CITADEL | CITADEL | Securities

TheDataOpen

EUROPE REGIONAL FALL 2021 DATA OPEN

TEAM 2

Headline sentiment and topic effect on online user engagement

Authors:
Ludwig Jonsson
Rohit Nag
Manith Adikari
Maria Julia Maristany

Date: November 7, 2021

Executive Report

Project Description

In the last two decades, with the advent of social networks, most firms have grown an increasingly strong presence online. Whether we are talking about technology, marketing, media, retail, politics, health or activism, **engagement** has become the standard metric to measure impact, and therefore, success, of products, campaigns or media content.

Elucidating the causal relations that drive engagement is critical to furthering our understanding of the dynamics and social impact of online content. For example, it can help us understand better parasocial relationships, the spread of ”fake news” on social media or the effectiveness of a political campaign. It can also help us to build predictive opinion models, to improve the effectiveness of marketing campaigns, or to understand better the effects of misinformation in the general population.

In particular, there is a growing body of research focused on the effect of the title, or headline of online content in engagement [1, 2, 3]. A limiting factor on this research has been access to enough data that can help draw meaningful conclusions. The dataset from the Upworthy Research Archive [4] provides an open access dataset that can be potentially exploited to answer scientific and statistical questions, and to further explore the effects of headlines and images in online engagement.

In the next sections we describe our goals in more detail, and the results that machine learning techniques and statistical analysis yielded.

Basic Data

The dataset used for our work, compiled by Matias, Munger et al [4], is a collection of 32,488 A/B tests conducted by Upworthy from January 24, 2013 through April 14, 2015. Each A/B test consists on variations of headlines, images and/or descriptive tests for articles accessible through the Upworthy website.

For each test, the dataset includes viewer responses to over 150,000 different packages in an experiment. There is a median of 4 packages per test. It is, however, important to note that only a small subset of tests explored all possible variations.

Results

Our aim was to identify trends in user-behaviours, namely click rates. This would enable us not only to gain insight into which type of A/B test is more effective to increase engagement, but also a first look into what the collection of available data can tell us about the interaction dynamics between the population analyzed with online content.

The data consists of a collection of packages, corresponding to a particular test, where one or multiple parameters were varied. For example, a single test could consist of four packages: each with a potentially different headline, image, excerpt, etc.

For every specific package, a few parameters showing a measure of effectiveness was also present. These included the number of user impressions, that is, views that the article received, and the number of viewers who clicked on that particular package.

We set out to identify the parameters which specifically affected clicks. In our case, we focused on headlines and images, as these were the main pieces of information shown to the user before reading the article.

A/B Testing

We explored the effect of a change or absence of images on user click rate. Intuitively, one might expect the comparison between including an image to not including one to heavily favour the former. However, this was not a valid comparison for the collection of available data as there was not enough data of packages with absent images to perform a robust statistical analysis. Instead, we explored the effect of a change in image content by comparing performance of different images through the means of A/B testing.

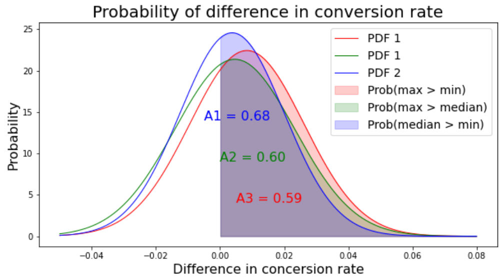


Figure 1: Probability of differences in conversion rates when comparing images.

Given that the actual, visual representation image was not part of the data set, we compared the top, average and worst performers based on click rates for the same image ID. The probability distribution of the difference in performance, as shown in figure 1, showed that the choice of image affected user click rates by approximately 65%. In other words, the top image in each group presented itself to be better than the worst one by 65%. As the image itself was not defined however, this result acts as a mere conclusion and a dead-end for any further statistical analysis.

