

**ISSS609 Text Analytics and Applications Project Report**

***Group 8***

***Jiang, Nan***

***Tey, Wei Jie***

***Wang, Yechun***

***Wang, Yuxin***

***Yao, Yu***

***Zhu, Chenzi***

***21 April 2019***

**Word Count: 5049**

Paper 001-2018/2019T1-G2.8

Convenient and Concise Digitalisation of News Articles

Jiang, Nan; Tey, Wei Jie; Wang, Yechun; Wang, Yuxin; Yao, Yu; Zhu, Chenzi

*Singapore Management University*

# Abstract

In an age of information overload, there is an explosion of online news articles from various publication centres, causing dilemma for readers to choose from. The fact that none of the publications stand out for readers signify that there are some issues with the current digitalization of news articles. Although readers prefer text news, a convenient platform with ease of access and concise articles, we often see a cluttered news website with detailed and long articles. Moreover, more publications seem to be adopting video news as they frequently include video news in the articles. These lead to a gap in readers’ wants and current solutions. In this research paper, the team seeks to explore the use of text analytics tools such as topic modelling and text summarization techniques to resolve the gaps. The results were pleasing as we were able to obtain many distinct topics and hence reduce the cluttering due to news headline display as well as to generate summary for news articles.

# Introduction

## Motivation

In today’s information age, we often face information overload in our daily lives. Similarly, there is an explosion of news articles from various publication centres, resulting in multiple selection dilemma by the users. This is especially so given the advancement in social media and technology and hence more news publication centres are adopting online or app version to release the latest news as shown in Figure 1 where the volume of online news consumption has rapidly increased (45% of US adults in 2017)[[1]](#footnote-1)..

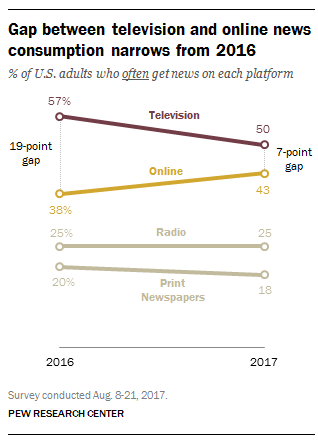


Figure 1: The gap between television and online news consumption is narrowing

This results in users having multiples choices to select from when they want to obtain the latest news.

While it may seem beneficial for the users initially, the information explosion results in repetitive and numerous articles written across all publications. As a result, users may often waste precious time reading similar articles. Moreover, the informative articles are usually long and detailed. These long form articles often turn frequent readers off as they seek a quick update of the daily news. This is not surprisingly that a large proportion (roughly 40%) of readers tend to skim through such articles instead of reading them thoroughly[[2]](#footnote-2).

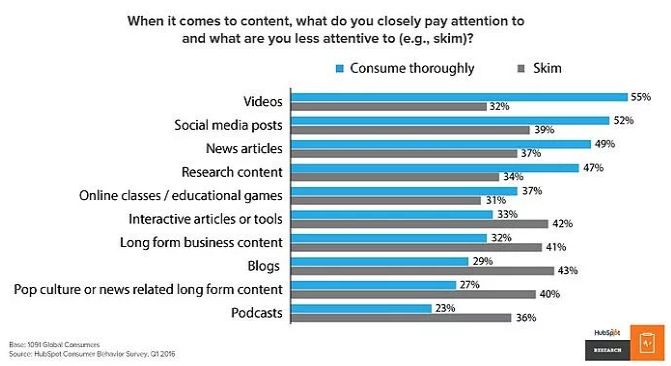


Figure 2: Percentage of readers who skimmed through content

All these highlight the problem of having varied options in the news publications area which require an easier access for users to obtain exactly what interest them as well as to ensure that the articles are not too long so that readers are not put off by it. While there are other ways of engaging users such as video given the popularity of YouTube, Instagram, Facebook and Twitter which frequently displays news videos, users still prefer text news across the globe as many find reading quicker and hence less time consuming[[3]](#footnote-3).

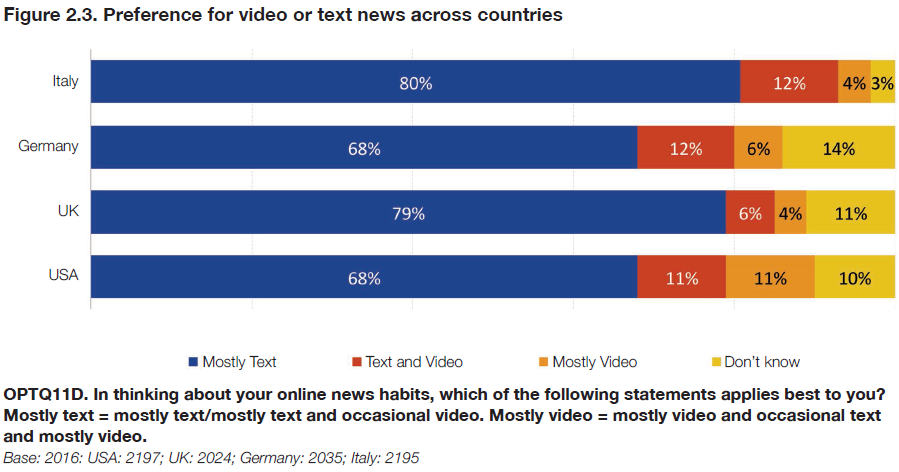


Figure 3: Preference for text over video news

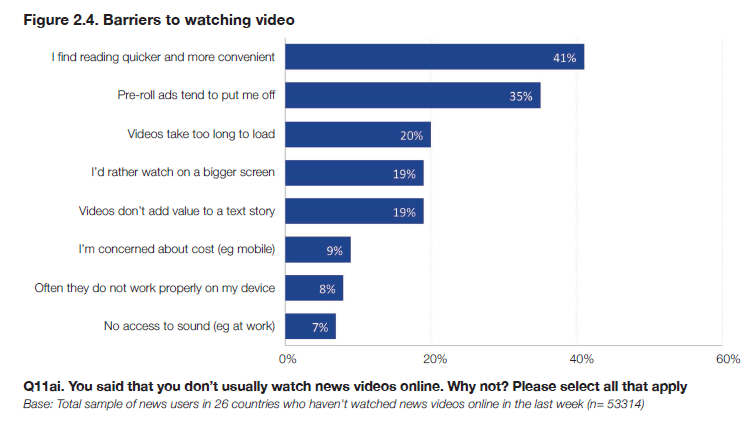


Figure 4: Reasons against Video News

Furthermore, readers are particular with the ease of navigation which indicates that the platform needs to be convenient. This is highlighted in the readers’ feedback on New York Times Website revamp where many readers hit back at the publication for complicating the website.[[4]](#footnote-4)

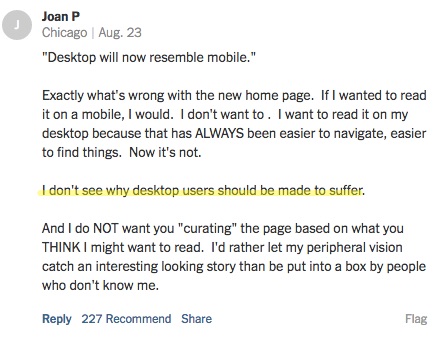


Figure 5: Feedback from one of the New York Times readers

In addition, current existing publications are not providing topical keywords despite displaying topics. The topics displayed are also generic and not comprehensive enough. Lastly, there is currently no function available in any publications that can summarize the articles.



Figure 6: Front Page Display of Publications (NYT & Guardian)

