Global Scientific JOURNALS

GSJ: Volume 11, Issue 3, March 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

ABSTRACT CORPUS ANALYSIS OF CAPSTONE PROJECTS USING LATENT DIRICHLET ALLOCATION ALGORITHM FOR AEMILIANUM COLLEGE INC.

RHODORA FAYE ALIM - BROSAS
JOSEFINA R. SARMIENTO

AEMILIANUM COLLEGE INC.
Rizal St., Piot, West District, Sorsogon City, Sorsogon, Philippines

Abstract. The study aimed to design and develop an Abstract Corpus Analysis of Capstone Projects Using a Latent Dirichlet Allocation Algorithm for Aemilianum College Inc. to categorize and evaluate the thematic corpus analysis model along Sustainable Development Goals of the United Nations and CHED CMO No. 07, a series of 2010; and, to evaluate the Capstone Project Corpus Analysis System for Aemilianum College Inc. using the industry software quality model – the ISO 25010 evaluation tool.

Based on the findings, the proposed corpus analysis system of the Capstone projects for Aemilianum College Inc. included features like data cleaning, thematic analysis showing the word coherence and automatic labels, and visualization to translate information into a visual context. The Capstone project conducted was within the Sustainable Development Goals and it met the research standard set for a Master in Information Technology (MIT) student. The evaluation made by the experts and users obtained an overall weighted average rating of 4.80 and 4.50 respectively which were both interpreted as far more than what is expected.

Furthermore, as warranted by conclusions, the system is recommended to be utilized by the target beneficiary. Likewise, it is also recommended to utilize an algorithm aside from LDA to generate thematic analysis. Also, there is a need to explore other classifiers and consider other word intrusion methods in evaluating the topic labels generated.

Key Words*: Abstract Corpus Analysis, Aemilianum College Inc., Algorithm, Capstone Project, Latent Dirichlet Allocation, Natural Language Processing, Theses Abstracts, Thesis, Project Study*

INTRODUCTION

Text corpora play an important role in the field of natural language processing. Many studies in natural language processing rely on training data in the form of annotated text corpora. The process of building these corpora is not easy, as there are currently no automated approaches that can accurately make text annotations that require human interpretation. As such, building a good text corpus necessitates the presence of human annotators to manually do the annotations and this requires a lot of time and effort (Tiam-Lee and See, 2015).

One text corpora approach is the manual data collection by experts by hiring a group of experts to annotate the sentiment of a text. Corpora built from this kind of approach show excellent quality in terms of the correctness of the annotations. The annotations also typically contain rich information (Breck, et. al. 2003). However, the cost and time required for this approach usually result in smaller corpora in limited domains (Cardie, et. al. 2015).

An automated text corpora approach is larger since the data collection is usually done by crawling through websites. The annotations are done automatically with computer algorithms that infer the sentiment of the text by observing certain features as shown in the works (Kaji, et. al. 2010). While corpora built from this approach are larger, the quality is not as good, since sentiment annotation is a task that requires human perception and is something that computer algorithms cannot do accurately yet.

The annotation of a corpus about the Oklahoma wildfires aims at the provision of broad-scale information as opposed to safety information mining about individuals. (Corveyt, et. al. 2010)

As cited by Kaufman (2010), Natural Language Processing (NLP) is the field of computer science, artificial intelligence, and linguistics that is concerned with the interactions between computers and human languages - is a growing field of studies and research. This field has a lot of real-world applications such as automatic summarization, named entity recognition, natural language understanding, question answering, sentiment analysis, speech recognition, information retrieval, and more. Studies in these fields require large data sets for training machine learning systems so that they can perform well and give reliable results. These data sets, usually called corpus or corpora in natural language processing, can be easily obtained from online resources.

According to Kong (2015), Natural language processing (NLP) deals with techniques in computer science for processing human language texts and speech. This is a difficult task, as human language is dynamic, flexible, and ambiguous. Example applications of NLP include information retrieval, information extraction, question answering, text data mining,

sentiment analysis, automatic translation, speech recognition, and synthesis. The impact of data abundance extends well beyond business, for example, political science, law, education, and public health are disciplines which becoming increasingly data-intensive toward data-driven discovery and decision-making. Data is not only becoming more available but also more understandable to computers. Computer tools for gleaning knowledge and insights from Big Data are fast gaining ground. At the forefront are the rapidly advancing techniques of artificial intelligence like natural-language processing, pattern recognition, and machine learning.

