Anusha Suresh. Topic Modeling for Email Subject Line Analysis. A Master's Paper for the M.S. in IS degree. May, 2020. 30 pages. Advisor: Christopher A. Lee

Email processing is an emerging area in natural language processing and machine learning. Archivists often must make judgements about the relevance and record status of email messages. This study is an attempt to streamline that process by testing subject line and message body analysis using topic modeling. Specifically, using the Enron Corpus and Latent Dirichlet Allocation, this study investigates the extent to which email subject lines can be used to predict the content of email messages to support efficient archival processing.

Headings:

Email Classification

Topic Modeling

TOPIC MODELING FOR EMAIL SUBJECT LINE ANALYSIS

by

Anusha Suresh

A Master's paper proposal submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2020

Approved by

_____

Christopher A. Lee

Table of Contents

Introduction

Collections of email are growing while the number of staff in archives is likely to remain constant or decrease over time. This situation impacts how quickly archivists can review and process collections for public consumption.

Analysis of the full content of email messages and their attachments can be expensive in terms of both human effort and computational resources. This study explores the viability of making email curation decisions based solely on the content of subject lines. While some subject lines clearly reflect the content of the associated message body content, there are numerous factors that complicate the relationships between subjects and message bodies, including multiple topics discussed in the same message or thread and topic drift within a thread over time. More generally, email behaviors are inconsistent and often difficult to predict. When creating subject lines, authors of email are often focused their immediate context rather than how others might access or understand the subject lines in the future. Individual messages within email discussion threads also often deviate from a single theme, with questions like “Are you free for lunch today?” in between text about what was discussed in a meeting earlier. The question about lunch can act as noise during the appraisal process. This study uses topic modeling, specifically Latent Dirichlet Allocation, to look at topic distributions across the Enron email corpus. The hope is that by configuring the algorithm to produce the most meaningful topics, one can map each subject line and each message to a topic and decide

how related the two email fields are. I investigate whether the subject line topic matches the message body topic. The Enron corpus serves as good study material because of how much it has been researched, the fact that it has been thoroughly redacted, and that it has the emails for multiple employees of the company, making for a rich document collection, and high chance of identifying meaningful topics.

The study uses a combination of python packages and modeling tools over the prepared data. All of the packages used, are open-source software readily available for download. I expected to see a number of false negatives and false positives because of the characteristics of the data and the behavior of LDA. Typically, this it is used for regular to larger text documents; emails vary so greatly in length that another issue is the model could crash; there are not enough words or patterns that can be pulled from the data for meaningful analysis. This study provides exploratory results on whether one can assess content similarity between subjects and messages.

Literature Review

Though research in using machine learning for curation of email has been limited, there are several areas of existing literature upon which my study builds: Topic Modeling, Latent Dirichlet Allocation (LDA), Applications of LDA, Short Text Topic Modeling, Mutual Information/Word Association, Comparative Analysis through LDA, and existing research in Email Processing.

Topic Modeling

This section introduces the idea of topic modeling and summarizes existing approaches. In "Reading the Tea Leaves," Chang et al (2009) describe the fundamentals of topic models and how humans interpret them. They introduce three common methods of topic modeling: probabilistic latent semantic indexing (pLSI), Latent Dirichlet allocation (LDA), and correlated topic model (CTM). In pLSI, the number of topics is proportional to the number of documents in a collection. LDA, which will be used in this study, divides the data into user-defined distributions. The number of topics, the number of words per topic, and the number of iterations over the data, are all a non-zero number that the researcher will choose. Usually this will occur iteratively after assessing the performance of the value. LDA does not generate named topics but instead generates lists of terms, and it is up to the researcher to infer what the terms have in

common. Lastly, CTM looks at the correlation between topics using a logistic distribution.

Over the years, researchers have attempted to improve the performance of the most popular topic modeling algorithms. One example is conceptualized latent Dirichlet allocation . “The basic assumption of most existing topic models is that each document is modeled as a probability distribution over topics, and each topic is directly a probability distribution over words […] In this novel assumption, each document is considered as a probability distribution over topics, each topic is a probability distribution over concepts, and each concept is a probability distribution over words” (Tang et al, 2018, p. 3456). Basically, it is looking at the topics to find related topics, then looking at the words to find related words, contextualizing the topics and requiring one less step for the human during the sense-making stage of topic analysis.

