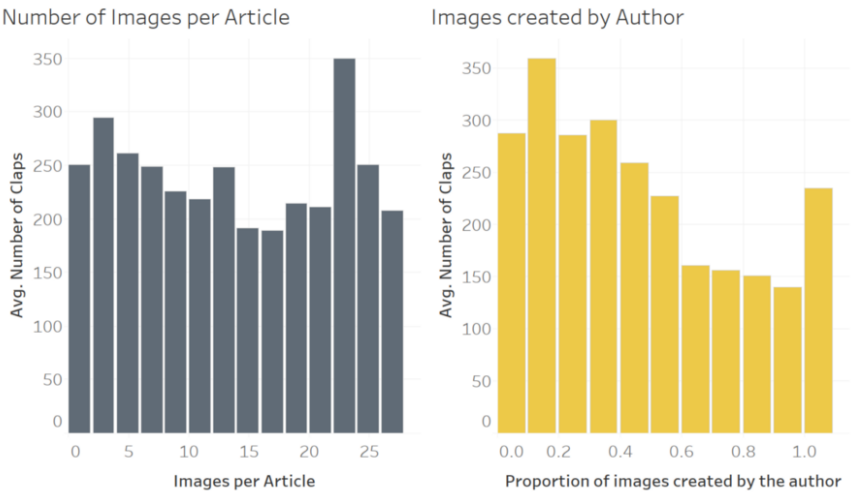
About a quarter of the images were created by the author himself, 32% come from various (internet) sources—about 12% are images from Unsplash.

With this labelling I would like to check whether the source of the images used has an influence on the performance of the article.

> *Or simply put: Is it worthwhile to create illustrations/graphs yourself?*

In the graph on the left, I have shown the relationship between the ***number of figures used*** in one article and the ***Clap Count***. On the right-hand side, the clap count over the proportion of images created by the author **(= (figures created by author) / (total number of figures))**.

The result is a bit disappointing. Neither the number of images used nor the source of the images seems to have a direct/clear influence on the performance. Other factors seem to have a greater influence.



Claps over Image Count per Article and Proportion of Self-Made images — Image by the author

. . .

What characteristics make a good headline?

That the headline has an influence on the number of clicks is considered proven. But what is a good headline? Are there some simple rules like using Capital Letters, Numbers or does the length play a major role?

I identified the following characteristics, which I used for correlation analysis with the ***Clap Count***:

1. **Headline Type** (Question, How-To, Guides, …)
2. **Numbers** used in Headline (e.g. *7 of the Most Used Regression Algorithms*)
3. **Year** used in Headline (e.g. *Hand-picked Resources for learning Deep Learning in 2020*)
4. **Headline Length** (20,50 or 150 characters?)

. . .

Headline Type

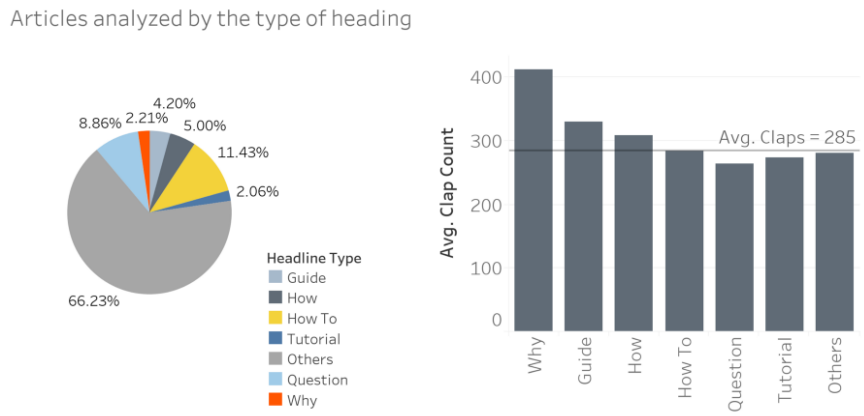
In the first step, I would like to find out which type of headline might perform better than others. Admittedly, this is not just about the

heading, the type of heading already reveals a lot about the article itself.

