Topic Modeling

Graduate Final Report

By:
Suchartee Kitisopakul

CWID: 891481764

Project Advisor: Paul Salvador Inventado

Department of Computer Science
California State University, Fullerton

December 11, 2019

Table of Contents

The List of Figures

The List of Tables

Section 1: Introduction

Problem Domain

Knowledge is power. News is one of the knowledge sources that humans can learn from. It keeps people up to date by providing information about events and what may affect public or individuals. Since news indicates the market/world trends, many people gain benefits from reading and analyzing news. Stocks is an investment that can fluctuate according to demand and supply in the world. Some security companies, such as Phatra Securities PCL, Bangkok, Thailand, appraise news as a valuable data source. Current world situation can affect the stocks. For example, the volume of rainfall that can lead to drought can negatively affect the agricultural companies. However, reading all available news is a burdened task. For thorough non-fiction reading, human takes about 175-300 words per minute or wpm (Brysbaert, 2019). Hence, reading and analyzing news from many available news sources (both local and international news) can be very time-consuming due to the length of each news and the amount of news available each day. It can also be a waste of time especially when spending time on some news that are irrelevant to the purposes or objectives.

Proposed Solution

Instead of reading through every news, this project tends to use the machine to relieve the problem by eliminating presumably irrelevant news. Machine can outperform humans in reading. Hence, the amount of reading time will be significantly decreased. To eliminate the irrelevant news, machine learning, more specifically, unsupervised learning technique will be used. Unsupervised learning requires no labelled data for machine to be trained, otherwise, the machine is learning by itself. Instead of examining all the data through keywords, finding the theme and then data that is related to that theme is more preferable (Blei, 2012). Therefore, labeling is not required for the data because the data can reveal the common traits as the patterns or themes. This thematic technique is called topic modeling. Topic modeling applies the idea of each data (news) contains latent (hidden) topics that come from the words within document (Zhou, Zou, & Chen, 2014). The topic modeling algorithm statistically discovers the theme from the large unstructured collections of data that may be impossible for humans to annotate nor to label (Blei, 2012). It can produce different patterns or themes by clustering key features of each narrative data (Pröllochs & Feuerriegel, 2018). Applying topic modeling will uncover the latent topics that the news are about. Figure 1 shows a diagram of the proposed solution pipeline.

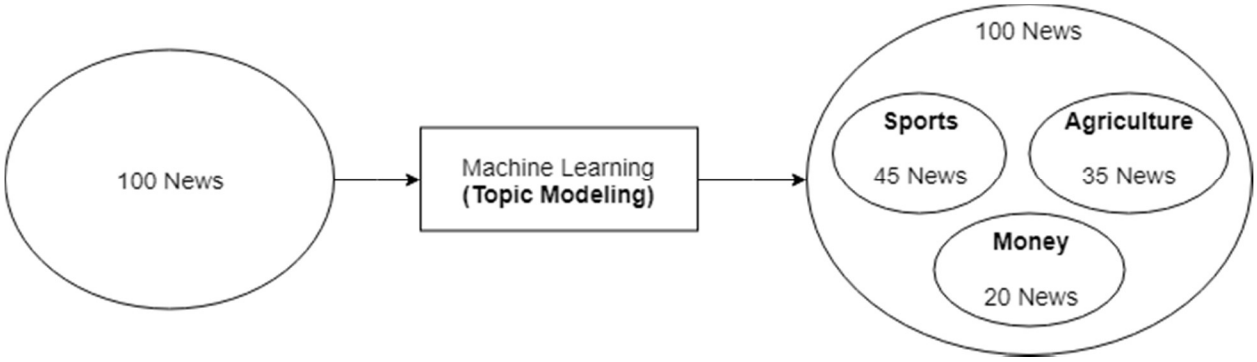

Figure 1 The Process Pipeline

According to Figure 1, there are 100 news released each day. Applying the topic modeling on all 100 news to find the themes of the news. At the end, there are 3 themes uncovered. From that the Phatra staff can then choose which news they should read for further analysis. If Phatra holds the Bank of

America securities, then the sports and agriculture themes are not relevant as much as money; instead of reading 100 news, the staff reads only 20 news. It shows that topic modeling can relieve the stated problems. In most cases, latent topics overlap with other topics. Hence, one document may have several latent topics. Giving the example from Figure 1, some documents may have both sports and money topics with probability of 40% and 60% respectively, or money and agricultural topics with probability of 20% and 80% respectively. Hence, the results would be the keywords for documents.