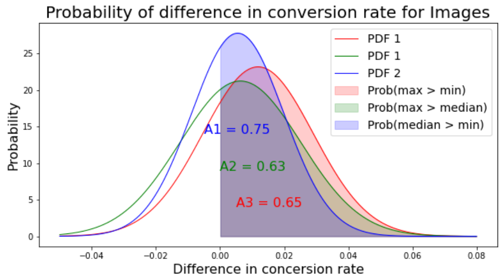


Figure 2: Probability of differences in conversion rates when comparing headlines.

On the other hand, a similar analysis on headlines (figure 2) not only showed a similar trend but more importantly paved the way for further exploration on identifying the particular aspect of the headline text which led to this user behaviour. This prompted sentimental analysis on the set of headline texts.

What is the relationship between sentiment and click rate?

To further explore trends between headlines and user-behaviour, we analyzed the **sentiment** of the headlines in the available data. This allowed us to discern how the variations in the tone, subjectivity headline phrasing could affect user behaviour.

To do this, we computed the *polarity* and *subjectivity* of the headlines, using the TextBlob NLP Python library. Polarity, that is, whether a headline reads as positive or negative, is defined as a value between -1 and 1, where -1 indicates an extremely negative viewpoint and 1 is a positive viewpoint. Subjectivity, on the other hand, is given as a value between 0 and 1, 0 being highly objective and 1 highly subjective.

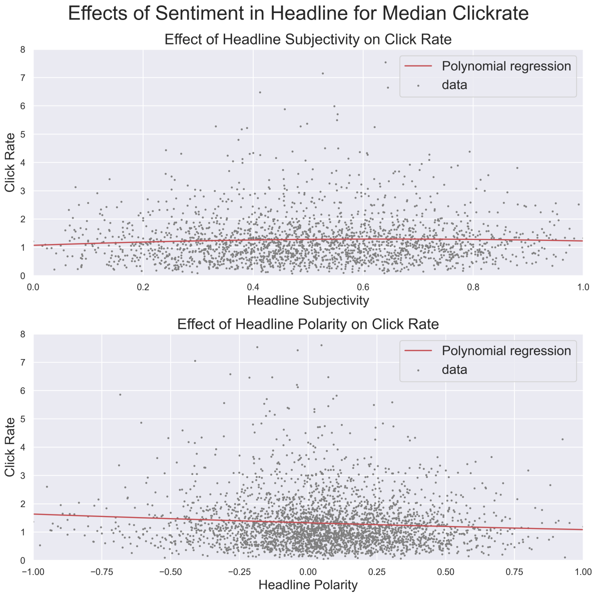


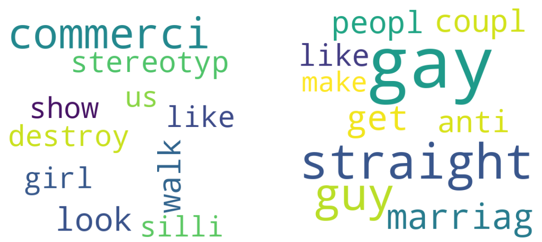
Figure 3: Graphs highlighting relationship between headline sentiment and click rate

As shown in figure 3, it is evident that most of the headlines have subjectivity and polarities clustered around 0.5 and 0.0 respectively, meaning that they are either neutral or slightly positive. The polynomial regression lines show no significant effect on click rate caused by polarity or subjectivity. However, it is relevant to highlight a weak negative trend between polarity and click rate, showing that negatively polarized headlines resulted in higher click rates.

Overall, the data suggests that the sentiment of the headline alone has minor to no effect on online engagement, represented as user click rate. To continue our analysis, we then decided to investigate the effects of the headline topic as an indicator of article content on user engagement.

Which topics lead to a higher click rate?

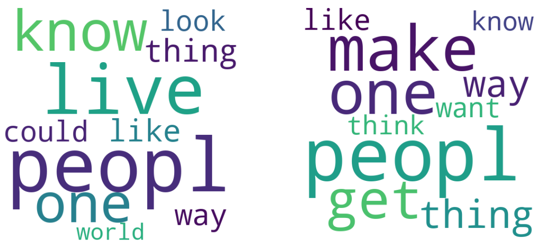
To categorize the headlines of the articles into topics, a machine learning algorithm called GSDMM was chosen. After categorizing each headline into 1 out of 50 topics, we chose the top 10 occurring words for the top 2 topics by median click rate and got the results in figure 4. Note that because of how we process that words, each word is showed in its 'stemmed' version. e.g. people → people.