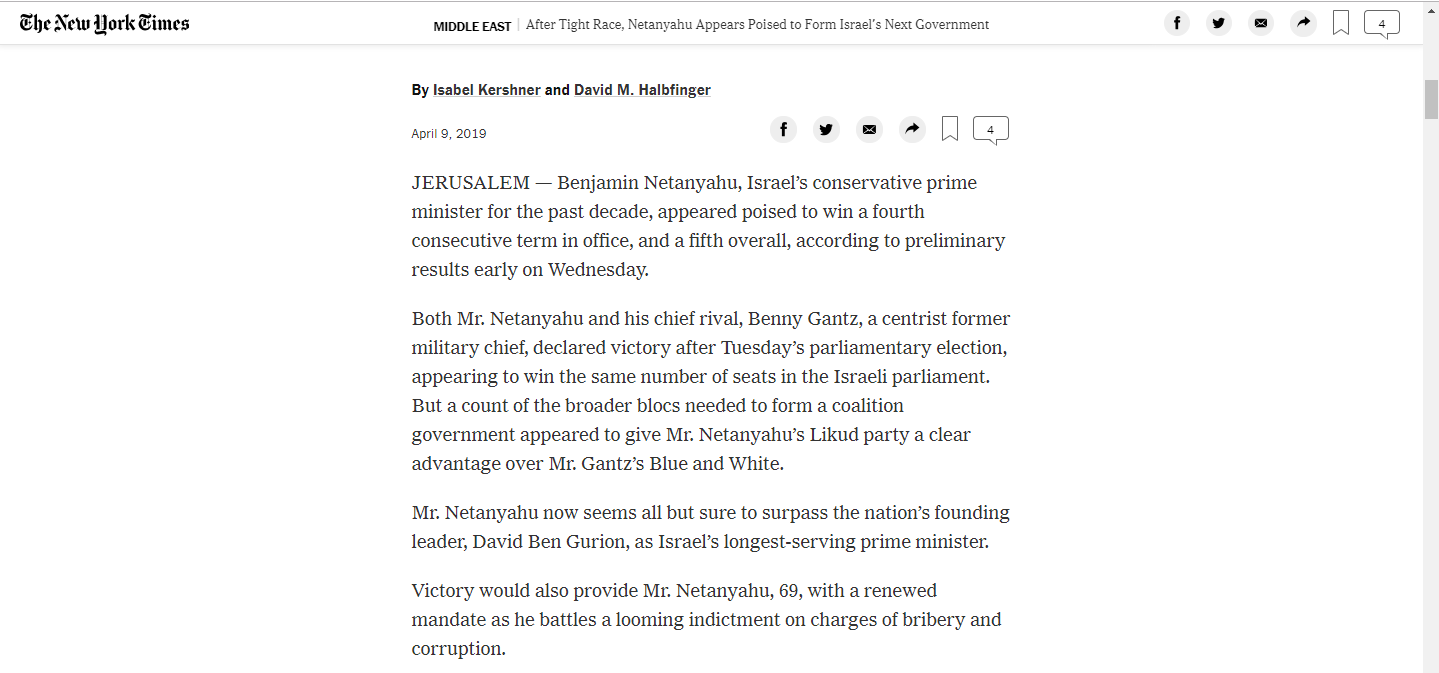


Figure 7: Example of News Articles

All these signify a rather interesting issue where content publications are not satisfying their audience preference when it comes to news reading. Thus, with a need to provide an easier access to relevant content and a shorter text form articles, the team aims to explore relevant text mining techniques in solving this issue.

## Research Objectives

As we aim to target users who frequently read news digitally and find it inconvenient to look through multiple apps, we seek to enhance convenience and ease of finding relevant summarized articles with a click of a button on a word cloud displaying various topics.

Thus, we plan to deploy mainly two ways in our project: a word cloud that display the current topics and articles summarization.

1. Topical Word Cloud

The overload of latest news articles from various publications leads to users being unable to quickly identify the topics that interest them. As our primary objective to provide convenience to news readers, we decide to create a word cloud where we will display all the current topics across the selected news publications, allowing readers to have access to the latest topics in a consolidated place.

In order to achieve this outcome, we will utilize the following text analytics tasks:

* Topic Analysis: As our objective is to display the current topics across all publications, we will need to group the news articles across all the publications and identify the major topics that occur across those articles. We plan to explore LDA model to cluster those documents into various topics.

1. Articles Summarization

As it is time consuming to read various long news articles, the team aims to provide users the option to read a summary of the articles. This will provide convenience to them as many readers tend to skim through long articles and hence may miss out on important points which may be crucial to understand the main gist of the articles.

In order to achieve this outcome, we will utilize the following text analytics tasks:

* Text Summarization: As our objective is to ensure a shorter article, we will summarize the existing articles so that users can quickly glance through and understand what the articles are about. We will explore both Abstractive and Extractive Summarization methods and evaluate them to select the best model for our eventual model input into the User Interface (UI).

## Paper outline

The rest of the papers are organized as follows. Section 2 outlines some of the past research done in this field while Section 3 will cover the data preparation and selection for our research purpose. Section 4 and 5 will cover this research paper’s methodology in detail, highlighting the use case of the respective text analytics task in this field. Moreover, the analysis of the results will be illustrated in Section 6 and 7 where the team will describe the things that went well and those that did not. We will end off with future works and conclusion in Section 8 and 9.

# Literature Review

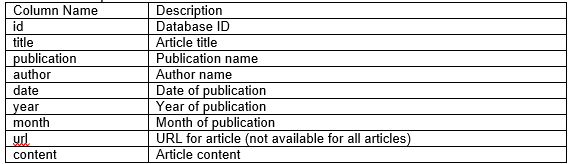
The use of text analytics tools in news publications is not a new dawn with many researchers deploying various methods to make news reading easier. The most popular research area has been the use of text mining in newspaper archives to provide convenient access to historical sources of data.[[5]](#footnote-5) This is essential as many historians and publishers tend to refer to historical newspapers during their works.[[6]](#footnote-6) With more readers going online, this proves the motivation to utilize text mining tools to simplify the news articles as well as to store them for future use.

Silipo (2019) explains that topic modelling is typically used in splitting articles into topics and that text summarization can be used wisely to extract the most meaningful words from text document to represent it. This forms the methodology of this research as the team draws ideas from this article to perform the two text analytics tasks mentioned above. While Silipo (2019) recommends the use of Latent Dirichlet Allocation (LDA) as a way for text summarization, we explore other forms of summarization in our research as we were encouraged by the fact that the types of text summarization used in the field vary from thematic features such as term frequencies to syntactic analysis approach.[[7]](#footnote-7)

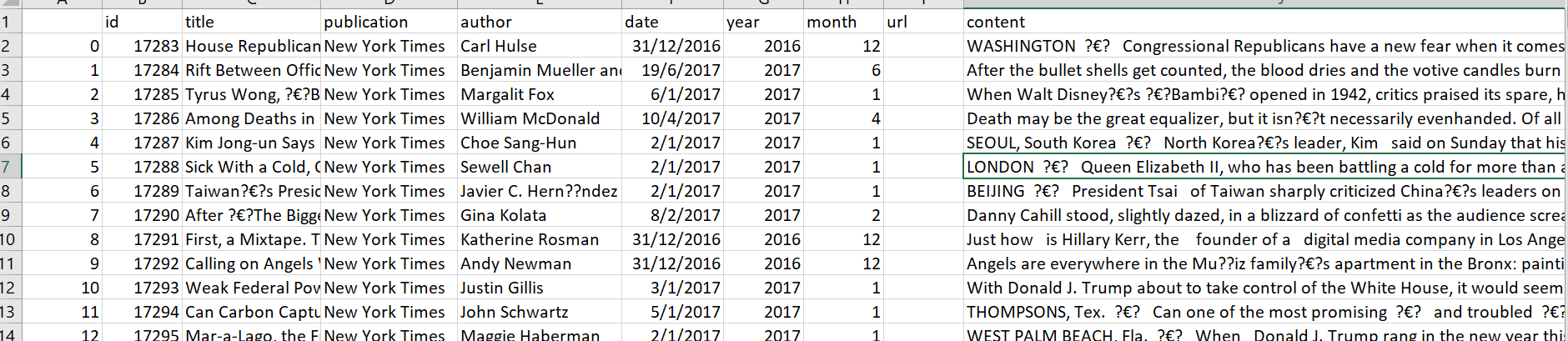
# Data Preparation

The dataset is sourced from Kaggle.[[8]](#footnote-8) There are more than 150,000 news from 15 US mainstream publications, such as New York Times, CNN, Business Insider, and National Review.[[9]](#footnote-9) The news in the dataset primarily falls between 2016 and July 2017.

The dataset field categories are as follows.



Below is a sample of the dataset:



## Data Challenges

Although we only require the article contents, the data set does pose some challenges for us.

1. There are more than 100 thousand articles. This will be computationally intensive when running through the text mining models. Due to time and machine specification constraints, we have taken a sample size of roughly 40 thousand articles.
2. Based on our business scenario usage, the ideal data will be real-time and fed into our model through an API. However, this will involve architecture design of backend systems. Hence, the team has taken this data set as a proxy of the global news articles to prove the concept and methodology at this stage.

## Data Selection

As mentioned above, we have extracted a sample from the data through selecting the most famous publications. Other ways could include selecting a shorter time period to observe. However, we believe that limiting to a few popular publications will have a less impactful outcome.

To do so, we look through various online sources to determine the top 4 most popular news publications. The reason behind choosing the most popular is to ensure that there are readerships for those publications, signifying an adequate level of reliability.[[10]](#footnote-10)

The final selected publications are Washington Post, Reuters, New York Times and Guardian.

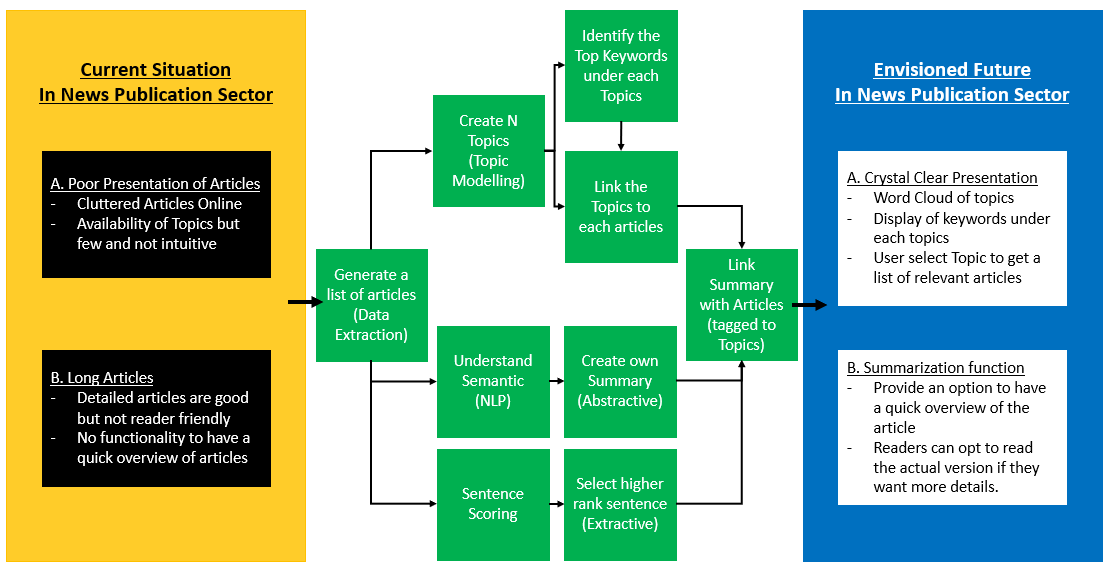
## Data Pre-Processing

As the tasks deal with article texts, we need to ensure the quality of the texts before inputting them into any models. Hence, we perform the following pre-processing techniques in general and use the relevant pre-processed files as the input for our respective models.