Section 2: Methodology

Data Preparation

The original links to access the news contents are from Mirsa (2018). The dataset is then retrieved from HuffPost website using Python BeautifulSoup library; 35,056 news are successfully retrieved out of 200,000 news. The fundamental cleaning process such as removing punctuation, numbers, and extra whitespaces and lowercasing are done prior the preprocessing step. The preprocessing steps also include tokenizing each news text to word corpus, stemming each corpus to base form of that word, removing stopwords, removing proper nouns, TF-IDF, and Bi-gram.

Stemming removes the derivational affixes. For instance, ‘cats’ is stemmed to ‘cat’ or ‘accessible’ to ‘access’. The project uses stopwords lists that are provided by Gensim and NLTK libraries. The proper noun tokens are tagged by Gensim library and are removed from the corpus. However, some proper nouns may not be properly taken out due to the uncanny way of news sentence structure. TF-IDF is applied to filter the corpus that appear in document too frequent or too rare. Instead of only a word (Uni-gram), this project uses two-adjacent words in a sequence (Bi-gram) for the corpus as well. In total, 70,000 corpus is reduced to 20,000 corpus for the implementation.

Implementations

There are a couple of topic modeling algorithms available nowadays such as Latent Dirichlet Allocation (LDA). This project implements the Latent Dirichlet Allocation (LDA). LDA is a topic model that statistically calculates the probability distribution among documents (Blei, 2012). Blei (2012) describes that LDA can identify the latent (hidden) theme within the collection of documents. It assumes that each document may contain multiple topics and that document can belong to different topics (Gui & Wang, 2017). LDA relies on two matrices to find latent topics; the word-topic matrix and the document-topic matrix. Number of topic must be defined by humans. Hence, humans need to find the optimal number of topic.

Topic coherence is the measurement used to calculate the score of the topic coherency. It calculates the sum of pairwise scores of each word in the top N words and represents the coherence score (Pleplé, 2013). This score represents the quality of top N words that the model generated. If the model is not good, it can imply that the number of topics that the humans input may be too large or too small. Figure 2 shows the coherence equation that is used to calculate the score and Figure 3 illustrates the comparing map of each top N words.

$$Coherence = \sum_{i<j} score(w_i, w_j)$$

Figure 2 Coherence Equation (Steven, Kegelmeyer, Andrzejewski, & Buttler, 2012)

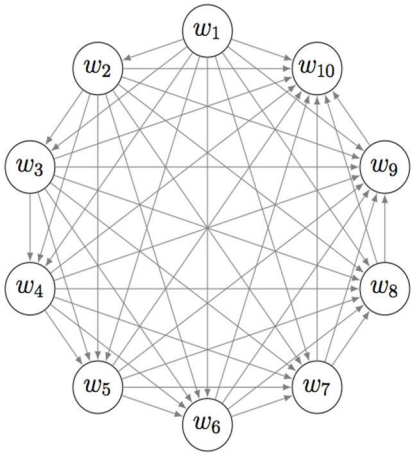

Figure 3 Topic Coherence (Pleplé, 2013)

Evaluation

To relieve the problem domain, the speed of machine reading all the news outperforms humans. However, to find the right number of topics and to evaluate the good model are different. Once the topics are revealed by the model, then the irrelevant news can be opted out so that the person in charge of analyzing news only focuses on the related news.

To find the right number of topics sending to the LDA model, the topic coherence is used. As described earlier in ‘Implementations’ section, topic coherence ensures the quality of model by

determining the quality of the top N words generated by the model. The score is used to see whether the number of topics is optimal or not.

There are many different automated metrics that Gensim library offers such as CV, Umass, UCI, and more. In this experiment, CV, Umass, and UCI are used to compute the coherence score.
UCI metric, or extrinsic measure, counts the word co-occurrence frequencies between external corpus, such as Wikipedia, with the top N words (Steven, Kegelmeyer, Andrzejewski, & Buttler, 2012).
Umass, or intrinsic measure, defines frequencies based on document co-occurrence (Steven, Kegelmeyer, Andrzejewski, & Buttler, 2012). CV calculates the cosine similarity measure and Normalized Pointwise Mutual Information (NPMI) with the top N words (Röder, Both, & Hinneburg, 2015).

These automated measurements solely may be inaccurate. Hence, human judgement comes to relieve the issue. To align with the model performance and to replicate the similar fashion, word intrusion is not utilized in the experiment. The human rating is the method used in this experiment as human judgement. In the same manner of the topic coherence where the machine compute the similarity of top N words from each topic. Human rating is done by five different people. Each person give scores of how well each topic group with other top N words.

Krippendorff's alpha or Krippendorff's coefficient is statistically reliability measurement for the agreement among multiple raters. Values ranges from 0 to 1 where 1 is a perfect agreement and 0 is perfect disagreement. The considerable alpha value starts from 0.667 to draw conclusions (Krippendorff, 2004). However, Krippendorf (2004) suggests that $\alpha \geq .800$is an acceptable level of agreement. This number is used to confirm the agreement of human judgement.