Researchers have found that “understanding the basic principles of these algorithms is essential in order to properly configure and use them. Hence, there is a need to understand how the results of topic models are created and to adapt the models to given data and tasks” (El-Assady et al, 2018). El-Assady et al. created a visualization of topics to aid the human sense-making process. Other researchers found several challenges associated with topic modeling, including predicting user behavior, visualization techniques, and image categorization (Jelodar et al 2019).

Topic modeling is an algorithmic approach to categorize document collections. It looks at the distributions of words in a collection and generates topics to which a human will need to add meaning.

Latent Dirichlet Allocation (LDA)

As mentioned previously, I have used LDA in this study. Being one of the most widely used and known methods of topic modeling, the documentation was helpful in the development of this study.

More specifically, LDA was developed as a framework in 2003 (Blei et al, 2003). Their goal was "to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks" (Blei et al, 2003, p. 993). More than 15 years later, this algorithm is applied in a multitude of domains, and, other researchers have built various tools and methods from this algorithm. For example, researchers came up with a modified version of LDA which "gives us the means to check the behavior of [Approximate Distributed] AD-LDA during execution, obtaining some assurance that our distributed implementation is not causing serious errors" (Ilher & Newman, 2012, p. 8). They were able to improve their performance metrics by reducing the sources of error in processing. Additionally, their modifications allowed them to improve reporting during each stage of processing giving them the opportunity to track errors and metrics throughout analysis. Running this algorithm over a large data set on many personal computers can take a long time depending on how many times the algorithm parses the data. Providing more reporting can assist with iterative judgements.

In "Unsupervised Latent Dirichlet Allocation for supervised question classification," the Momtazi states, "Our method first uses unsupervised topic modeling to extract topics from a large amount of unlabeled data. The learned topics are then used in the training phase to find their association with the available category labels in the

training data. The category mixture of topics is finally used to predict the label of unseen data" (2019, p. 380). Previously, we discussed the two-fold distribution method that is utilized in LDA, the distribution of words topics then the topics in the documents, this specific model leverages an unsupervised learning approach, where a "gold-standard"/set of tagged data is not needed. The data gathered through this approach can then be used as a training set or validation set in order to gauge performance. While it is outside the scope of my study, it would be interesting for future research to apply Momtazi's approach to classification of email.

Applications of LDA

Processing of email collections is a relatively recent development, with many prominent cultural institutions exploring the best approaches. We can learn important lessons from the application of topic modeling to other types of materials.

Blog posts contain a wealth of information, and researchers have applied topic modeling to blog content to predict voting preferences. They "used probabilistic topic models to cluster the ground truth documents for each candidate into different underlying latent themes. The same topic models were then applied on the blog collection and the 'orientation' of each of the blogs with different themes of the election candidate speeches was performed using KL [Kullback-Leibler] divergence of the topic distribution over the overlapping vocabularies" (Das et al, 2009, p. 85). By taking this two-fold approach, the researchers found that issues such as tone and sarcasm are not easily detected during natural language processing.

Other researchers spent time analyzing research abstracts to extra topics using LDA. Once they extracted their topics, they performed regression analysis to identify “hot” and “cold” topics (those that matched with human topic judgements and those that did not). (Fang et al, 2018). This y paves the way potentially to improving search results, improving search recommendations, and speeding up the search process.

Lu et al. (2016) found that multiple channel latent Dirichlet allocation (the use of multiple variables for data generation) was helpful in determining medical decisions based on partial medical records. This variation of LDA allowed the researchers to identify links between the variables that they considered to aid in clinical decisions. They had a combination of four variables that went into the algorithms learning and then found correlations between diagnoses and medications. This analysis outdid several other methods including the common k-nearest neighbor (a simple similarity measure for classification tasks).

Researchers have used LDA for email thread identification. The goal for these researchers was a two-fold clustering approach; emails were first grouped together generally and then grouped again to identify specific email threads. (Sharaff & Nagwani, 2016). The approach itself was creative in identifying relationships between emails, however the issue here is understanding the significance of the email threads – if the thread remained in a single email cluster or if the thread stretched across topics or clusters.