1. Tokenization: to split documents into tokens
2. Make all words into lower case: to ensure that identical words are not mistreated as different words due to them being in different case form.
3. Take only alphabets: to remove noise from potential non-alphabetic words that can distort the results.
4. Stopwords: to remove common words such as “the” that are considered noises as they do not provide useful information in identifying topics. These stopwords also tend to be among the top few most frequently used words and hence, we have eliminated them.

# Solution Overview

The current issue boils down to an inconvenient platform due to cluttering and long articles. Hence, we aim to address these two issues through topic modelling and text summarization techniques. Below is a depiction of our solution.



# Methodology

To provide a clearer topical display of current articles as well as summarization functionality to have concise articles, we have explored the use of Latent Dirichlet Allocation (LDA) algorithm to perform topic modelling as well as two forms of summarization techniques (Abstractive and Extractive). These text analytics tasks were performed in Jupyter Notebook through Python programming.

Jupyter Notebook is an open-source web software that allows users to create documents displaying codes, visualizations and text using Python[[11]](#footnote-11).

## Topic Modelling[[12]](#footnote-12)

Typical text clustering algorithms group documents into hard clusters which give one label to to one cluster. This means that an article can only be tagged to one label. However, in the case of news publications, we often observe that the news articles are comprised of various topics within one article. Thus, there is a need to use a soft clustering technique where one document can have multiple labels.

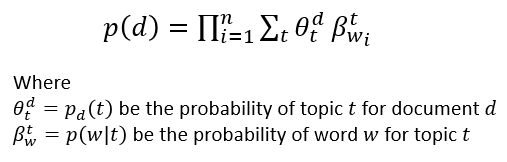
Therefore, the team has chosen the use of topic modelling as it identifies topics mentioned in the articles by understanding the co-occurrence of a set of words in the articles. Amidst the different topic modelling algorithm such as Probalisitc latent semantic analysis (PLSA), we have utilised LDA model in our topic modelling analysis.

LDA approach considers each documents as a distribution of topics and each topics as a distribution of words. It is an unsurpervised machine learning technique that finds the best distributions which maximise the probability of generating all the articles.



Figure 8: LDA Model

Generic Formula for LDA:



Parameters of LDA model include (alpha) and (eta). We did not adjust any of these parameters. Alpha is the document-topic density while eta is the topic-word density.

Through the Gensim package in Python, we construct the LDA model, considering the following key factors[[13]](#footnote-13) in obtaining a good segregation topic.

1. The quality of text processing.
2. The variety of topics the text talks about.
3. The choice of topic modeling algorithm.
4. The number of topics fed to the algorithm.
5. The algorithms tuning parameters.

In our methodology model run, we utilized the pre-processed data and turn them into vectors. This is done by first creating a dictionary representation of the articles and then create a bag of words representation of the articles. The final parameters are then fed into the LDA model.

As we need to input the number of topics (k) for LDA model, we have run a loop function to create multiple runs with varying k. To identify the optimal k, we will evaluate the outcome of each run with evaluation methods discussed in the later session.

Through our model run, we noticed that the stopwords function in NLTK was insufficient to remove all the noisy words and hence, we added a further list to the stopwords function and rerun in order to get a more representative topic segmentation.

|  |
| --- |
| Code for LDA model |
|  |

## Articles Summarisation

We utilize two main methods: Abstractive and Extractive Summarization. The average score method is one dimensional, looking at the raw frequency of words when weighing the importance while the TD-IDF method is bi-dimensional, considering both the rare and important words.

### 5.2.1 Abstractive Summarization

The underlying methodology is the seq2seq models[[14]](#footnote-14). This Natural Language Processing (NLP technique) utilizes TensorFlow for summarization of textual data. TensorFlow is an open source library created by Google for numerical computation and machine learning.[[15]](#footnote-15)

This method utilizes slightly different preprocessing techniques. A look up dictionary is created to map each unique word to a number, adding in 4 special tags. EOS indicates end of sentence. SOS symbolises start of sentence. PAD is the placeholder to fit the sentence into a standard batch length. UNK refers the word that is unkown in my dictionary.

|  |
| --- |
| Code for Lookup Dictionary |
|  |

In the embedding step, we use pre-trained embedding from TensorFlow hub, which is generated from a large collection of Wikipedia articles.

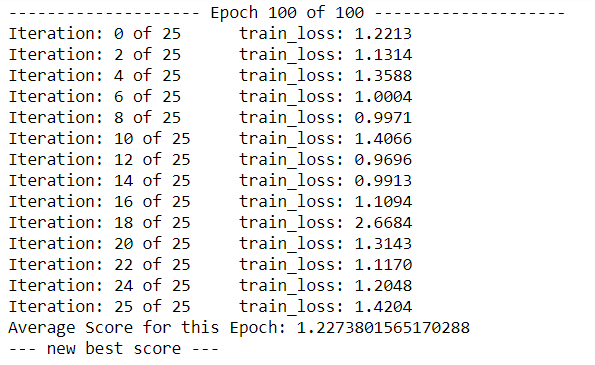
|  |
| --- |
| Code for Embeddings |
|  |

After these processing, each article has been seperated into words, and each summary has been added to unique tags. We have also succesfully tokenise the words into ids.

In our training model, the following parameters are defined.

|  |
| --- |
| Code for Abstractive Model |
|  |

In the training, we have run an epoch of 100 and has achieve a best score of 1.23 as shown below.



### 5.2.2 Extractive Summarization

Extractive method is by its name, which extracts important sentences from articles and formulate them into a summary through placing weights into each sentence. For this research, we explore different algorithms to define the weights of the sentences and further rank them based on importance and similarity among each other.

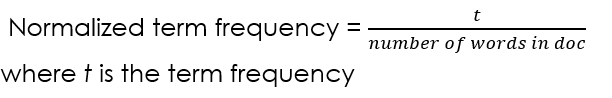
Method 1: Average score threshold

Based on the preprocessed data, we do the following processes to generate the summary.

First, we created the word frequency table to weigh the importance of each word except the stopwords.

|  |
| --- |
| Code for Word Frequency table |
| page1image859973040 |

Then we score the sentences by summing the frequency of every non-stop word in a sentence before dividing by total number of words in a sentence (Term Frequency). A potential issue with this is that long sentences will have an advantage over short sentences to get a high score. To solve this, we perform normalization, dividing every sentence score by the number of words in the sentence.



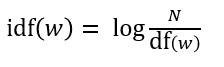
Equation 1: Normalization of Text

Next, we find the average score of the sentence value and set the threshold (1.3x) to select sentences in forming the summary before finally generating the summary.

|  |
| --- |
| Code for Average Score |
| page2image738101776 |
| Code for Summary Generation |
| page2image738102112 |

Method 2: TF-IDF

Similar to the average score method, we created the word frequency matrix for each sentence first before calculating the term frequency for each word in the sentence.



**Equation 2: IDF Equation**

The additional step in this methodology is to calculate the IDF value for each word in the text. It calculates word value based on the text, not on the sentence. Next, we calculated the TF-IDF value for each word by the following algorithm.

|  |
| --- |
| Code for TD-IDF |
| page4image738019232 |

Finally, we calculated the average score of the sentences values and set the threshold before selecting the qualified sentences to generate summary.

Comparison between Average Score and TF-IDF

The average score model and the TF-IDF model are similar in their implementation from weighing the words to selecting the sentence higher than the threshold to generate the summary. The main difference is the method used to weight the word. Moreover, TF-IDF model takes the uniqueness of word into consideration in the IDF.

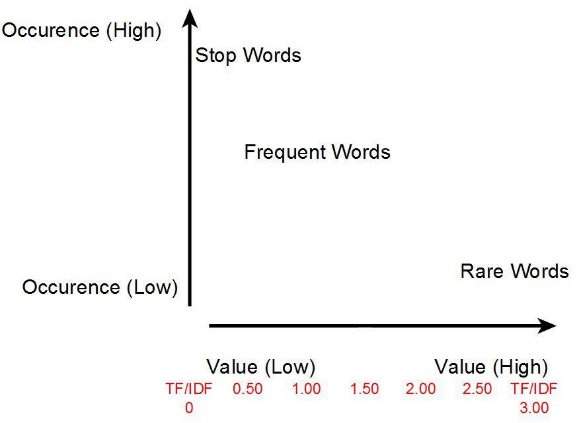


Figure 9: Frequency of Words in TF-IDF

# Results & Analysis

As the team has run through various models, we have analyzed the outputs as well as evaluated the models so as to determine their effectiveness and usefulness. This is essential as we need to identify the optimal model in each case (topic modelling and article summarization) for the implementation in the UI.