(a) Topic 44 (b) Topic 41

Figure 4: Top ten words occurring in Topic 44 and Topic 41. Larger words corresponds to a higher frequency of that word in the topic.

In comparison, the top 2 least performing topics, in terms of click rate, contained the following words presented in figure 5.



(a) Topic 14 (b) Topic 11

Figure 5: Top ten words occurring in Topic 14 and Topic 11. Larger words corresponds to a higher frequency of that word in the topic.

Topic	Median click rate	Compared to all
Topic 41	0.01731	+39.8%
Topic 44	0.01743	+41%
Topic 14	0.00767	-56.6%
Topic 11	0.00841	-49.2%
All topics	0.01333	0%

Table 1: Median click rate for different Topics as well as the comparison to the median click rate for all topics. All decimal values were rounded to nearest 5th decimal.

As can be seen from figure 4, especially from Topic 41 (b), there is a clear trend of different key words that in-turn lead to a click rate about 40% higher than the median for all topics (see Table 2).

It should be noted that 64% of the traffic came from USA, and during this time period there was a lot of public debate regarding gay marriage, which in mid 2015 culminated in the country-wide legalization of same-sex marriage. For reference, the next highest traffic, at 8%, came from Canada. This means that Topic 41, shown in Figure 4 (b), would probably be popular at the time.

As a general comparison between Topic 44 and Topics 11 & 14, we can see that Topic 44 contain 'stronger' words than Topics 11 and 14. Such as: Destroy, Stereotype, Silly, and Commercial. Whilst Topics 11 and 14 contain quite vague and general words.

Summary and Perspectives

Overall, the data available suggests that A/B tests are slightly effective on increasing user engagement. However, due to the high p-value obtained (3̃0%) while performing the analysing—p-value being, in this case, a measure on the certainty on which we can distinguish between high performing and low performing tests—out hypothesis needs further testing, possibly with a larger, and more systematic, data set.

Further analysis of the article headlines suggests that the sentiment of a headline has little to no effect on driving user engagement, while the headline topic is much more effective, with changes of topic reflecting on a much higher variation on user click rate as a representative of online engagement.

Perspectives for this work, together with other possible avenues of analysis, can be found in B.

Technical Report

Methods

Throughout our analysis, we used many different statistical and computational techniques. This included an initial data familiarisation using A/B testing. After confirming our initial hypothesis on the effect of headline content on click rates, we used TextBlob to analyse the sentimental value of every headline in an attempt to identify the aspect responsible for causing the increase. Finally, we explored different avenues of topic modelling, to be able to determine the topic of a particular headline. For this we used machine learning techniques, finally settling on Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture Model.

A/B Testing Significance Analysis

The dataset contains quantitative data for viewer responses to A/B tests deployed on articles found in the Insight website. As a first step, to preliminary analyze the data, we analyzed the effectiveness of these tests to impact online engagement.
On an A/B test, test subjects are shown different options for a similar thing, and their choice gets recorded. In the tests we have available on the dataset, the parameters that changed between test subjects changed were mainly headlines, images or both. A first insight into the binomial distributions can be seen in figure , while an analysis on the distribution of the average click rate, which is well approximated by a normal distribution, in figure . In these, we do not distinguish between single or simultaneous changes on headlines/images.

Since an A/B test is not necessarily between just two options (there are occurrences of sometimes more than 6 different options for the same article), we needed a way to scale it down so we would have 3 distributions to compare. We did this by taking the min, median, and max in terms of click-rate, clicks, and impressions for each article.