I've classified the headings into the following groups:

1. **How-to** (e.g. *How To Avoid Writing Sloppy SQL*)
2. **Guide** (e.g. *A Product Manager's Guide to Machine Learning: Balanced Scorecard*)
3. **Tutorials** (e.g. *A Tutorial On Creating Data Pipeline For Object Detection Using PyTorch And…*)
4. **How** (e.g. *How Artificial Intelligence is Transforming Finance and Banking*)
5. **Questions** (e.g. *Should I learn Julia?*)
6. **Why** (e.g. *Why Data Science might just not be worth it*)

Around 34% of the headings could be assigned to one of these groups. The graph on the right-hand side shows the correlation analysis between **headline type** and the **clap count**. Headlines of the "Why" and "How" type, which usually give an answer to a specific question, clearly perform above average. The same applies to guides.



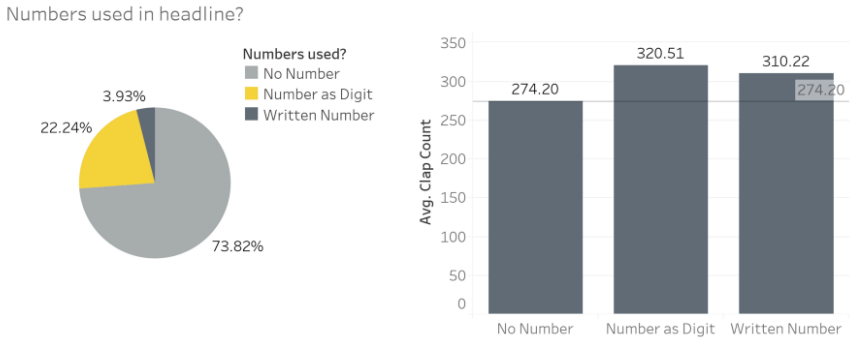
Headline Types used and Clap Count over Headline Type— Image by the author

. . .

Numbers in headlines

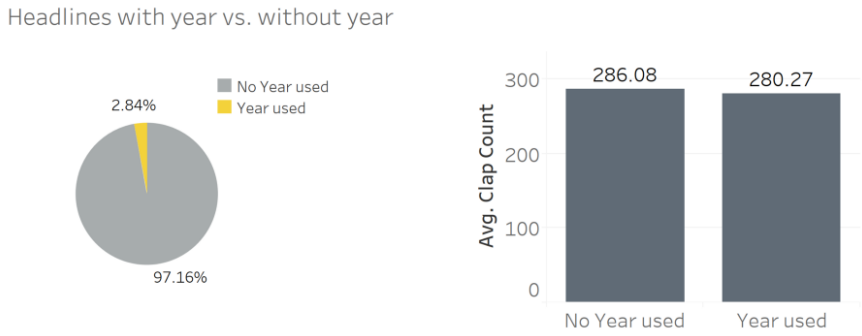
Various articles on Medium recommend using numbers in the headline to increase the probability that user click on your article. I am assuming that this should be reflected in the ***number of claps*** for the TDS articles as well.

And indeed, articles with a numbers in the headline, perform above average – the ***Clap Count*** is around 16 % higher.



Clap Count depending on whether a number was used in the heading—Image by the author

I also noticed that a lot articles use ***year numbers*** in their Headline (e.g. *Data Science Tools you should learn in 2021*). **However, the available data do not show any positive impact.**



Clap Count depending on whether a year was used in the heading—Image by the author

. . .