## Topic Modelling

Since the use of topic modelling is to obtain the topics mentioned in the articles, we will need to evaluate the topics in terms of clarity and independence. Clarity is defined as words in a topic should be understandable and related to each other. Independence refers to the degree of non overlapping fundamental keywords across topics, resulting in a distinct cut of topics. In this report, the team explores purely the use of traditional LDA model. Although LDA model is an unsupervised machine learning method, LDA model users will still have to indicate the number of topics. Hence, there is also a need to determine and evaluate the optimal number of topics.

### 6.1.1 Model Evaluation

In order to obtain the optimal number of topics as well as to ensure the topics are clear and independent, the team has to perform evaluation methods and compare the results. We looked through a few evaluation methods in our analysis, mainly perplexity, topic coherence and human evaluation through coherence rating.

Traditionally, perplexity is often used in evaluating topic models where a better model assigns high probability to words that actually occur through the measurement of log-likelihood. Hence, the lower the perplexity score, the better the model. However, this predictive likelihood and human judgement are often not correlated and hence it might be not helpful to use this method on a standalone basis when evaluating topic models[[16]](#footnote-16).

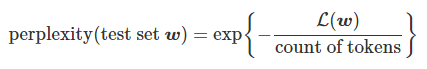


Figure 10: Perplexity Equation

On the other hand, human evaluation technique through coherence requires human participants to rate the coherence of the topic. However, this is time consuming and not an effective way to determine the optimal number of topics as there could be multiple iterations. Hence, this is only ideal when paired up with other methods to determine between two closely matched optimal number of topics models.

Due to the nature of the below mentioned methods, the team leans towards the use of topic coherence method in determining the optimal number of topics as well as its clarity and independence. Topic coherence is applied to the top N words from each topic and the pairwise word similarity score of the words in the topic is determined. This means that the higher the coherence rating, the more similar the words are in each topic which satisfy the definition of our clarity requirement. Thus, topic coherence measures the human interpretability of a topic model in essence[[17]](#footnote-17).

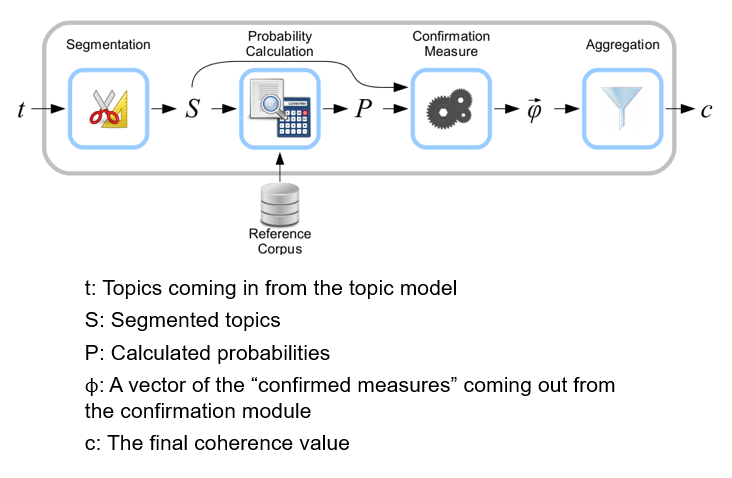


Figure 11: Coherence Methodology

### 6.1.2 Insights and Analysis

The team has performed the LDA model run twice (once with the NLTK original stopwords function and the other with added words into the stopwords function).

The optimal number of topics chosen is based on topic coherence score unless it is unclear. Then the team will use other evaluation in conjunction to make a final judgment.

For the first run with the original NLTK stopwords function, we were able to obtain an optimal number of topics as 20 with a coherence score of 0.49.

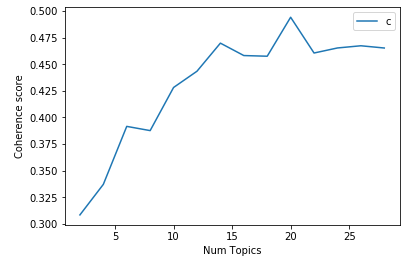
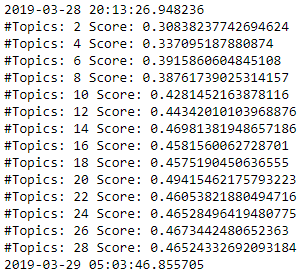


Figure 12: Coherence score comparison for varying k

While the result appears pleasing, a further investigation into the distribution of words within each topic highlights a lot of noises where typical words such as would, like and also appear repetitively across a few topics. This is likely due to a phenomenon which the team terms it as the “Overlaps”. There could be situations where the stopwords could not remove all the noises in the articles as stopwords function remove the most frequently used words. However, there is this region after the stopwords region where the words appear multiple times. These words are likely to belong to the noise words category but were not included in the stopwords package. Therefore, the team has performed an inclusion of a list of words before running the LDA model again.

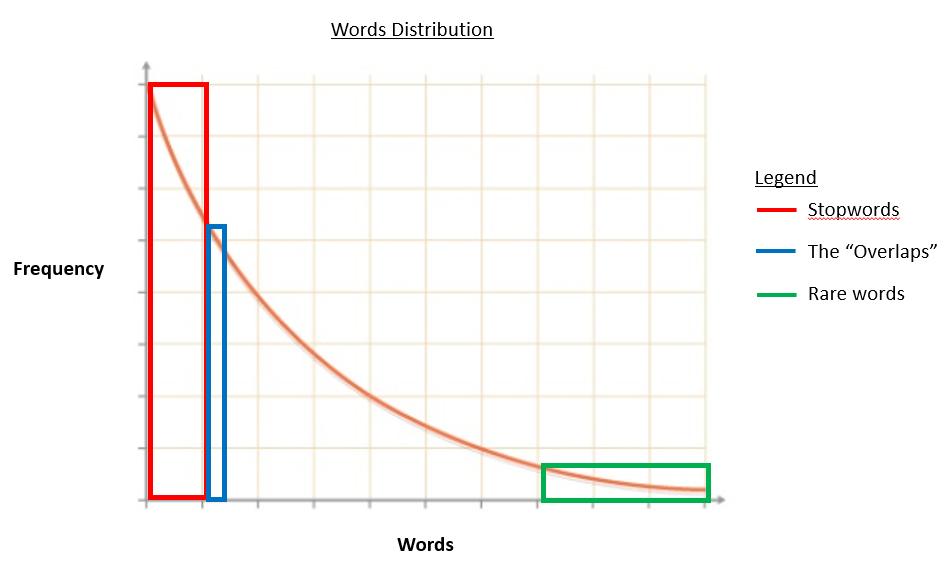


Figure 13: Depiction of typical word distribution in documents

The removal of additional noise words resulted in an improvement in coherence score as seen in the new run output. This is understandable given that the noise words are not directly relatable to a lot of the words in the topics.

However, we faced issues with choosing the optimal k topics in the new run as the coherence score displays three spikes in the run (10, 22, 28 topics).

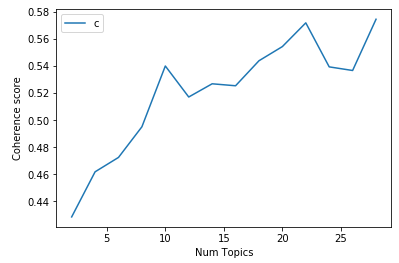
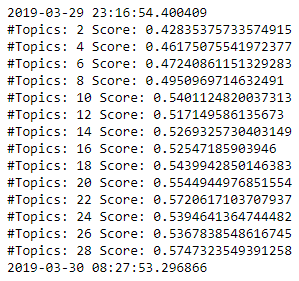


Figure 14: Coherence score comparison for varying k (run 2)

To resolve this dilemma, we utilise the perplexity score as well as human judgement to compare the three options. For the perplexity result using NLTK perplexity function, it is a negtive logartihm function and hence we need to look at the absolute score.

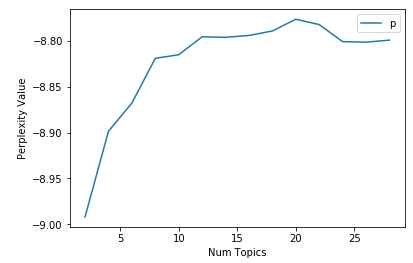
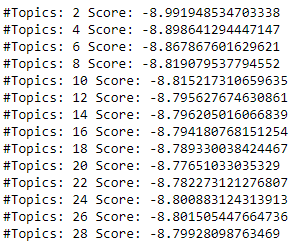


Figure 15: Perplexity score (run 2)

Through the combination of evaluation method, the final conclusion is to **use the 22 topics model.**

After selecting the k, we ran various plots such as the dominant topic per article, the topic that appeared the most among the articles to find out if there is any trend occuring. It is interesting to note that there seems to be a lot of articles on US politics. This is likely due to the run up to the 2016 US Presidential Election.

While the most dominant topic in an article may not work well for some articles, it was able to match many articles closely[[18]](#footnote-18). This highlights the effectiveness of LDA as a soft cluster techniques as the purpose of such topic modelling algorithm is to allow multiple appriopriate label to be tagged to a single article. Hence, there could be an article where multiple topics reside in that article which is our initial objective of this research.



Figure 16: Snapshot of Most Dominant Topic per Article

Next, we look into the potential overlaps between topics (Independence) as well as the topic distribution among all the articles using the pyLDAvis function (Figure 13). The size of the circle represents the topic representation among the articles i.e. the bigger the more frequent the topic appears in all the articles. The distance between the circles represents the dissimilarity in terms of topics based on the words. We observe that politics is once again the more frequent topics for the selected observation period.