Scopes & Limitations

Misra's (2018) news are from Huffpost, American news website, from 2012 to 2018. This dataset is labelled, however, the label will not be used in the project since the topic modeling is unsupervised machine learning. Also, the project mainly runs on Jupyter Notebook as it supports documentation and

python language. If the dataset is humongous, other tools may be needed for better performance such as Hadoop or Spark.

There are few limitations that have occurred during the project. The given URL links to access the news contents are not available at the time of the project. Furthermore, some URLs are removed or are redirected to another URL that has a different set of HTML elements. The method used to retrieve the news content from the given URLs in this project is the BeautifulSoup Python library. This library allows to pull the data from the HTML files based. Hence, the specific element must be given in order to find the location of the news content in each URL. Since, the HTML elements location is changed, the news content cannot be retrieved properly. Therefore, only 35,056 news can be retrieved from the original size of approximately 200,000 news.

Section 3: Results

In the experiments, 9 LDA models (Model_0 to Model_9) are created with different number of topics; from 2 topics to 10 topics. Each model shows top 10 words that represent each topic. The topic coherence measure is applied for determining the optimal number of topics; both intrinsic and extrinsic measures are used. CV, Umass, and UCI are the automated metrics. The only problem with these metrics are they have different intervals. As seen in Figure 4, the trend is there, but one of them is rather straight compared to others.

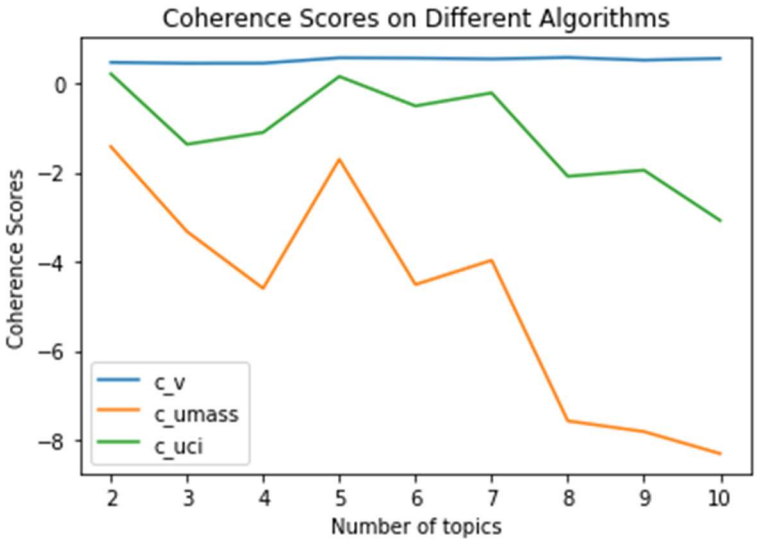

Figure 4 Topic Coherence Metrics without Normalization

CV is known to have the range from [-1, 1] while others have [-∞, ∞]. Hence, CV is shown as a straight line whereas other metrics keep declining or inclining for other cases. Normalization is required in order to compare these metrics. Tanh function would do the trick as seen in Figure 5.

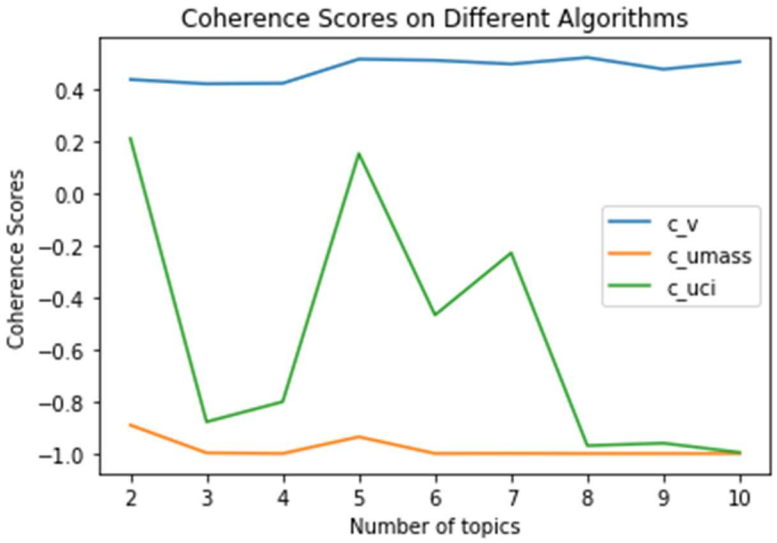

Figure 5 Topic Coherence Metrics after Applying Tanh Function

Tanh function normalizes their intervals to [-1, 1]. However, it does not tell much about which one is good as some metrics portray the straight line. This goes back to the same problem. This is because Tanh normalization is done so that they all have the same interval. Rather than focusing on the same interval numbers, focusing on the ratio each value actually sits on. For instance, the maximum value represents 100% and minimum value represents 0%. With this normalization, 3 metrics are comparable as seen in Figure 6.