Short Text Topic Modeling

One of the limitations of common topic modeling approaches is that they require text within documents to be relatively long. Short text topic modeling is known to be effective for texts that are less than 50 words, including tweets, headlines, and stack overflow questions (Amrouche, 2019). Some researchers have found that associating each text with a small number of topics in addition to learned sematic word relations yielded the best results (Li et al, 2017-2018). Others have tried to solve the sparseness problem (too little data to find patterns or develop meaningful topics) by taking a global view, using hidden topics in larger datasets and applying the learned knowledge to the short texts to fill in topical gaps (Phan et al, 2011). Another approach to address the sparsity problem is fuzzy topic modeling, which uses a bag of words frequency approach to cluster words. The clustering is further improved by looking at word co-occurrence to create fuzzy clusters (Rashid et al, 2019). Another approach is word vector analysis using external data, which accounts for how a word fits into and relates to other words in a larger collection (Yu & Qiu, 2019). One can also use document clustering instead of word clustering, based on the premise that "words that appear in long documents can enrich short text context and improve the clustering performance for short texts" (Yan et al, 2017).

Mutual Information/Word Association

For this study, I considered both mutual information and word association. Mutual information is a determination of whether two words relate to each other based on

their probabilities of occurring together. Word association determines the probability of one word occurring if another word has already occurred in the text.

Some researchers have found that looking at both statistical and semantic information have better results in relationship mining (Yang et al, 2011). Others find that mutual information is only useful in some contexts, and serves as a good measure in only certain situations; "Bayesian classifiers are more suitable to the cases when cost terms are exactly known for trade-off of error types and reject types. Mutual-information classifiers are capable of objectively balancing error types and reject types automatically without employing cost terms" (Hu, 2011, p. 16-17). Other researchers have found that mutual information along with other concepts (including redundancy) were better at understanding different types of independence and dependencies within their data (Hong & Kim, 2011). Others have considered threshold intervals, which can improve term extraction methods (Bin & Shichao, 2011). One study used document clustering which they found "improve the performance of cross-domain learning for text categorization" (Zhuang et al, 2011, p. 113).

Comparative Analysis through LDA

In one study, researchers wanted to know about the matching between a web service description and the web service request. They went through a process of "semantic service categorization and enhancement" to get their data to a point where LDA would be successful in assessing semantic similarity (Mhatre & Jadhav, 2017, p. 951). Another interesting method was when researchers wanted to understand two co-

occurring documents, which in this case were comments to a news article.
Their method involves using short texts and longer texts to learn from each other, and use their proposed algorithm to predict topics and features (Yang et al, 2018).

Two other studies have used LDA for opinion mining in competing product reviews (Wang et al, 2007) and for facial recognition (Kim et al, 2005). These studies have innovative goals and methods in their use of LDA to compare data within a collection.

Existing Research on Email Processing

One of the most active areas of email research has been spam detection. Most traditional spam filtering methods work with ~90% accuracy. However, by developing a two-step filtration process, researchers have been able to achieve higher accuracy. This approach starts with traditional methods to catch most common spam, and then a more personalized step by looking at user background information (Youn, 2014). Another method for spam detection uses session header information to consider trends in common junk mail analysis (Wang & Chen, 2007).

Other research has focused on making sense of email collections. One study, using information from Indian internet users, looked at email to understand feelings and conversations around net neutrality. Here, LDA was used as part of a sentiment analysis assessment of the collection (Jayathilaka et al, 2016). In 2015, researchers used LDA to categorize and make sense of email. Their hope is that users can take email processing into their own hands and decide how many topics and clusters they

need for their purposes (Hong & Moh, 2015). In a similar case, the researcher's goal was subject and folder classification through an n-gram analysis model (Alsmadi & Alhami, 2015). With the growth of email collections, it is becoming increasingly important to benchmark this research.