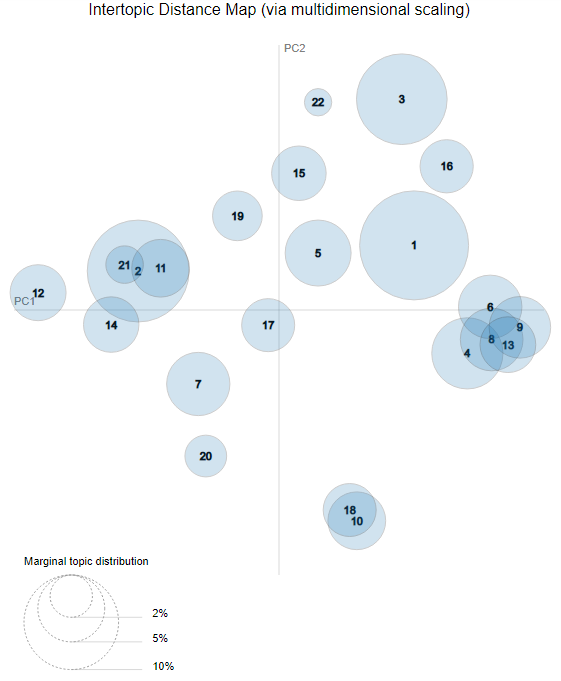


Figure 17: Topic Distribution Visualization

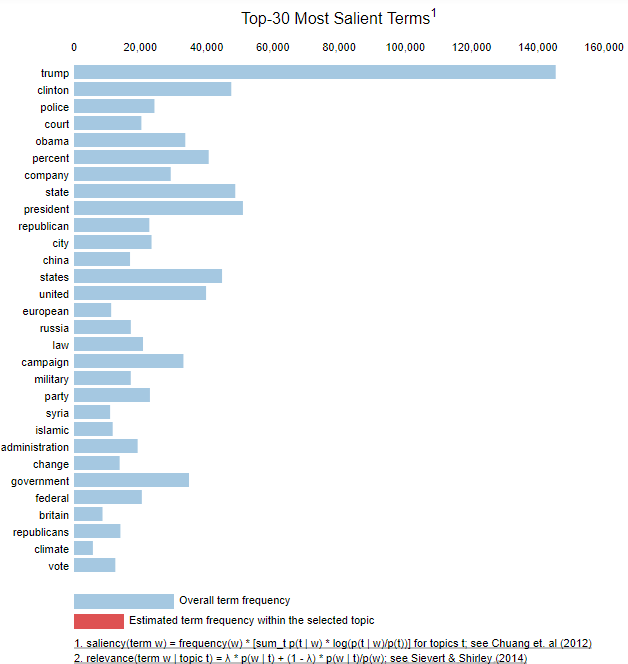


Figure 18: Topic Distribution Top Salient Terms

While most of the topics are relatively far apart from one another (indicating the independence of the topics), we do see some topics overlapping with the others. We investigated those cases using the top 10 words in each of the overlapping topics and noticed that there are indeed similar in terms of broad generic topics. However, our 22 topics split seems to perform segmentation into specific sub topics. This potentially alerts us that there could be an issue if we select a large k as topics might be overlapping. While we do see some overlapping, a cross check against a smaller k (in our case 10 topics) highlights that the 10 topics are too broad and encompasses a wide variety of words[[19]](#footnote-19). As a result, we continue to stick with our 22 topics model, bearing in mind that there were overlapping topics.

Next, the team analyzed the top 10 words per topic to derive the topic labels[[20]](#footnote-20). While most labels were generally easy to identify, we did face issues labelling the overlapping topics due to the relative similarity in words found in those topics.

Topic Names include:

Sports, Election, War, Entertainment, Community, Politics, Food, Crime, Business, Trivia, EU Politics, Science, Woman, Climate, Terrorism, US Politics, Technology, Accidents, Law, US Election, Arts, US Law

We discovered that some of the top 10 words have relatively lower frequency in the topic which enforces the idea that large number of topics k may lead to subtopic appearing.[[21]](#footnote-21)

## Article Summarization

### 6.2.1 Model Evaluation

Human Evaluation is utilized in this text analytics task as the other methods such as Rogue requires a sample of the correct summary to compare against.

When comparing across Abstractive and Extractive methods, we came to a quick conclusion to select Extractive method.

Due to the sheer size and intensive computing power required to run the NLP-based Abstractive method, we can only train 4000 articles. This leads to a poor outcome of the method as we noticed an error rate of 33%. Moreover, this technique produces a line summary which may be too short for readers.[[22]](#footnote-22)

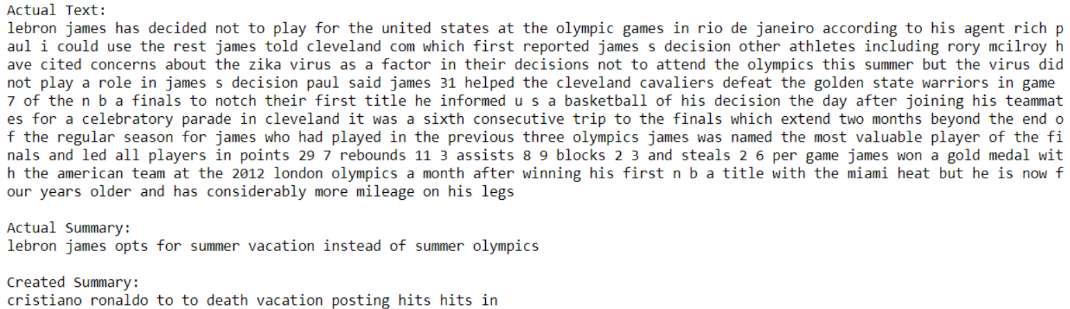


Figure 19: Bad Examples of Abstractive Summarization

We believe the discripancy in the quality of the summary generated is likely due to the insufficient training data. Due to the large amount of meaningless and unreadable data, we decide to explore abstractive method as future work and adopt extractive summary method instead.

When evaluating the Extractive Summarization, we use the following criteria to score the summary:

1) Grammar

2) Sentences’ order

3) Semantic analysis

4) Overall content

Both methods perform well in terms of grammar and sentence order. This proves that extractive method is able to match the articles concisely and accurately.

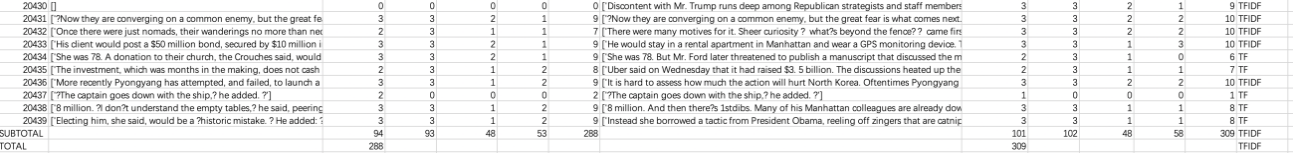
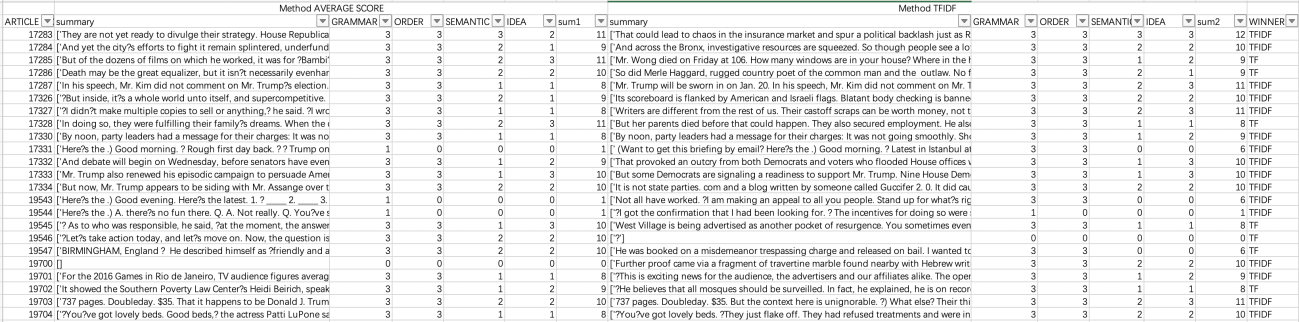


Figure 20: Evaluation Scores for Extractive Summarization

Between the two methods, the TFIDF method seems to perform better.

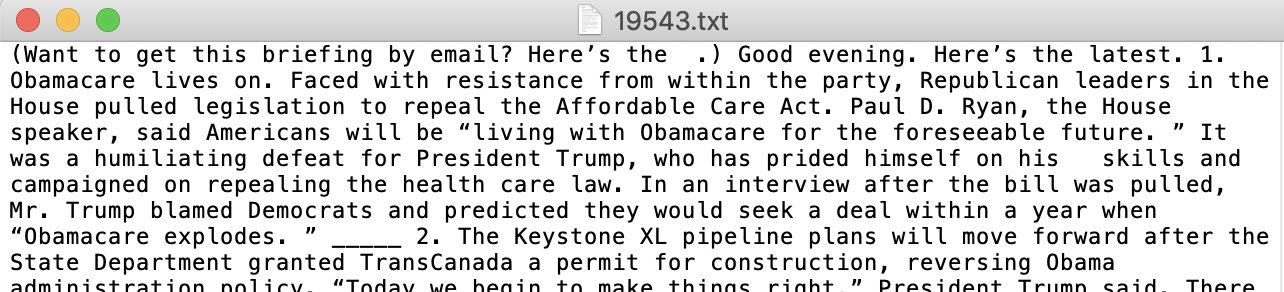
### 6.2.2 Insights

While most of the texts are able to generate rather concise and accurate summary, there are situations where the summary is full of dashes or even blanks in both methods.

page5image861212992

Figure 21: Example of summary with issues

We noticed that these are mainly from briefing articles such as the one shown below. This is likely due to the text form observed in such articles where there are space and annotations found.