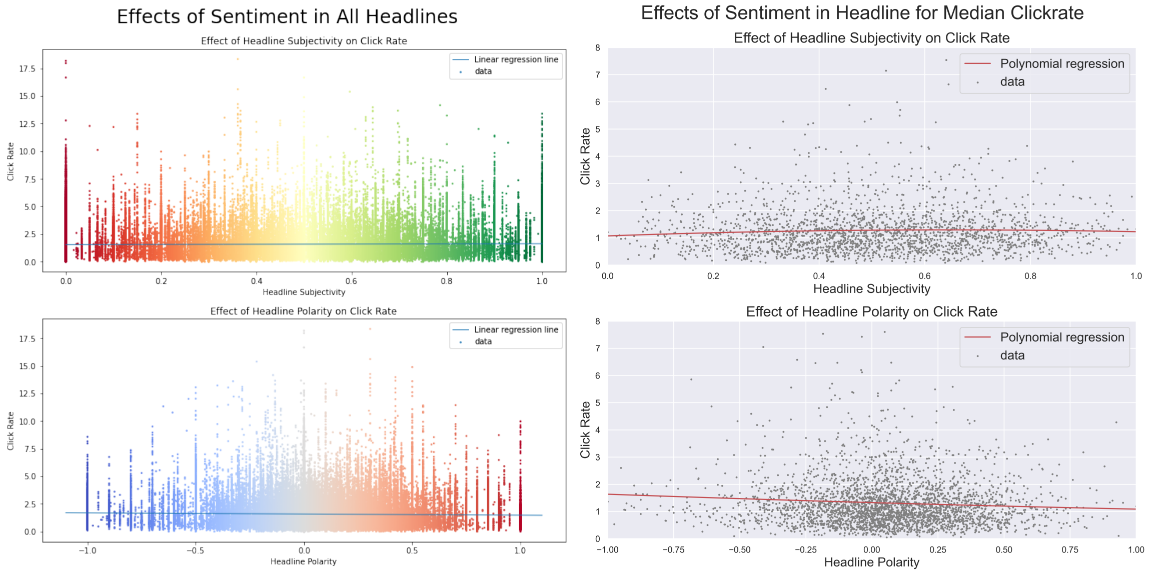
Using these values, we created the four plots shown in figure 7, as well as Figure 2 and 1. The Normal distributions were made using the mean and std of each gathered min, max, and median data. And the Binomial distribution was created using the pmf of the mean of clicks.

TextBlob: Simplified Text Processing and Sentiment Analysis

The data was imported into a Python IDE (Integrated development environment) and manipulated using pandas, a Python software library for data analysis. This allowed the data to be held within tabular data structures for pre-processing and further analysis. Additional libraries were utilised to execute specific functions, such as NLP and data visualisation.
TextBlob is a Python library that was developed with the primary aim of natural language processing (NLP). It has many features and potential applications:the main feature we adopted as a tool of analysis for our data was that of **sentiment**. This query return a tuple of the form (polarity, subjectivity). In this context, polarity is a number ranging from -1 to 1, which indicates the negativity/positivity of the sentence analyzed. Similarly, subjectivity is a number that ranges from 0 to 1, where 0.0 indicates that the sentence is very objective, and 1 very subjective.
The data seemed to show some form of pattern at first sight. However, upon closer analysis it was evident that the pyramid shape was formed from a higher concentration of data being included in the dataset for middle values (i.e for subjectivity = 0.5 and polarity = 0). As expected, the data contained a lot of overlap. For each value of polarity or subjectivity, there were multiple points. To get a unique data point for every value, the mean of each value was taken and plotted. A polynomial regression line fit on top of this data showed a slight trend but not conclusive.

Latent Dirichlet Allocation

Latent Dirichlet Allocation [5] (LDA) is a probabilistic model for collection generation, used in machine learning for text classification, or more generally, topic discovering. It is an unsupervised algorithm, consisting of a three-level hierarchical Bayesian model, and focused on finding statistical co-occurrence patterns in a set of training documents to elucidate their semantic structure. Once this structure is found, a new document can be ex-



(a) Scatter plot showing headline sentiment (subjectivity in the top image, polarity at the bottom) against user click rate. The analysis shows little to no effect on sentiment against click rate. In this preliminary analysis, the data is skewed since we considered all packages, which meant that we had duplicates in headlines. See figure 6(b) for the effect on the median of the user statistics corresponding to each unique headline.