Some researchers, like Benhardus and Kalita (2013), obtained their data by streaming online data sources like Twitter and Facebook. Streaming provides a larger and more real-time data set. When dealing with streamed data, one of the first things that the researcher has to do before using it is clean it. The cleaning process involves preprocessing the data by applying different operations on it so that they are in a uniform and normalized structure when they are used. Such operations include tokenization, stemming, removal of certain components like URLs or emoticons, etc.

The following process may be considered in the study: Separating the tweets that contained safety information from the huge number of irrelevant tweets; extracting important information, such as person names and locations from highly domain-specific text included in the tweets; and, verifying and delivering this information to the people that need it. (Neubig, et. al. 2021).

Lots of document collections are well organized in a hierarchical structure, and such a structure can help users browse and understand these collections. Meanwhile, there are a large number of plain document collections loosely organized, and it is difficult for users to understand them effectively. In this paper, we study how to automatically integrate latent topics in a plain collection with the topics in a hierarchically structured collection. We propose to use semi-supervised topic modeling to solve the problem in a principled way. (Mao, et. al. 2021). The experiments show that the proposed method can generate both meaningful latent topics and expand high-quality hierarchical topic structures.

Yi and Allan (2008) explore the utility of different types of topic models, both probabilistic and not, for retrieval purposes. They showed that: (1) topic models are effective for document smoothing; (2) more elaborate topic models that capture topic dependencies provide no additional gains; (3) smoothing documents by using their similar documents is as effective as smoothing them by using topic models; (4) topics discovered on the whole corpus are too coarse-grained to be useful for query expansion. Experiments to measure topic models' ability to predict held-out likelihood confirm past results on small corpora, but suggest that simple approaches to topic models are better for large corpora.

Masada, et. al. (2009) paper presents a new Bayesian topical trend analysis. We regard the parameters of topic Dirichlet priors in latent Dirichlet allocation as a function of document timestamps and optimize the parameters by a gradient-based algorithm. Since the method gives similar hyper parameters to the documents having similar timestamps, topic assignment in collapsed Gibbs sampling is affected by timestamp similarities. We compute TFIDF-based document similarities by using a result of collapsed Gibbs sampling and evaluate our proposal by link detection task of Topic Detection and Tracking.

Latent Dirichlet allocation (LDA) is a generative probabilistic model that can be used for text corpora. LDA considers each document in the corpus as a mixture of a set of topics.

Each topic is in turn considered a mixture of words. When LDA is run over a corpus of online reviews about a specific product, the list of topics it returns can be considered as probable features of the product. In the implementation, the LDA model of the "gensim" package is used (Blei, et. al. 2015).

As large-scale text data become available on the Web, textual errors in a corpus are often inevitable (e.g., digitizing historic documents). Due to the calculation of frequencies of words, however, such textual errors can significantly impact the accuracy of statistical models such as the popular Latent Dirichlet Allocation (LDA) model. To address such an issue, Yang and Lee's paper, proposed two novel extensions to LDA (i.e., TE-LDA and TDE-LDA): (1) The TE-LDA model incorporates textual errors into the term generation process, and (2) The TDE-LDA model extends TE-LDA further by taking into account topic dependency to leverage on semantic connections among consecutive words even if parts are typos. Using both real and synthetic data sets with varying degrees of "errors", our TDE-LDA model outperforms (1) the traditional LDA model by 16%-39% (real) and 20%-63% (synthetic); and (2) the state-of-the-art N-Grams model by 11%-27% (real) and 16%- 54% (synthetic). (Yang and Lee, 2013)

With this, the proposed study intends to make a text scanning through a thematic corpus analysis of the Master in Information Technology Capstone Project of Aemilianum College Inc. using the Latent Dirichlet Allocation (LDA) method for the last twenty (20) years. The proposed system will provide a listing of the key topics commonly conducted by the students in MIT Thesis based on the Abstract copies in order to give an idea to the school administrators if said topics are in adherence to the suggested areas mandated by CHED and in accordance to the Sustainable Development Goals (SDG) agenda of the United Nations.