Moreover, we also observed that a lot of blank summaries are due to initial articles processing. There are articles with inherent news-specific data issues such as no punctuation to classify sentence, three continuous blanks and unreadable formats. Hence, the team reprocessed the data to generate the new summary.

|  |
| --- |
| Code for new pre-processing |
| page6image861388848 |

Lastly, we have an interesting observation where some texts can only be summarized by the TF-IDF method. Our initial guess was that these corresponding articles are too short and hence are affected by the threshold we set to get the summary. However, we notice that other shorter articles are able to display reasonable summaries and TF-IDF method works. Hence, this deduction is overthrown.

We explore the original text and found out that the intrinsic nature of news articles being a multi-labels text form results in such an outcome.





Figure 22: Example of Average Scoring Method Weakness

# Limitations

The preliminary assessment of the methodology illustrates favourable outcome as we are able to identify relevant topic segmentation as well as summarizing the relevant articles. However, there are still limitations that the team faces during the period of testing.

For Topic Modelling:

* Overlapping of topics cause issue with labelling the topics as words can be very similar for the top 10 words.
* Identifying the optimal number of topics is subjective in this case as we do not see a consistent plateau even at 28 topics. This could be due to the nature of the data where the LDA models could identify sub topics among the articles.
* There could be real world implementation issue even though LDA model is an unsupervised machine learning technique. This is because it requires one to practice discretion in giving the generated topics a name. Hence, unless publications take an archived history dataset to set up the topics and do not show topics based on latest news, users will have to wait for the generation of such UI and cannot obtain immediate news update.
* We only assume unigram words in our LDA model.

For Summarization:

* For Abstractive method, the difficulty level for machines to understand the ‘meaning’ of words using neural encoder-decoder model remains high. One reason is because there are lots of rare words (appeared only once) in our dataset and these rare words are not replaceable and of crucial importance in understanding a sentence. Another reason is news would require a strong contextual knowledge to accurately interpret the underlying meaning in articles.
* For Extractive methods, they do not take into account of the context. Moreover, there are situations where we were not able to find the most important word accurately using the weight method as well as to determine the ideal threshold.

# Future works

This paper explores basic test analytics task in solving a real-world issue of clustered and long news publications. While the initial outcome looks favorable, there are rooms for improvements for future works.

1. We can explore the use of other types of topic modelling techniques such as using bigrams or trigrams as input for LDA model or the use of LDA and Latent Semantic Indexing combination model or LDA mallet.
2. We can explore other tools to automatically label the generated topics so that it can be used real-time with the latest news articles.
3. For abstractive summary method, we currently have limited knowledge on machine learning and neural network. In the limited time of preparation, we mostly looked at the mature code written by professionals in github and tried to apply the mature code into our own dataset. We plan to look deeply into the field of machine learning and modify existing algorithms.
4. For extractive summary method, we can explore the use of graphing theory instead of the frequency-driven approach as it displays articles as connected graph and indicate the similarity between articles. This can potentially resolve the issue of multiple topics in news articles.[[23]](#footnote-23)

# Conclusion

The main issue with current news digitalization revolves around an inconvenient platform where users are unable to easily find what they want to read as well as the fact that articles are usually detailed and long form. In this research, we have explored the use of text mining techniques, particularly topic modelling and text summarization, to resolve the gap between the current situation and readers’ preferences. While we were successful in achieving the topics split and summarization, we do see limitations with our methodologies which will require further enhancement for it to perform better.

# Project Refelctions

Jiang, Nan

Teamwork is quite important in such a big project. Thank all my team members! Combing the power of human and machine is very exciting!

Tey, Wei Jie

Analysis of unstructured data is useful in the case of news articles summary. However, there are limitations currently to the things machine can do as seen in this project

Wang, Yechun

Machine learning is really interesting and worth learning. Despite its powerful functionalities, it is so difficult.

Wang, Yuxin

Manual evaluation is still very crucial in judging text analytics results.

Yao, Yu

For text analysis, no separate analysis model, like topic modelling and summarization can influence interactively.

Zhu, Chenzi

For text summarization, extraction doesn't perform well in general, different word-score methods didn't make much discrimination.

# References

(n.d.). Retrieved from Jupyter: https://jupyter.org/

(n.d.). Retrieved from Kaggle: https://www.kaggle.com/snapcrack/all-the-news/

(n.d.). Retrieved from Tensorflow: https://www.tensorflow.org/

An, M. (2016, June 25). *The Future of Content Marketing: How People Are Changing the Way They Read, Interact, and Engage With Content*. Retrieved from HubSpot Research: https://blog.hubspot.com/news-trends/the-future-of-content-marketing

BIALIK, K., & MATSA, K. E. (2017, October 4). *Key trends in social and digital news media*. Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2017/10/04/key-trends-in-social-and-digital-news-media/

CHEN, A. (2014, Jan 08). *There Can Never Be Enough White Space*. Retrieved from Slate: https://slate.com/culture/2014/01/new-york-times-website-redesign-reviewed.html

Feeney, M. (2017). *What Can Text Mining Reveal about the Use of Newspapers in Research?* IFLA International News Media Conference.

Glader, P. (2017, February 01). *10 Journalism Brands Where You Find Real Facts Rather Than Alternative Facts*. Retrieved from Forbes: https://www.forbes.com/sites/berlinschoolofcreativeleadership/2017/02/01/10-journalism-brands-where-you-will-find-real-facts-rather-than-alternative-facts/#43d45635e9b5

GOTTIPATI, S. (n.d.). Text Mining and Language Processing: Topic Modelling. Singapore.

Kalogeropoulos, A., & Cherubini, F. (n.d.). *The Future of Online News Video*. Retrieved from Digital News Publications: http://www.digitalnewsreport.org/publications/2016/future-online-news-video/

LAWLOR, J. (2018, January 10). *The top 10 newspaper publications in the US*. Retrieved from Muck Rack: https://muckrack.com/blog/2018/01/10/the-top-10-newspaper-publications-in-the-us

MYERS, G. (2017, February 20). *10 Most Reliable News Sources*. Retrieved from https://www.toptenz.net/10-reliable-news-sources.php

Pleplé, Q. (2013, May). Retrieved from http://qpleple.com/perplexity-to-evaluate-topic-models/

Röder, M., Both, A., & Hinneburg, A. (n.d.). *Exploring the Space of Topic Coherence Measures.* Retrieved from http://svn.aksw.org/papers/2015/WSDM\_Topic\_Evaluation/public.pdf

Schmied, T. (n.d.). Retrieved from GitHub: https://github.com/thomasschmied/Text\_Summarization\_with\_Tensorflow

Sciforce. (2019, January 23). *Towards Automatic Text Summarization: Extractive Methods*. Retrieved from Medium: https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715

Silipo, R. (2019, February 18). *AI Zone: LDA for Text Summarization and Topic Detection*. Retrieved from DZone: https://dzone.com/articles/lda-for-text-summarization-and-topic-detection

Singh, D. (2018, March 31). *10 Most Popular Newspapers In The World You Should Read If You Can*. Retrieved from https://topyaps.com/top-10-newspapers-of-the-world/

Strether, L. (2018, August 29). *Why I Hate The New York Times Front Page Redesign*. Retrieved from naked capitalism: https://www.nakedcapitalism.com/2018/08/why-i-hate-the-new-york-times-front-page-redesign.html

*Topic Modeling with Gensim (Python)*. (n.d.). Retrieved from Machine Learning Plus: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

Wong, L. (n.d.). *Automatic News Summarization and Extraction.* Imperial College Dept. of Computing.

Yang, T.-I., Torget, A. J., & Mihalcea, R. (n.d.). *Topic Modeling on Historical Newspapers.*

Yzaguirre, A., Smit, M., & Warren, R. (2016). *Newspaper archives + text mining = rich sources of historical geo-spatial data.* IOP Conference Series: Earth and Environmental Science.

# Acknowledgments

We would like to thank Kaggle for sharing and hosting the data sample. We would also like to express our utmost gratitude towards Prof. Swapna Gottipati, Associate Professor of Information Systems (Education) at Singapore Management University (SMU) for her valuable mentoring and advice. This research is done with the support from SMU as part of the Masters of IT in Business course.

# Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Jiang, Nan

Enterprise: Singapore Management University

E-mail: nan.jiang.2018@mitb.smu.edu.sg

Name: Tey, Wei Jie

Enterprise: Singapore Management University

E-mail: weijie.tey.2018@mitb.smu.edu.sg

Name: Wang, Yechun

Enterprise: Singapore Management University

E-mail: yechun.wang.2018@mitb.smu.edu.sg

Name: Wang, Yuxin

Enterprise: Singapore Management University

E-mail: yuxin.wang.2018@mitb.smu.edu.sg

Name: Yao, Yu

Enterprise: Singapore Management University

E-mail: yu.yao.2018@mitb.smu.edu.sg

Name: Zhu, Chenzi

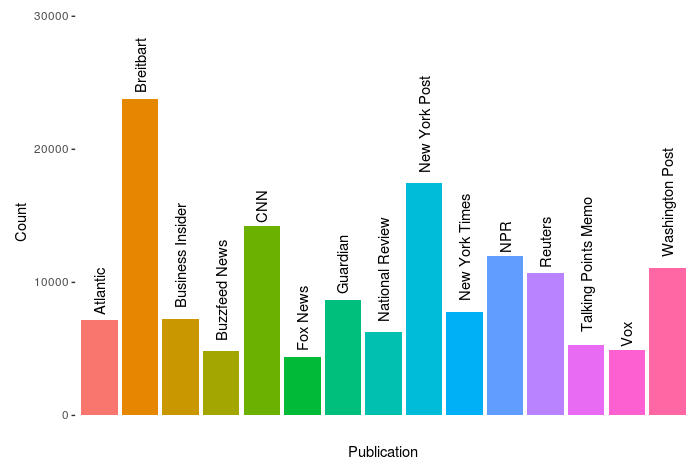
Enterprise: Singapore Management University

E-mail: chenzi.zhu.2018@mitb.smu.edu.sg

The respective brand and product names are trademarks of their respective companies.

# Appendix

## Appendix I: Breakdown of Publications



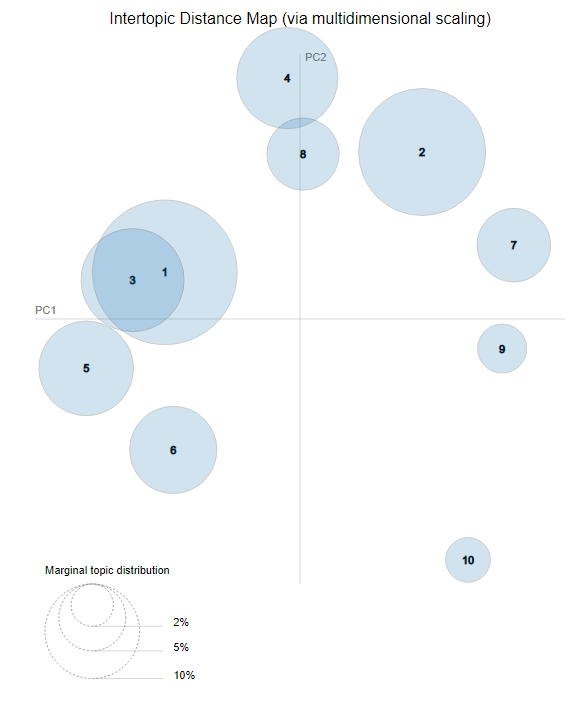
## Appendix II: Dominant Topic in An Article



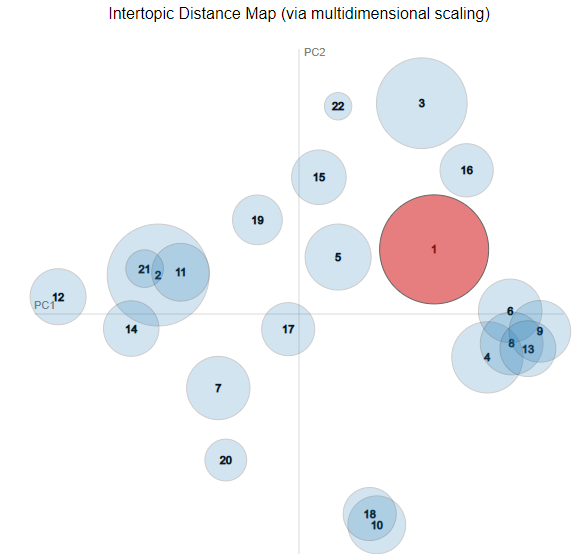
## Appendix III: The Frequency of Appearance for the Dominant topic

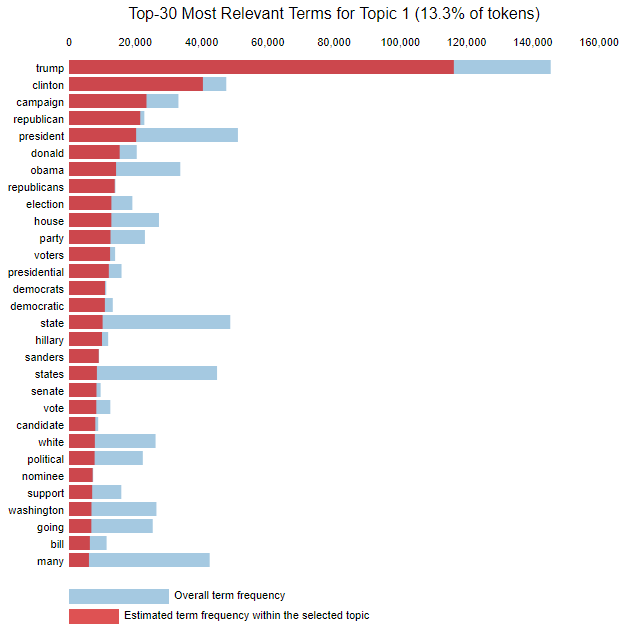


## Appendix IV: Topic Distribution Visualization for 14 topics model



## Appendix V: US Election most popular

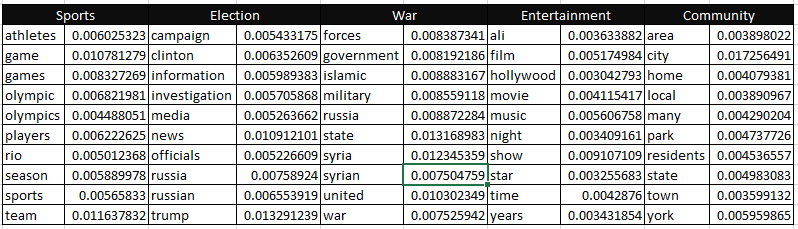


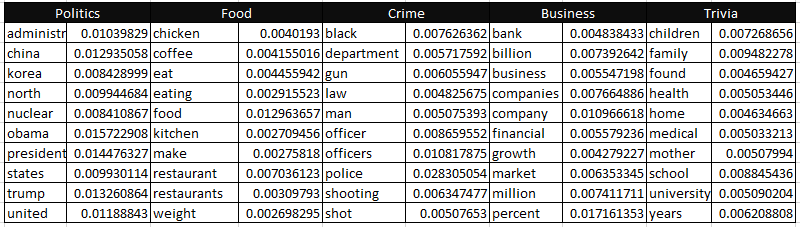


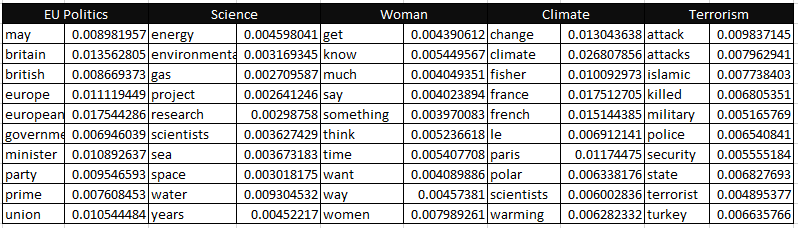
## Appendix VI: Identifying the topic Labels