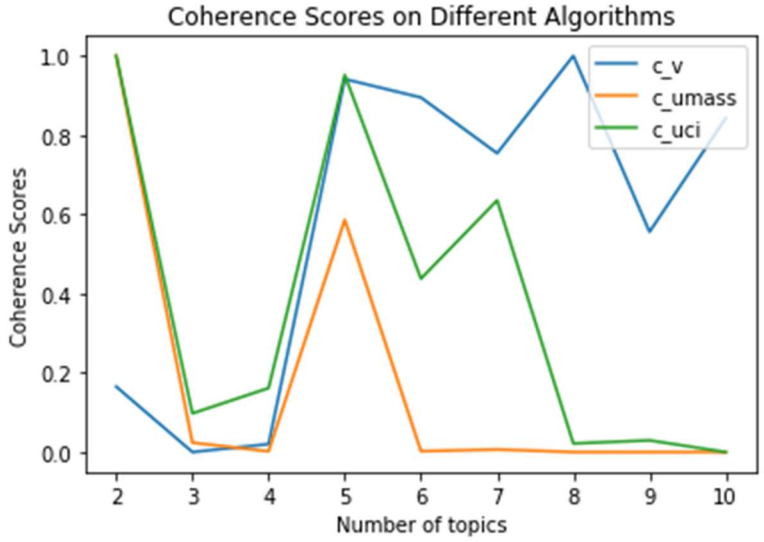

Figure 6 Topic Coherence Metrics after Normalization

Based on several automated evaluations, the possibly final model is Model_3 (5 topics) as it shows the significant evaluation score. However, humans still need to decide what the optimal number of topics is. As for humans to judge and decide, top 10 words comes to help humans understanding the words behind each topic. Table 1 shows top 10 words of all 5 topics in Model_3.

Table 1 Number of Documents per Topic for Model_3

Ranking	Topic	Number of Documents	Top 10 Words
1	Topic_2	15,818	['election, president, state, country, vote, campaign, police, government, attack, official']
2	Topic_4	11,188	['film, woman, video, love, star, movie, like, fan, host, actor']
3	Topic_1	7,758	['company, percent, health_care, tax, million, climate_change, state, school, plan, woman']
4	Topic_0	271	['investigation, email, intelligence, information, probe, campaign, special_counsel, committee, hacking, document']
5	Topic_3	21	['percent, marijuana, margin_error, based_margin, survey, cannabis, model_based, assumption_wrong, survey_error, error_rest']

Human judgement is performed to assure whether the automated metrics are reliable. 5 people rate all 9 models whether each model's top 10 words of each topic belong together in the same topic or not. Once more, ratio normalization is used because each rater has different minimum score and maximum score. The normalized result is shown in Figure 7.

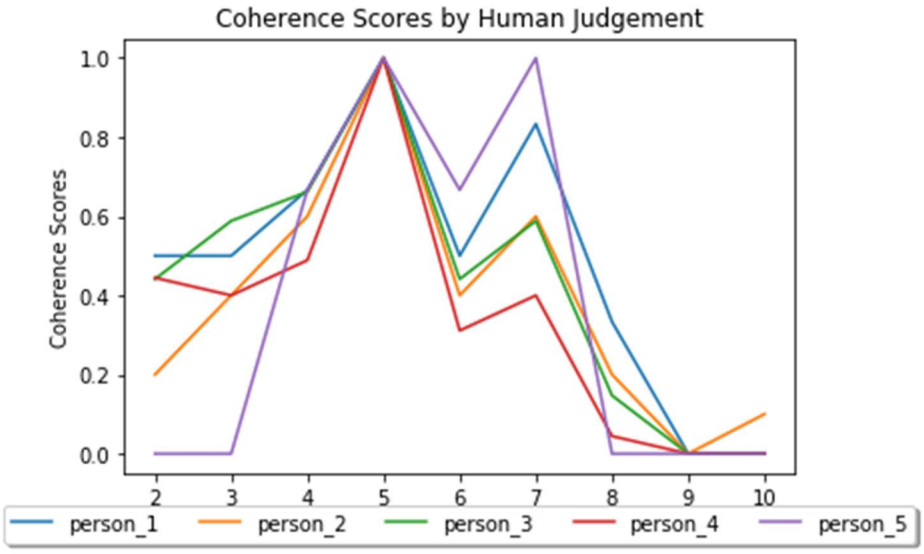

Figure 7 Coherence Scores by Human Judgement

Having many people to rate can sometimes unreliable. Krippendorff alpha value is used as a reliability measure for human judgement. The calculated Krippendorff value (alpha value) is 0.82. It is just above the acceptable value as suggested by Krippendorf (2004). This means all 5 raters have similar agreement. As Figure 7 shows, most of the raters have similar judgement towards the coherence of each topic for each model. The trend of average human judgement is illustrated in Figure 8. Again, Model_3 (5 topics) is still on the lead.