Specific Objectives

Specifically, the study aimed to:

1. Design and develop a corpus analysis system of the Capstone projects for Aemilianum College Inc. with the following features:
    - 1.1 Pre-processing
    - 1.2 Thematic Analysis using Latent Dirichlet Allocation (LDA)
    - 1.3 Visualization
2. Categorize and evaluate the thematic corpus analysis model along:
    - 2.1 Sustainable Development Goals of the United Nations
    - 2.2 CHED CMO No. No. 07, series of 2010
3. To evaluate the Capstone Project Corpus Analysis System for Aemilianum College Inc. using the industry software quality model – the ISO 25010 evaluation tool in terms of:
    - 3.1 Functional suitability
    - 3.2 Performance efficiency
    - 3.3 Usability
    - 3.4 Reliability
    - 3.5 Security
    - 3.6 Maintainability

3.7 Portability

Table 3.5. - Project Development Time Frame

Activities	Month 1				Month 2				Month 3				Month 4				Month 5				Month 6			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
a. Requirements Planning																								
▪ Review related literature & studies																								
▪ Permission & meeting with the Librarian																								
▪ Collect MIT Capstone abstracts																								
b. User Design																								
▪ Design the system structure																								
▪ Develop the proposed screen layout																								
▪ Design the NLP Processes																								
▪ Prepare work plan																								
c. Construction																								
▪ Actual built of the proposed tool																								
d. Cut-Over																								
▪ Technical expert's evaluation																								
▪ Statistical treatment of evaluation ratings																								
▪ Pilot testing																								

The proposed system was built in six (6) months duration as presented in Table 3.5. This observed the agile development processes using the Rapid Application Development (RAD) approach. The requirements planning phase was one in five (5) weeks that includes the review of related literature and studies, especially on Natural Language Processing processes, seeking permission and conduct of meetings with the school Librarian, and collecting the MIT Capstone abstracts as input data in the proposed system.

The user design phase was conducted in eight (8) weeks' time. The activities include the designing of the system structure, development of the proposed screen layout, designing the NLP processes, and preparation of the work plan until the delivery of the proposed project. Next is the construction phase which is the actual coding of the system and this entails 12 months duration. The last phase is the cut-over which is the conduct of the technical expert's evaluation, computation of the evaluation results from the expert's evaluation, and pilot testing. This will be done in the last month of the project development timeline.

Requirements Planning

The students enrolled in the MIT program taking up the capstone project course intend to develop among its students the knowledge of Information Technology (IT) concepts, techniques and principles, and skills in using IT to provide solutions to problems of the institutions in particular. This Capstone project served as the culminating activity among Master in Information Technology (MIT) students that required an output useful in the development of IT solutions.

The Aemilianum College Inc. had a clear agenda on the given topics for the Capstone project that were anchored on the institution's philosophy and framework and must be aligned with the Sustainable Development Goal of the United Nations and thematic areas of the Commission on Higher Education (CHED) for IT education to answer the needs of the industry. The project may be but not be limited to application development that focuses on software engineering processes or application design that focuses on the effective testing procedure or a study on application development processes. The concerned graduate student was the one responsible for the development of the Capstone project.

With this current set-up, the proposed study is intended to categorize the topics being proposed and/or conducted by the graduate students taking the MIT program. The first category to consider was the Sustainable Development Goals (SDG). The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and into the future. At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership as follows: (1) No Poverty, (2) Zero Hunger, (3) Good Health and Well-Being, (4) Quality Education, (5) Gender Equality, (6) Clean Water and Sanitation, (7) Affordable and Clean Energy, (8) Decent Work and Economic Growth, (9) Industry, Innovation, and Infrastructure, (10) Reduced Inequalities, (11) Sustainable Cities and Communities, (12) Responsible Consumption and Production, (13) Climate Action, (14) Life Below Water, (15) Life on Land, (16) Peace, Justice and Strong Institutions, and. (17) Partnership for Goals.

The SDG recognizes that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth – all while tackling climate change and working to preserve our oceans and forests (United Nations, n.d.). As to the CHED thematic areas, the study may be categorized into (1) system integration, (2) application development, and (3) software engineering.

In determining the requirements of the proposed system, figure 4.1 depicted the Context Diagram Data Flow. It has two (2) external entities – the admin and public users. The admin was in charge of managing the analysis of the corpus and other settings of the system like adding stop words and logged files. The admin is capable of uploading the Capstone abstract to the proposed system entitled Abstract Corpus Analysis of Capstone Project Using Latent Dirichlet Allocation algorithm for Aemilianum College, Inc. Dashboard data was made available among the public users.