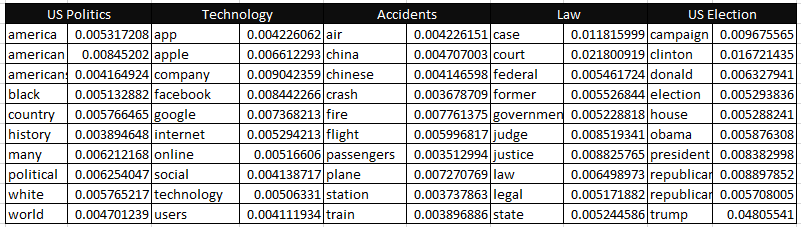


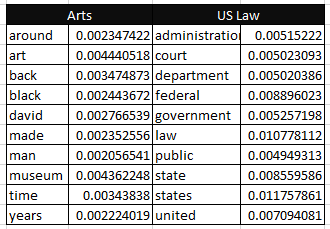
## Appendix VII: Top 10 words per topic





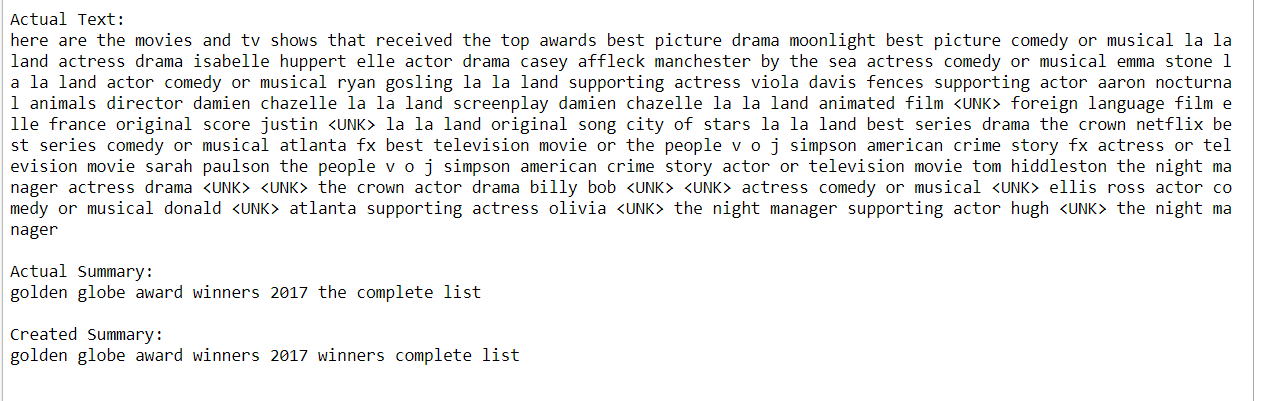


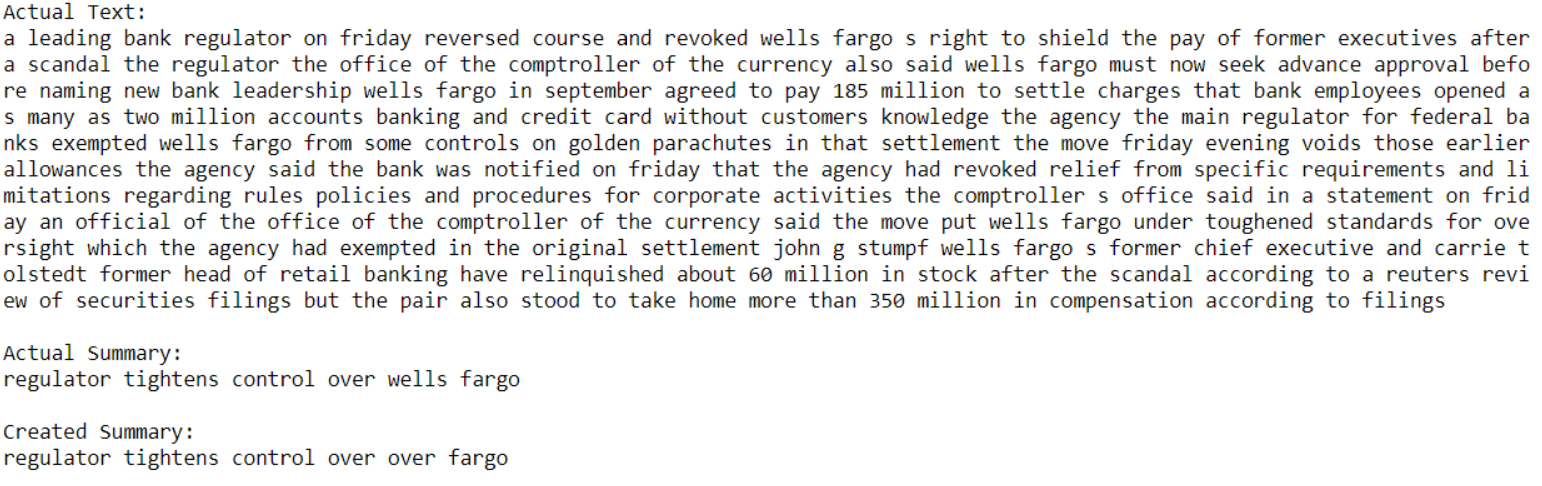




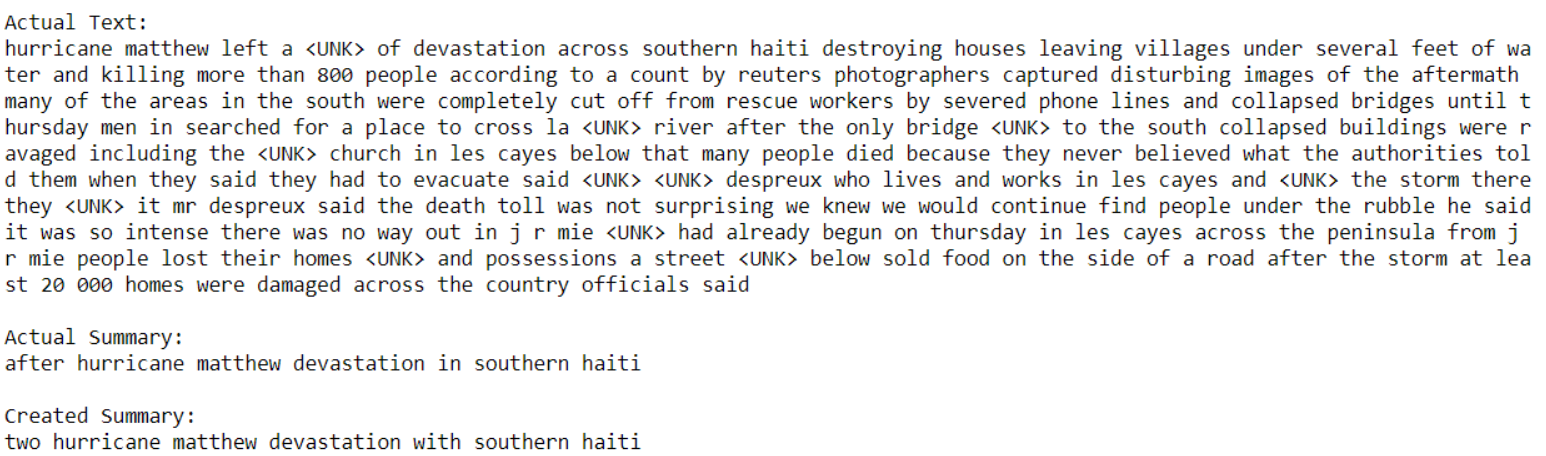
## Appendix VIII: Sample of Abstractive Summarization

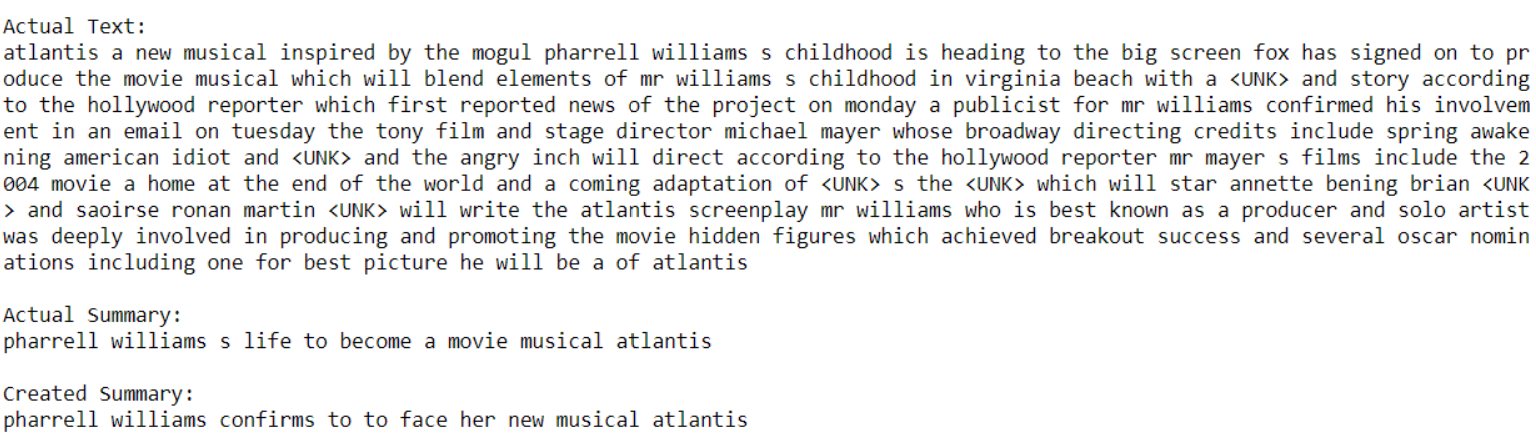
**Good examples:**





**Readable examples:**





1. (BIALIK & MATSA, 2017) [↑](#footnote-ref-1)
2. (An, 2016) [↑](#footnote-ref-2)
3. (Kalogeropoulos & Cherubini, n.d.) [↑](#footnote-ref-3)
4. (CHEN, 2014) (Strether, 2018) [↑](#footnote-ref-4)
5. (Yzaguirre, Smit, & Warren, 2016) (Yang, Torget, & Mihalcea) [↑](#footnote-ref-5)
6. (Feeney, 2017) [↑](#footnote-ref-6)
7. (Wong) [↑](#footnote-ref-7)
8. (Kaggle, n.d.) [↑](#footnote-ref-8)
9. Appendix I: Breakdown of Publications [↑](#footnote-ref-9)
10. (Glader, 2017) (LAWLOR, 2018) (Singh, 2018) (MYERS, 2017) [↑](#footnote-ref-10)
11. (Jupyter, n.d.) [↑](#footnote-ref-11)
12. (GOTTIPATI) [↑](#footnote-ref-12)
13. (Topic Modeling with Gensim (Python), n.d.) [↑](#footnote-ref-13)
14. (Schmied, n.d.) [↑](#footnote-ref-14)
15. (Tensorflow, n.d.) [↑](#footnote-ref-15)
16. (Pleplé, 2013) [↑](#footnote-ref-16)
17. (Röder, Both, & Hinneburg) [↑](#footnote-ref-17)
18. Appendix I: Dominant Topic in An Article [↑](#footnote-ref-18)
19. Appendix III: Topic Distribution Visualization for 14 topics model [↑](#footnote-ref-19)
20. Appendix V: Identifying the topic Labels [↑](#footnote-ref-20)
21. Appendix VI: Top 10 words per topic [↑](#footnote-ref-21)
22. Appendix VIII: Sample of Abstractive Summarization [↑](#footnote-ref-22)
23. (Sciforce, 2019) [↑](#footnote-ref-23)