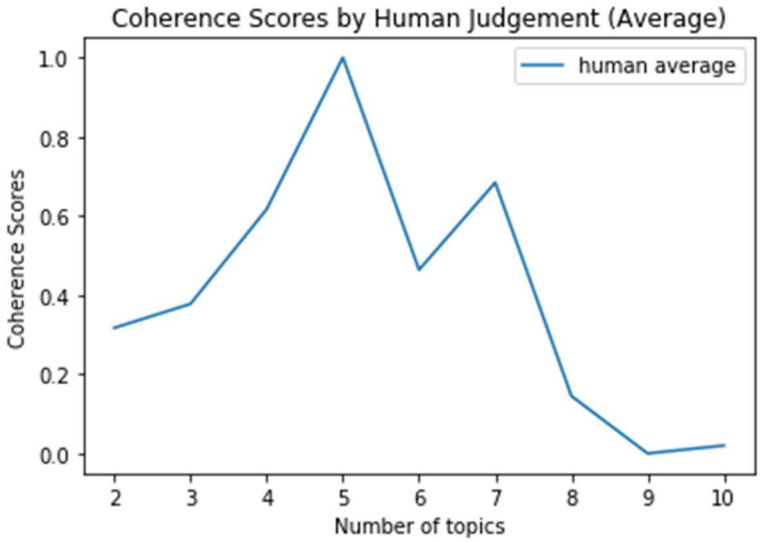

Figure 8 Average Human Judgement

Figure 9 displays all metrics used for these models. The average human judgement (red line) is somewhat aligned with the automated metrics. In addition, each of automated metrics are calculated with human judgement using Krippendorff's alpha to make a better conclusion. Figure 10 to Figure 13 illustrate the results from each metric. Krippendorff's alpha value for average automated metrics and average human judgement is 0.478, for CV and average human judgement is 0.14, for Umass and average human judgement is 0.188, and for UCI and average human judgement is 0.667. Though the value is lower than the acceptable, this score is enough to draw a conclusion as mentioned by Krippendorff (2004). Figure 13 shows the UCI metric with the average human judgement are being compared, they are nearly aligned.