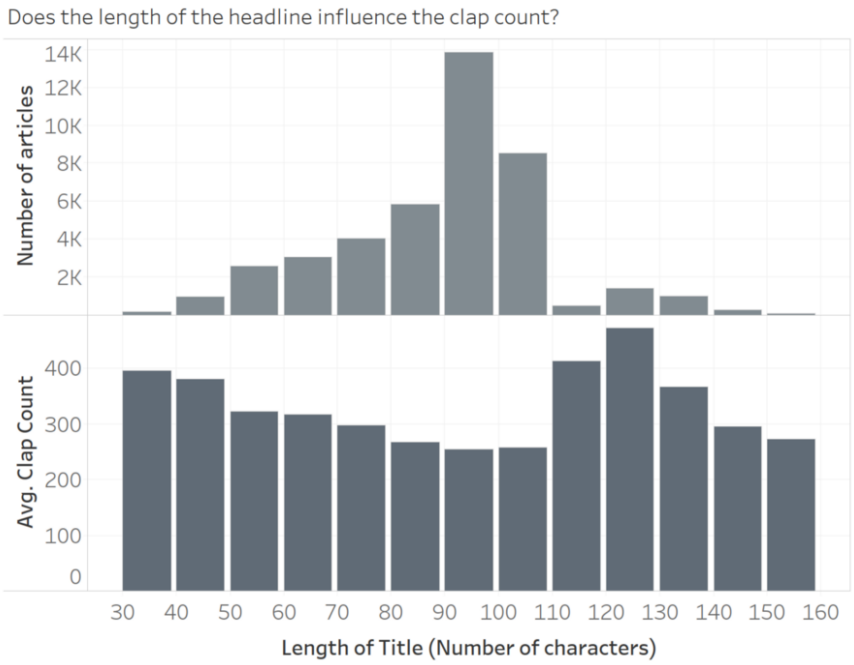
Length of the headline

I often read that headlines should be short and neat. But does that really have an impact on performance?

The majority of the headings have a length of 50–110 characters. Headlines with a length of more than 110 characters, on the other hand, are used less.

Up to a length of 110, there is a clear drop in the average ***Clap Count***. So for the majority of the articles the thesis of the short headlines seems to apply. However, for headlines that are even longer, the clap count increases again.

I personally assume that artificially lengthening the headline without gaining additional information is rather counterproductive.



Clap Count over Length of Title — Image by the author

. . .

Summary

I have again summarised the most important findings:

- The number of TDS articles published has grown exponentially in recent years, with this growth unlikely to continue in 2021.

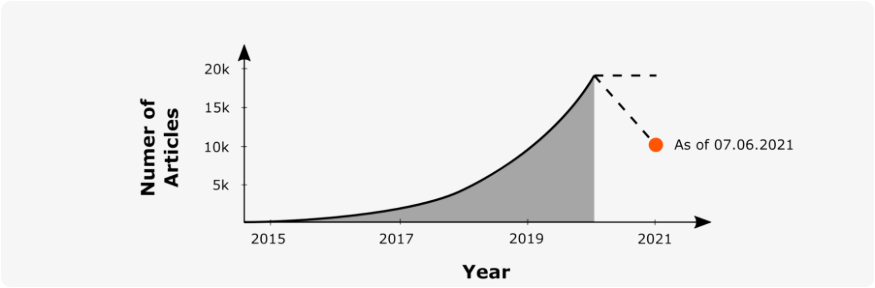
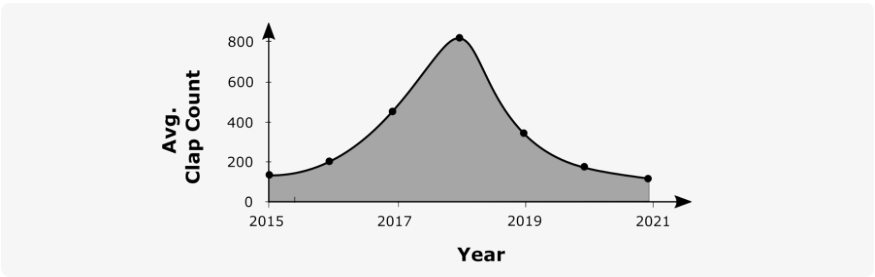


Image by the author

- Articles could achieve strongly different clap counts depending on the year they were published—articles in 2018 stand out exceptionally here.