Research Questions and Hypotheses

The purpose of the study is to understand if email subject lines are good indicators of the purpose or intent of the email message body. The primary question is how reliably one can determine the purpose and importance of an email message based solely on the content of the subject line. This could aid archivists during the appraisal process by speeding up the time spent reading and tagging email. Additionally, I was also interested in comparing both the scope and number of topics of subject lines and email message bodies.

Methodology

Data Collection and Processing

The dataset being studied is comprised of email messages found within the ENRON Corpus. The whole corpus is upwards of 700,000 messages from approximately 150 email accounts. The ENRON corpus is readily available online, and the messages are in PST, a format used by Microsoft Outlook. The corpus has already been redacted, so the risk of exposing any private data is low. I used NLTK, String, re, and libratom to conduct the following pre-processing of the text: stripping header information, appending only the subject line to the body text, removing the Enron disclaimers, splitting and casing the text, removing stop words, removing frequently occurring words, removing punctuation, removing numbers, tokenizing the text, removing days of the week, months of the year, small words, email addresses and websites. These decisions came after multiple rounds of testing and the realization that many of these were the most frequently occurring terms in the dataset. A common source of redundancies was when an entire email thread was stored within the body of a single message (so header material in a body of a reply, was counted as part of the body of that message). Subject lines were usually 10 words or fewer. Each message in the Enron corpus has a unique identifier. The identifier was important during the testing and evaluation phases of this study, in which I assessed the relevance of approximately 100 messages.

Data Analysis Methods

For topic modeling, I used gensim's LDA algorithm (gensim being a widely popular data science package for python). As mentioned, the text was processed to include both the subject line and body. This was to ensure that the same topics were generated, meaning there were not different topics created for the subjects and messages. The combination of the subjects and messages had a uniform set of topics which later on allowed me to go back and assign topics to each individual subject and message. LDA works best with large amounts of text.

I ran the LDA algorithm over the data several times to find the optimal number of topics and optimal number of words per topic. There is no best way to determine the number of models. Options like hierarchical Dirichlet Process, and covariance modeling exist to aid in this process, but were limited here due to processing constraints. Thus, the number was decided after numerous trials of testing various numbers for k topics and deciding how many of them made sense. I decided if the number of words in the topics and if those words were made sense together. I tested various numbers of topics between 10 and 60 with the number of words per topic between 10 and 50.The main iteration discussed here will be 35 topics with 20 words in each topic. At this point, subject lines and message bodies were considered separately for the first time in this project. The program was developed such that each email was parsed, and topic distributions were assigned to each subject line and each message. The topic distributions show the amount to which each topic appeared for the particular subject or message. This is where the unique identifier was essential; each unique identifier came with a topic distribution for subjects and for message bodies. There was one issue with the unique identifiers where

they were unique to specific pst files. For example, Chris Germany has 3 files to his name, so some identifiers repeat across his files. This usually was not a problem as many of the people had only one file to their name. For analysis, random.choice() chose a pst file at random. Then random.sample() chose 100 messages without replacement. This results in 100 messages chosen from Sara Shackleton's collection. I then determined how many of those subjects and messages were in true agreement after analyzing the most significant topic in each distribution (if the same topic appeared with higher significance for both the subject and message).

Results and Discussion

I manually analyzed each of the messages from the 100-message sample to draw the following results. I found that 13% of the data was in true agreement between both the subject line and message body where the subjects and messages had the same significant topic. I simply looked to see that the one topic drew maximum significance from both sets of topic distributions to make this judgement. Another 3% were not true agreements; messages had a single significant topic that matched one of two significant topics for the subjects. There were two subjects because there was not a significant enough difference to select a singular topic. 44% of results were in complete disagreement; the most significant topic that was assigned to the subject was not the most significant topic that was assigned to the message. Additionally, there were results that ended up in a gray area. For example, 6 of the 100 messages had all to all relationships, there was even distribution across all the topics for both subjects and messages There was an additional portion of the results that were inconclusive, of the whole 34% of the data showed that while a subject could have one significant topic, the message might have an even all-round distribution, about 21% of total results. The same issue was found with even topic distributions for the subject to a significant topic in the message(13%).All of this information sheds light on how the emails behave, the limitations of this study, future improvements, and how this information could be used in the appraisal process.