(b) Graphs highlighting relationship between headline sentiment and click rate. To account for the fact that there were repeated headlines with the same test ID, for each unique headline we considered the median click rate. The analysis suggests very slight trends favoring negative polarity, and slight headline subjectivity as drivers of user engagement.

Figure 6: Plots showing the effects on user engagement of sentiment in the article headline.

pressed in this representation, and queried for topical similarity against other documents. In this context, a topic is a collection of terms (in our case, words), and a document, a headline, lede or excerpt.

Using LDA for our data to find topics for the different headlines was tried and tested. However, because our Headlines consist of less than 50 words, they can be classified as short-texts and thus LDA will not be able to find definate topics for each headline. Short-Texts are infamously one of LDA's weaknesses. See this article from TowardsDataScience [6].

In this report, the LDA algorithm was implemented in the open-source Python library Gensim [7]. See A for more details.

GSDMM

The Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture Model, or GSADMM, is another Bayesian clustering model [8]. To understand how this algorithm works, the analogy given by Wang et al in their work proves useful:

Let us imagine that we are in a film class, where the students have to arrange themselves into groups according to their movie tastes. To simplify things, the professor asks them to quickly write down a couple of the movies they have recently watched. Now each student is effectively labeled with a preliminary, albeit imperfect, list of movies that represent their taste. The clustering algorithm then works as follows: the professor will randomly assign the students to K different tables (categories). In the next step, the students will move, with certain probability, to a different table. The probability of moving will depend on two things: the size of the table, and the movie interests of the student in such table. Essentially, the bigger the table, and the more similar the taste, the more likely for a student to make the move.

One important difference which makes GSDMM more favorable towards short text topic modelling is that it assumes each text belongs to only 1 topic.

Compare this to LDA which assumes that a text can belong to many topics.

We used a codebase created by GitHub user rwalk[9] to run the GSDMM Topic Modelling. We trained a number of models with different hyperparameters and then chose one which seemed to produce more obvious topic clusters.

Text Pre-Processing

For both LDA and GSDMM we applied some simple language processing in order to turn the strings of text input to our models into something our models could better work with. This was done using the Python libraries nltk, contractions, and gensim. Steps would include:

- Make all words lowercase and expand any contracted words. e.g. They're → they are.
- Remove any words which are numerical or English 'stop' words, such as '1' or 'and'.
- Remove any words which are not greater than 1 character in length.
- Lastly each word is either Lemmatized using WordNetLemmatizer, or Stemmed using SnowballStemmer. Depending on the model targetted.
- Optionally, if using LDA we also need to turn each string of words into a vector format. This was done using the 'Bag-of-Words' or BOW method using Gensim's doc2bow() method.

It was chosen to use Lemmatization for the LDA and Stemming for GSDMM. This was based on the fact that many guides found online chose to use Lemmatization for LDA and Stemming for GSDMM.

Visualisation

In order to visualise data, we used the Python libraries Seaborn and Matplotlib for all plots and graphs. For the word clouds shown in Figure 4 and 5, we used the Python library wordcloud.

Results Discussion

GSDMM for Topic Modelling

In figure 8 we can see how each topic fares in terms of its median click-rate for articles there. In order to test for the Topic Coherence of our trained model, we used the CoherenceModel which comes shipped with the Python library Gensim. Since our model is not completely compatible with it, we had to pick which words to use. This ended up being the Top-N words, ranked by frequency in each topic/cluster. The results are presented in Table 2.

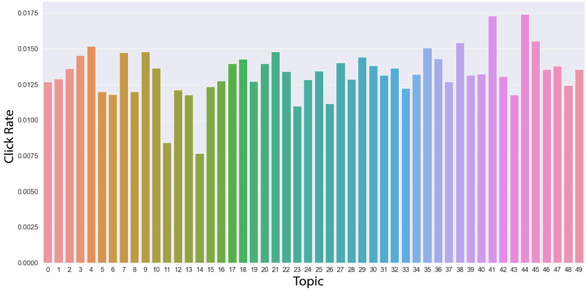


Figure 8: Median click rate for each topic found using GSDMM. It used Hyperparameters $\alpha = 0.1, \beta = 0.1$ and was run with 30 iterations.