- The average Clap Count generally increases with the reading time. However, a straight line can only roughly approximate this increase, in reality there are several local minima and maxima

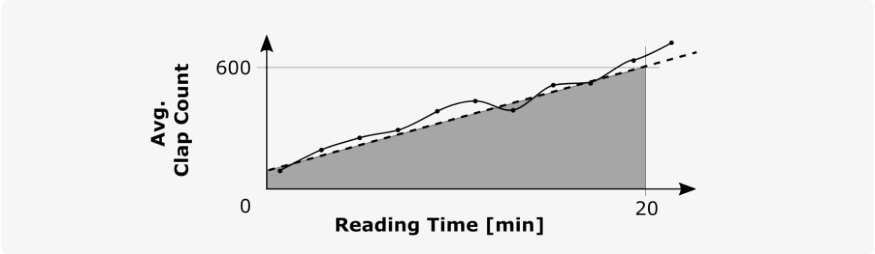


Image by the author

- Articles with a small number of claps usually have a higher claps/voters ratio (nearly 6 claps per voter) than articles with multiple voters (around 4 claps per voter).

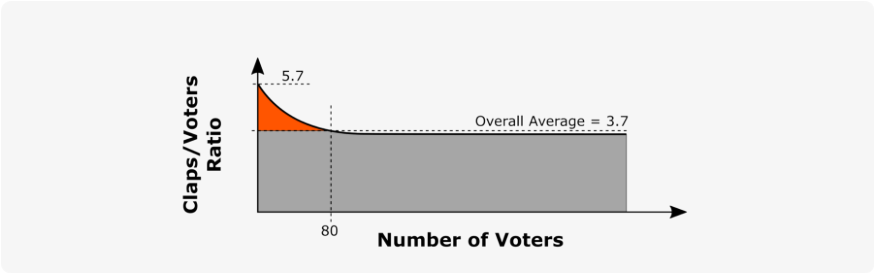


Image by the author

- Articles that can be assigned to the “Why”, “How” or “Guide” type score above average. They usually refer to a specific problem and describe how this can be solved/implemented.

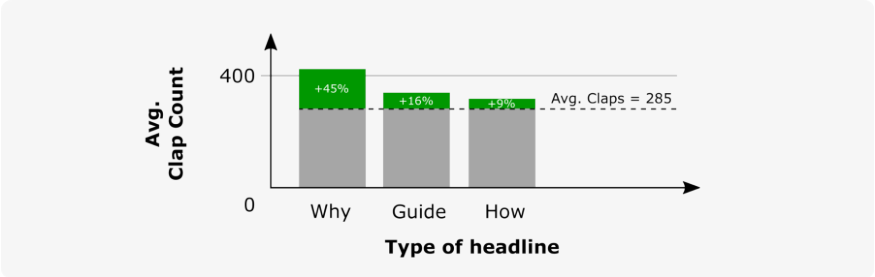


Image by the author

- A good 25% of all articles use numbers in their headline, and thus actually perform about 16% (13%) better than articles without numbers in the headline.

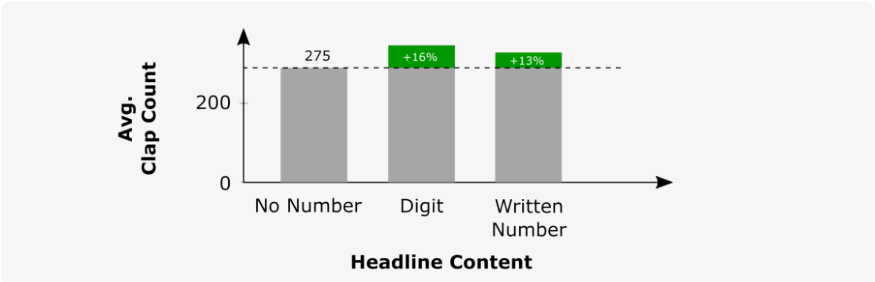


Image by the author

In fact, the results have helped me a lot to assess my own work more realistically. I hope the analysis makes it a little more transparent for you too.

Thanks for reading! If you liked the content, feel free to leave a clap or two, or follow me on Medium.

. . .

References

[Sem21] **The Anatomy of Top Performing Articles: Successful vs. Invisible Content—Semrush Study,** https://www.semrush.com/blog/anatomy-of-top-performing-articles/#header5