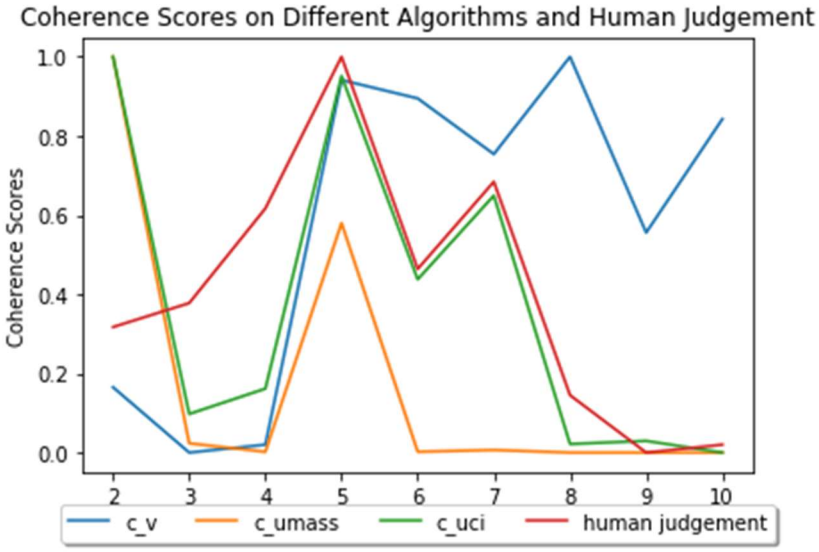

Figure 9 Topic Coherence on All Metrics

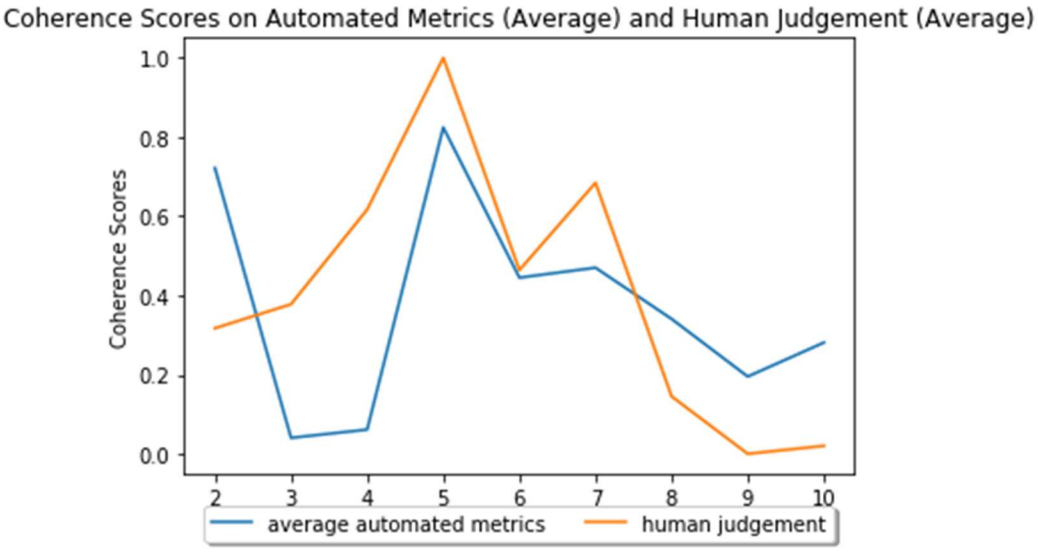

Figure 10 Topic Coherence on Average Automated Metrics (CV, Umass, UCI) and Human Judgement