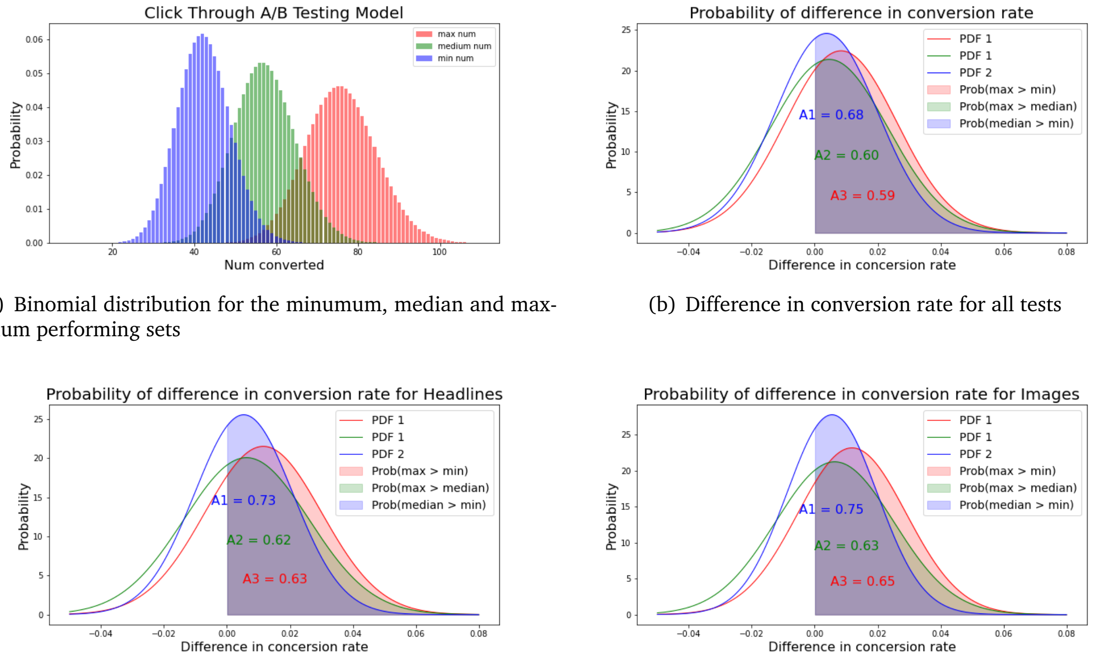
Number of words	Topic Coherence
Top 10 Words	0.31205
Top 8 Words	0.35779
Top 5 Words	0.46183

Table 2: Topic Mutual Coeherence tested on the Top-N words (by frequency) for our GSDMM model. All decimal values were rounded to nearest 5th decimal.

As can be seen from the data in Table 2, across all Topics/Clusters we find quite bad Mutual Coherence between topics when using 8 or more words. However, when we limit the top words to be 5, we reach a more acceptable score. This can then referenced to how the word clouds shown in figure 4 and 5 look. I.e. the larger words show much more significance than the smaller ones.

A/B testing P Value

The minimum, median and maximum values were assumed to be three discrete options for the user to



(a) Binomial distribution for the minumum, median and maximum performing sets

(b) Difference in conversion rate for all tests

(c) Difference in conversion rate for tests with varying headlines only

(d) Difference in conversion rate for tests with varying images only

Figure 7: A/B Testing significance analysis graphs

click on. With this assumption, the hypothetical binomial plots could be plotted using their respective click rates. To validate the results from A/B tests, both the z-value and P-value was calculated. The P-value is used to confirm the significance of the results when compared to the null hypothesis where there is no relationship between the categories being tested. Essentially, it gives a measure of the probability that the observed results were caused by unlikely observations.