Study Limitations

The results of this study suggest several areas for improvement. The first is that the text was preprocessed in multiple ways expect for stemming. Stemming is the process of reducing the total number of words in the corpus by reducing the words to just their stems. This also reduces the number of redundancies in the dictionary that you created, leading to an improved method of determining a richer set of topics.

Another issue, within the data itself, is how each message was formatted, especially when it came to email threads. Each thread, whether it be a reply or forward had more "header" information within it which was counted as part of the message body. So, while the specific content is not an official header/metadata for the message, it is considered content due to the timing of the email – the most recent email in the thread has the right metadata in the right place. The extraneous "header" information was removed during the second round of pre-processing when frequently occurring words were also removed. However, this step still left behind other indiscernible content.

Due to the way outlook structures the messages, whitespace was never removed accurately. There were many instances where words joined together, and many where they were split in the middle. This is a problem both with the email structure, and how return events (pushing the enter button) are encoded and decoded within the text. Anomalous phrases were relatively rare, but they do seem to add and create noise in the dataset (another reason why the number of topics was not optimal).

Essentially, the issues came down to problems with preprocessing and the dataset itself. The final issue was that the size of the dataset has a correlation with the noise

within it. If the dataset is very large, but also very noisy, then it will produce more faulty results, also indicated by this study. .

Email Behaviors

After observing some messages in particular, there is more to say about email behaviors and people's behaviors towards email. In an analysis of several of the absolute matches, there was a word to word match between what was in the subject line and what was in the content. Occasionally, this match was because of entire threads being contained within message bodies, in which can the subject line matched that same subject line appearing within the body. This is the same reason 3% of the pairs had relative agreements where the most significant topic in the content matched one of two marginally(difference of a thousandth) different topics in the subject. One of these relative matches was because terms within the subject and message were related or could be considered synonymous (specifically the relationship between "equity" in the subject and "number of shares" in the content). It is important to note that many of these numbers were also followed up with manual analysis of the messages. When I observed the actual messages, in these relative matches, the matches did not make sense even though the topics did. Thus, while the positives were accurate, in that they matched with the topics and by my evaluation, they were not much of a reflection of whether the subject line was purely indicative of the body.

Based on an analysis of a random subset of nonmatches, 44 messages to be exact, a handful of them tuned out to be matches. I reviewed each of the messages to determine if they are true negatives as indicated by the topic distribution, and I was able to find several things not indicated previously. There would be a word for word match in the

subject and content but the topic distributions would widely differ. Another issue with the non-matches was that there were equal distributions over the topics for subjects and/or messages. In most cases, when the message was not assigned a topic, this was because the body content was actually empty (draft messages or messages that included attachments but no associated message content). Reading through other messages, many of them were less than 30 words, excluding their inline subject information(parts of threads). All 100 of the emails analyzed had subject lines. In most cases, they were one word or a short phrase, which usually match something in the content or synonymous to it. There was one specific case where there was a match, but the subject line, which was one word, was not an exact content match, it was appeared to be a related word. It appears that communication at Enron was often quick, in short bursts, with short subject lines and message bodies. Email collections with longer subject lines and (especially) longer message bodies may not contain such a large number of direct words matches between subject and body.

Future Improvements

This dataset would have benefitted from stemming during the pre-processing steps. It would have greatly reduced the noise in the data and made sure that the set to create the topics was as meaningful as possible. Another step that could be used to produce better results would be to reconsider how the emails are structured. In the analysis, it is clear that because multiple sets of header information that are counted as text can create noise in the data, removing this information would not only change how the topics are created, but it also could impact whether or not the topics are true matches. The difference being if the matches are created based on content words being exact

matches, or semantic relationships between them. It would also be interesting to look at the relationships between the topics themselves. For now, there are few topics, but many of them are clear as to what they might be about. Understanding if there were actual similarities with the topics is more informative considering that many results had marginally similar topic distributions (take for example message 221692 for which the topic distribution for the subject line was (Topic 8, 0.34295702), (Topic 10, 0.34260932)). This could either mean two topics are related or that the content within the email are so closely related that they do not fit into one category. This information then can be used to make decisions about key players in the corpus and the types of information they discuss (can you determine their position or department or role in an organization by going through their email -aside from their signatures).