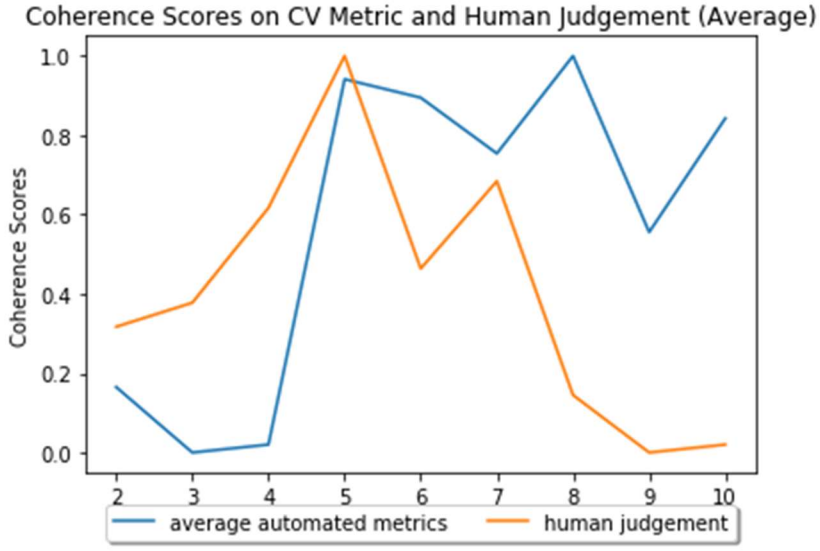

Figure 11 Topic Coherence on CV Metric and Human Judgement

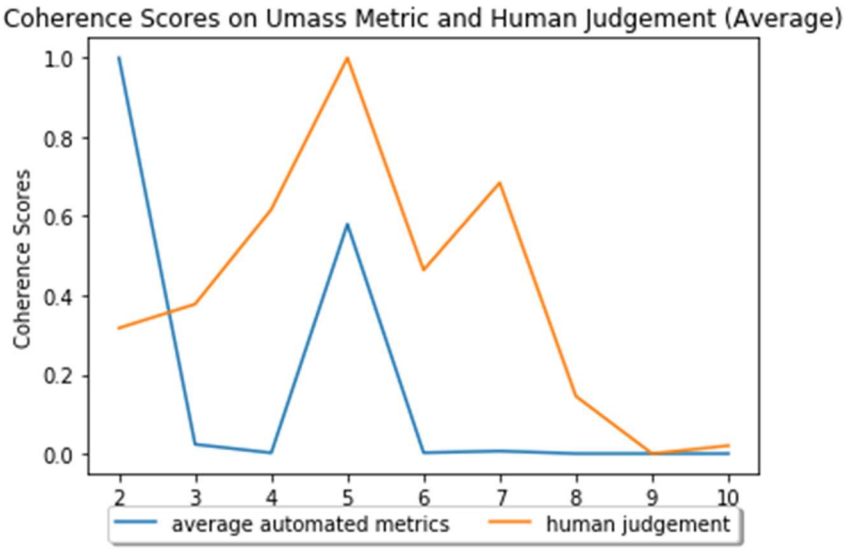

Figure 12 Topic Coherence on Umass Metric and Human Judgement

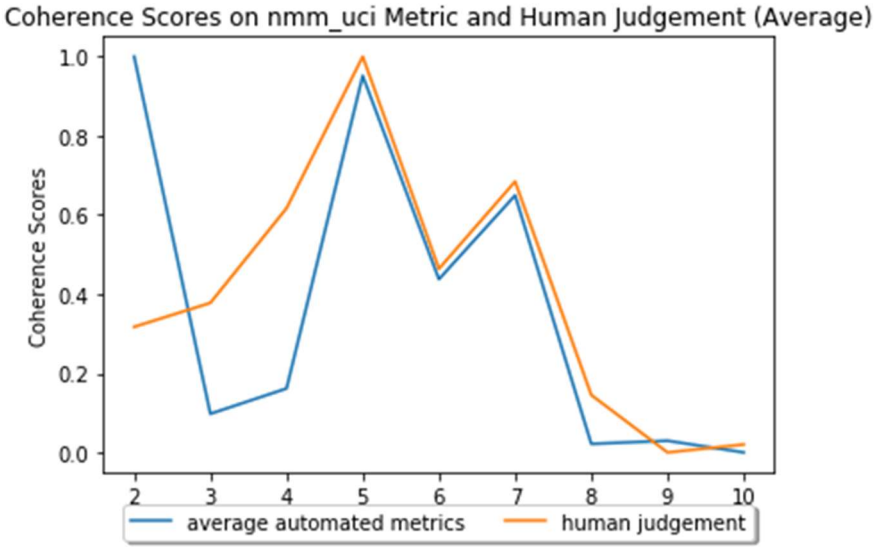

Figure 13 Topic Coherence on UCI Metric and Human Judgement

Section 4: Discussion and Conclusion

Discussion

Figures shown in the results section illustrate the results of this project. It is tricky when it comes to finding the optimal number of topics. Since the machine cannot really decide how many number of topics the model should generate, human decision comes to involve. This project tries to solve the domain problem by reducing the human interaction as well as to resolve the issues. A few number of topics generally broad and vague while many topics can be too specific. Each topic can be overlapped; one document can belong to many topics. Model_0 (2 topics) has the high coherence score and it means that each topic is fairly coherent. Top 10 words can describe and shape the story of that topic individually. However, 2 topics in Model_0 are totally different from each other and that makes the generated topic too vague. In contrast, Model_9 (10 topics) is too specific or the clustered words of each topic do not make semantics sense. Hence, choosing the number of topics need reliable evaluation.

Finding the good coherence score is delicate. Coherence score is defined by calculating the pairwise words among top 10 words. The full coherence score is not the best choice to choose the optimal number of topics. When the score is full, it implies that those two words are almost the same. For instance, compare the word ‘cat’ and the word ‘cat’; the coherence score is full because those two words are identical. So, the justification for coherence score should be fairly high in between. According to the automated coherence scores, Model_3 (5 topics) shows the promising result. To see whether the suggested model is good enough, the project also include human judgement for better adjudication. After normalization, the graph shows that the human judgement reasonably aligns with the automated metrics. Without normalization, it would have been hard to compare among these metrics. As the Krippendorff alpha of 0.82 for human judgement and automated metrics, it implies that the automated metrics are trustworthy to use. To suggest which metrics to use for the future work, or other projects, Krippendorff’s between each metric and human judgement has been made. The average coherence score of all automated metrics shows that not all metrics are good to use. However, UCI metric reveals the best Krippendorff’s

value of 0.667 out of other 2 metrics. Though the suggested alpha value by Krippendorff is 0.8, 0.667 is sufficient to draw a conclusion. That result can ratify that UCI metric is the best metric to use rather than Umass and CV coherence algorithms. UCI's alpha value aligns well with the human judgement because the graph illustrates that Model_3 is a potential final model, this statistically confirms that Model_3 is a strong conclusion. Interestingly, UCI is the extrinsic measure. Extrinsic measure compares the pairwise of words with external source; In this case, it is Wikipedia. From the results, UCI measure performs better to find the coherence of topic. One possible reason is intrinsic measure compares with its own data. The vocabulary pool is limited. Also, intrinsic measure compare words among the generated corpus that could possibly be biased. Pleplé (2013) claims that intrinsic measure rather correlates with human judgement.

Most of the data are about politics as Table 1 has shown the number of documents per topic. Data used for topic modeling has a very big impact on the model. Different data can produce different reesults. Since the data from HuffPost is mostly about politics (13,944 news out of 35,056 news), the result from LDA model is reasonably accurate and aligned with the actual data.