A known mathematical expression shown in equation 1 shows that the difference of random numbers, normally distribution is also a normal distribution. Using this, the the difference in conversion rate normal distribution could be plotted as shown in figure and .

$$P(b-a) = \mathcal{N}(\mu_B - \mu_A, \sqrt{\sigma_A^2 + \sigma_B^2}) \qquad (1)$$

For the probability distributions shown in figure , the, the p value was found to be 0.27, 0.38 and 0.37 for the three cases respectively. The A/B test results were therefore inclusive with the available data.

References

[1] Linda Lai and Audun Farbrot. What makes you click? the effect of question headlines on readership in computer-mediated communication. *Social Influence*, 9(4):289–299, 2014. pages 2

[2] Lewandowsky S. Chang E. P. Pillai R. Ecker, U. K. H. The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology*, 20:323–335, 2014. pages 2

[3] Vaibhav Kumar, Mrinal Dhar, Dhruv Khattar, Yash Kumar Lal, Abhimanshu Mishra, Manish Shrivastava, and Vasudeva Varma. Swde : A sub-word and document embedding based engine for clickbait detection, 2018. pages 2

[4] Munger K. Le Quere M.A. et al. Matias, J.N. The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. *Nature Sci Data*, 8:195, 2021. pages 2

[5] David M Blei, Andrew Y Ng, and Jordan Edu. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. pages 5

[6] Richard Pelgrim. Short-text topic modelling: Lda vs gsdmm. `https://towardsdatascience.com`. pages 6

[7] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`. pages 6

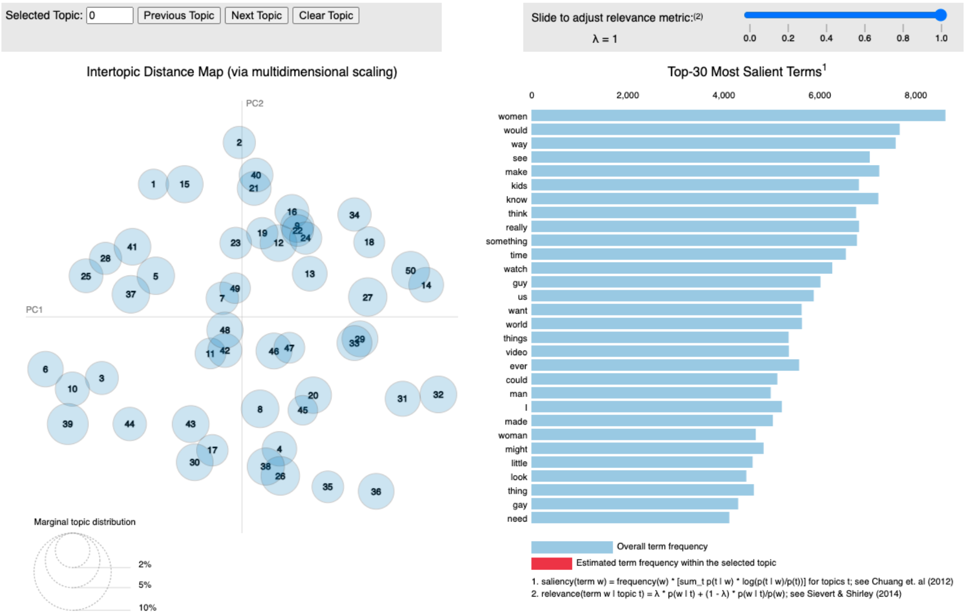
[8] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 233–242, New York, NY, USA, 2014. Association for Computing Machinery. pages 6

[9] Ryan Walker. Gsdmm: Short text clustering. `https://github.com/rwalk/gsdmm`. pages 7

Appendices

A Topic Modelling with LDA: Challenges and Discussion

While LDA is one of the most widely used, optimized, and effective tools for topic modelling that is available nowadays, it has severe shortcomings. Convergence for large data sets such as ours fails when the average length of the textual data is short. In essence, LDA does not work very well with large sparse data. LDA assumes that every string of text or document belongs a bit to all clusters (with a probability of belonging to each), which e.g. proves troublesome when you have e.g. 8 words belonging to 50 topics. In Figures 9 and 10 we visualise the clusters that LDA, using first 50 and later 10 clusters, gave us using the Python library pyLDAvis. In Figure 11 we see how LDA fared when trying to map topics to click-rate. Similar to how we did in GSDMM in Figure 8.