This study faced certain limitations in both processing and the data itself. However, it also serves as a gateway for understanding the populations found within email corpora, which I discuss in the next section.

Archival Appraisal Process

This study aimed to see if there is a way to determine if an email is a record in a faster way than having someone manually read and tag emails. By understanding if the subject is predictive of message content, accurately, then the appraisal process can be streamlined. While the results of this study are inconclusive, they are paving the way for more questions. How can emails be classified together? Are all messages that fall under the same topic, actually about the topic? Are the subject lines actually meaningful in record analysis?

Another set of issues relate to personally identifiable information (PII). PII detection and redaction is very important as most email processing occurs for public consumption. Understanding if there might be PII based on the content of the subject line streamline not only record management, but also PII Redaction. There are existing tools that remove PII, but the question is whether or not one can predict whether a message contains PII and needs to be redacted. There are a number of other questions that can be asked, but ultimately it comes down to the goals of the archival institutions and the laws and mandates they abide by to process and publish this information.

Conclusion

This study was conducted to see if emails subject lines were indicative of the message body through the use of Latent Dirichlet Allocation. After an analysis of 100 random messages, the results were deemed inconclusive. There were several positive matches based on containing the exact same words or phrases. There were several mismatches due to empty message bodies (simple emails that only contained attachments or drafts) and many false negatives due to topic distribution, in other words, the model found them to be mismatches, while I found them to be matches. The negatives made the analysis interesting because there were cases where words in the subject lines had no direct relevance to message content. There were other cases where're even though there were exact matches between the subject and message, the other terms, including those that were synonymous, were identified as mismatches by the model. Many of the issues in this study, in the future, could be improved with richer text processing methods to filter out noise. Additionally, it is possible that working with a smaller dataset, for training, might lead to more meaningful topic creation. This study has planted seeds for more questions to be asked during the appraisal process, whether it be during processing or analysis of a corpus. This study has shown that more meaningful results are possible, with several tweaks, and more questions to be asked along the way.

References

Alsmadi, I., & Alhami, I. (2015). Clustering and classification of email contents. Journal of King Saud University - Computer and Information Sciences, 27(1), 46-57. doi:10.1016/j.jksuci.2014.03.014

Amrouche, M. (2019, August 22). Short Text Topic Modeling. Retrieved from https://towardsdatascience.com/short-text-topic-modeling-70e50a57c883.

Bin, Y., & Shichao, C. (2011). Term extraction method based on mutual information with threshold interval. Paper presented at the , 227(4) 186-194. doi:10.1007/978-3-642-23226-8_25

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(4-5), 993-1022.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. Paper presented at the 288-296.

Das, P., Srihari, R., & Mukund, S. (2009). Discovering voter preferences in blogs using mixtures of topic models. Paper presented at the 85-92. doi:10.1145/1568296.1568311

El-Assady, M., Sevastjanova, R., Sperrle, F., Keim, D., & Collins, C. (2018). Progressive learning of topic modeling parameters: A visual analytics framework. IEEE Transactions on Visualization and Computer Graphics, 24(1), 382-391. doi:10.1109/TVCG.2017.2745080

Fang, D., Yang, H., Gao, B., & Li, X. (2018). Discovering research topics from library electronic references using latent dirichlet allocation. *Library Hi Tech, 36*(3), 400-410. doi:10.1108/LHT-06-2017-0132

Hong, C. S., & Kim, B. J. (2011;2009;). Mutual information and redundancy for categorical data. Statistical Papers, 52(1), 17-31. doi:10.1007/s00362-009-0196-x

Hong, H., & Moh, T. (2015). Effective topic modeling for email. Paper presented at the 342-349. doi:10.1109/HPCSim.2015.7237060

Hu, B. (2011). What are the differences between bayesian classifiers and mutual-information classifiers? doi:10.1109/TNNLS.2013.2274799

Ihler, A., & Newman, D. (2012). Understanding errors in approximate distributed latent dirichlet allocation. IEEE Transactions on Knowledge and Data Engineering, 24(5), 952-960. doi:10.1109/TKDE.2011.29