Conclusion

As the goals of reducing human labor and time, topic modeling can help resolve the issue. Topic modeling is unsupervised machine learning where there is no correct answer to the result. Human judgement is the important role in order to decide whether the model is good or not. Even though topic modeling performs cluster-like manner, clustering and topic modeling are different. Clustering partitions similar data into groups while topic modeling finds the hidden topic among common data. Also, clustering does not consider documents to be a member of different clusters. Meaning that, one document can only belong to one cluster while topic modeling allows multiple documents to be in the topic. In other words, one document can consist of multiple hidden topics.

The project uses LDA topic model to generate 9 topic models from 35,056 news data from HuffPost website; the models starts from 2 to 10 topics incrementally. The results are evaluated with both automated coherence metrics, such as intrinsic and extrinsic measures, and human judgment. All evaluations are normalized for better comparison. Tanh function is not quite a good choice to normalize in the project. Instead, scaling all the values into ratio helps comparing as a big picture. To ensure the reliability among raters, Krippendorff's alpha value is used to measure the agreements. 0.8 or more is an acceptable value for the reliability, Krippendorff's value for this project's human ratings is 0.82. Therefore, the human judgement is adequately satisfying. With the automated and human evaluations, Model_3 (5 topics) is the best model for this project. To find which automated measure to be used in the future, one possibly answer is UCI due to the experiment done in this project. UCI measure correlates with human judgement with the Krippendorff's alpha of 0.667 more than other automated metrics; this value is ample to draw the conclusion. For future projects, UCI automated metric can be used to justify the model rather than having humans to judge as it can moderately give similar agreement to humans. This way, human interaction can be reduced.

In conclusion, LDA model used for topic modeling fairly generates a decent model. However, choosing the right number of topics is not an easy task. Topic coherence metrics can help deciding the

optimal number as it works quite similar to human judgement. Though, each automated metrics have differences among them, for example, the algorithms or intrinsic/extrinsic measures, it is better to justify the optimal number of topics altogether with all available metrics. As seen from Figure 13, UCI metric has proved to be the closest metric to human judgement. As the goal of the project of finding ways to reduce human involvement, UCI is conclusively the evaluation metric for topic modeling. As a result, the project can use LDA topic model to reduce reading time and human labor using just UCI metric to evaluate the model as it is most close to the human judgeement.

Future Works

Topic modeling is like an open-ended research. There are many possible ways to work on and there are still many that are uncovered in today's research. As mentioned in the Scope & Limitations and Discussion sections, more dataset will need better tools such as Hadoop or Spark to process the huge amount of data more efficiently. Depending on the dataset given to the model, the model generates different results. Those results may give better insights or useful information for analyzing.

Preprocessing with existing libraries such as Gensim and NLTK is to maintain consistency on the project. Other available tools or services such as NER (Stanford Named Entity Recognizer) to distinguish the proper noun from the data. Some proper nouns can be a regular noun and the used library cannot identify them correctly, for instance, Person name or Mascot Name.

Lastly, finding the categories of top N words can be helpful for the research. The project implements the similarity findings using WordNet to help categorizing. It does not give the pleasant results. For instance, the top 10 words: ['gun', 'shooting', 'police', 'attack', 'killed', 'mass_shooting', 'gunman', 'firearm', 'suspect', 'gun_control'] are categorized as 'Entertainment' which is morally wrong. Furthermore, he hyponyms of each word in the top N words can be misleading to different entity. For instance, 'honeymoon', 'great_year', 'golden_age', 'half_life' have the same hyponym which is 'time_period'. Because natural language is complex, using similarity by hyponyms/hypernyms is not enough to categorize what are those top N words are about. One software called 'Wmatrix' may make things easier as it linguistically compares and analyzes corpus.

Conclusively, there are many things to explore for topic modeling. Human natural language is very complex and complicated. Creating the best model is quite challenging. Therefore, topic modeling still has a big research area to look at.

References

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM,55*(4), 77-84.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. Journal of Memory and Language, 109, Journal of Memory and Language, December 2019, Vol.109.

Gui, J., & Wang, Q. (2017). Topic modeling of news based on spark Mllib. *2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2018*, 224-228.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.).

Misra, R. (2018, May). News category dataset. *ResearchGate*.

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

Pröllochs, N., & Feuerriegel, S. (2018). Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. Information & Management, Information & Management.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 399-408.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. Exploring Topic Coherence over Many Models and Many Topics. (2012, May). *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 952–961.

Zhao, W., Zou, W., & Chen, J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. *BMC Bioinformatics, 15*(Suppl 11), S11.