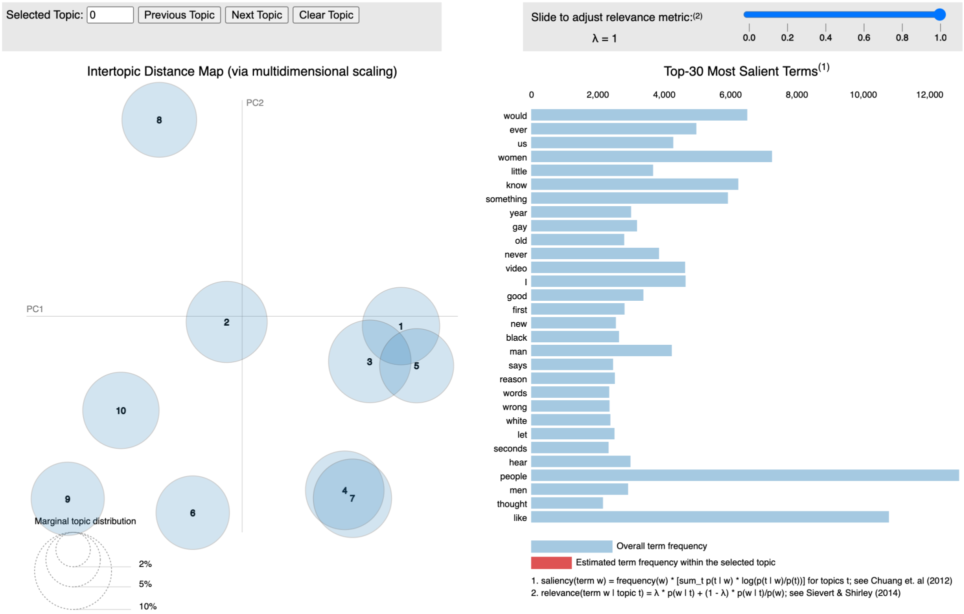

Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t l w) * log(p(t l w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w l topic t) = λ * p(w l t) + (1 - λ) * p(w l t)/p(w); see Sievert & Shirley (2014)

Figure 9: Visualisation of Topics/Clusters found by using LDA with 50 clusters specified.



Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10: Visualisation of Topics/Clusters found by using LDA with 10 clusters specified.

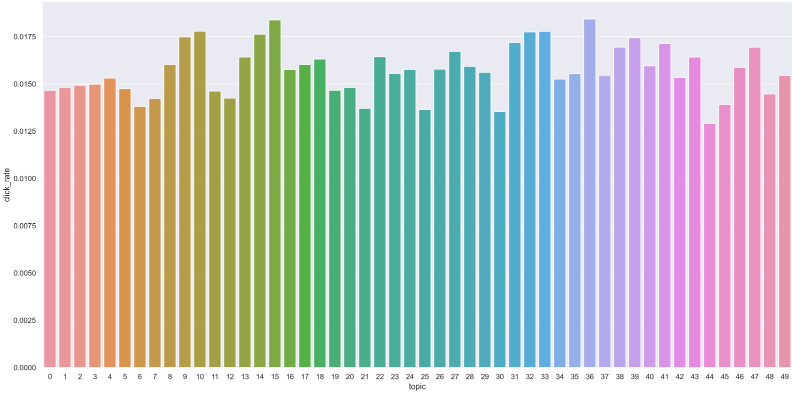


Figure 11: Median click rate for each topic found using LDA.

B Future research areas

Another possible way of categorizing the data includes using semi-unsupervised learning to distinguish between articles true to their headline against those that would qualify as to clickbait. These approach would involve to manually label a small set of articles, distinguishing them between real and clickbait, to then use an unsupervised technique to label the rest of the data set. Then, a supervised model can be run to try and predict if the article is likely to be clickbait.