Jayathilaka, K. M. P. N, Weerasinghe, A. R., & Wijesekara, W. M. L. K. N. (2016). Making sense of large volumes of unstructured email responses. Paper presented at the 35-40. doi:10.1109/ICTER.2016.7829896

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211. doi:10.1007/s11042-018-6894-4

Kim, T., Kim, H., Hwang, W., & Kittler, J. (2005). Component-based LDA face description for image retrieval and MPEG-7 standardisation. Image and Vision Computing, 23(7), 631-642. doi:10.1016/j.imavis.2005.02.005

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2017;2018;). Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Transactions on Information Systems (TOIS), 36(2), 1-30. doi:10.1145/3091108

Lu, H., Wei, C., & Hsiao, F. (2016). Modeling healthcare data using multiple-channel latent dirichlet allocation. Journal of Biomedical Informatics, 60, 210-223. doi:10.1016/j.jbi.2016.02.003

Mhatre, P. J., & Jadhav, V. (2017). A SEMANTIC MODEL FOR WEB SEARCH AND DATA MATCHING USING LDA. International Journal of Advanced Research in Computer Science, 8(7) Retrieved from http://libproxy.lib.unc.edu/login?url=https://search.proquest.com/docview/1931100927?accountid=14244

Momtazi, S. (2018). Unsupervised latent dirichlet allocation for supervised question classification. Information Processing and Management, 54(3), 380-393. doi:10.1016/j.ipm.2018.01.001

Phan, X., Nguyen, C., Le, D., Nguyen, L., Horiguchi, S., & Ha, Q. (2011). A hidden topic-based framework toward building applications with short web documents. IEEE Transactions on Knowledge and Data Engineering, 23(7), 961-976. doi:10.1109/TKDE.2010.27

Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. Information Processing and Management, 56(6), 102060. doi:10.1016/j.ipm.2019.102060

Sharaff, A., & Nagwani, N. K. (2016). Email thread identification using latent dirichlet allocation and non-negative matrix factorization based clustering techniques. Journal of Information Science, 42(2), 200-212. doi:10.1177/0165551515587854

Tang, Y., Mao, X., Huang, H., Shi, X., & Wen, G. (2018). Conceptualization topic modeling. Multimedia Tools and Applications, 77(3), 3455-3471. doi:10.1007/s11042-017-5145-4

Wang, C., & Chen, S. (2007). Using header session messages to anti-spamming. Computers & Security, 26(5), 381-390. doi:10.1016/j.cose.2006.12.012

Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent dirichlet allocation. Electronic Commerce Research and Applications, 29, 142-156. doi:10.1016/j.elerap.2018.04.003

Yan, Y., Huang, R., Ma, C., Xu, L., Ding, Z., Wang, R., . . . Liu, B. (2017). Improving document clustering for short texts by long documents via a dirichlet multinomial allocation model. Paper presented at the , 10366 626-641. doi:10.1007/978-3-319-63579-8_47

Yang, G., Mabu, S., Shimada, K., & Hirasawa, K. (2011). A novel evolutionary method to search interesting association rules by keywords. Expert Systems with Applications, 38(10), 13378-13385. doi:10.1016/j.eswa.2011.04.166

Yang, Y., Wang, F., Zhang, J., Xu, J., & Yu, P. S. (2018). A topic model for co-occurring normal documents and short texts. World Wide Web, 21(2), 487-513. doi:10.1007/s11280-017-0467-8

Youn, S. (2014). SPONGY (SPam ONtoloGY): Email classification using two-level dynamic ontology. Scientific World Journal, 2014, 414583-11. doi:10.1155/2014/414583

Yu, J., & Qiu, L. (2019). ULW-DMM: An effective topic modeling method for microblog short text. IEEE Access, 7, 884-893. doi:10.1109/ACCESS.2018.2885987

Zhuang, F., Luo, P., Xiong, H., He, Q., Xiong, Y., & Shi, Z. (2011). Exploiting associations between word clusters and document classes for cross-domain text categorization. Statistical Analysis and Data Mining, 4(1), 100-114. doi:10.1002/sam.10099