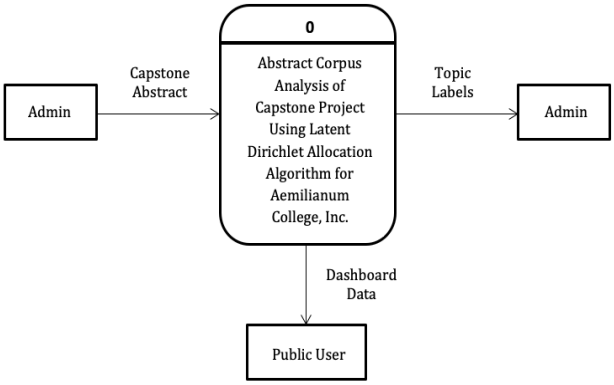

Figure 4.1. - Data Flow Diagram – Context

Detailed requirement specifications on the processes of the proposed system was shown in Figure 4.2. The system allowed the uploading of the Capstone Abstract to undergo pre-processing. Pre-processing includes cleaning such as removing unnecessary words normally comprising of pronouns, conjunctions, or prepositions generally in English and Filipino. These also included irrelevant data like words that contain punctuation and special characters ("#", "/", ":", others), and responses with less than 2 words. The basis of which was the listing of stop words being provided. After pre-processing, the cleaned data were available for download.

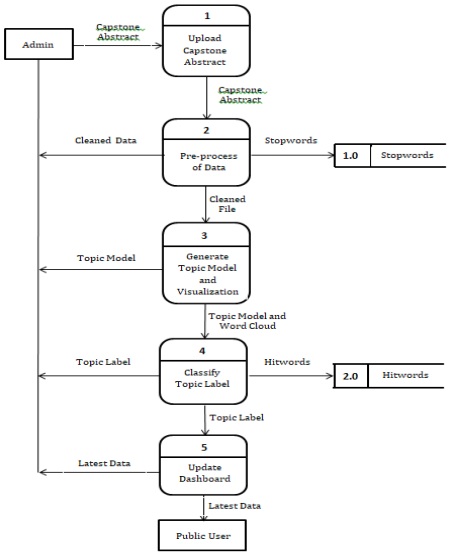

Figure 4.2. - Data Flow Diagram – Level 0

Afterward, the cleaned file could be processed to generate the topic models and visualization of the cleaned corpus uploaded. The topic labels could automatically be generated based on the pre-defined lists of his words. The dashboard could be updated with the latest data available.

To describe further the requirement planning of the proposed system, Use-case diagrams were also drawn to describe the high-level functions and scope of a system. These diagrams also identified the interactions between the system and its actors. The use cases and actors in use-case diagrams described what the system does and how the actors used it, but not how the system operates internally.

Figure 4.3 was the Use case diagram of the proposed system. The actors here were the admin and public users. The figure depicted that the admin is capable of uploading the Capstone abstracts which shall be subjected to pre-processing. Afterward, the generation of topic labels and visualization could be made. The topic labels could be generated automatically based on the two (2) categories given. Further, the admin is capable of customizing the settings such as adding the hit words, adding labels, and adding users. For monitoring purposes, the admin could also view the user log. While the other actor which is the public user could view the dashboard with some important details of the institution.

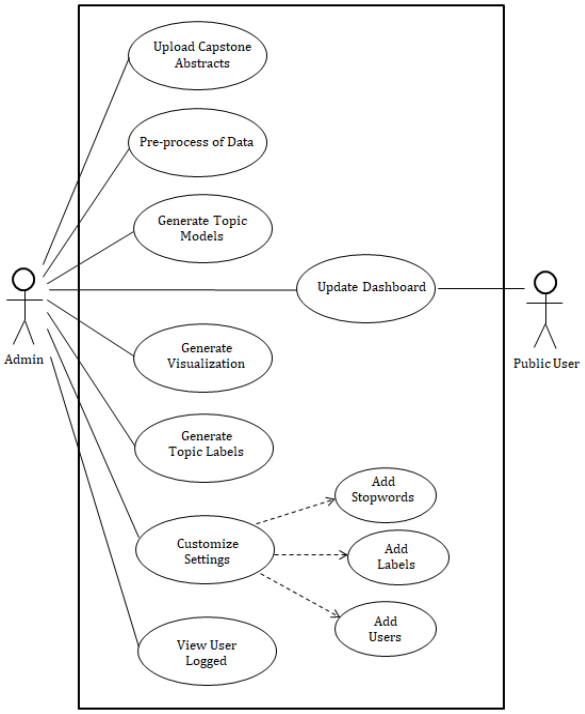

Figure 4.3. - Use Case

Another Unified Modelling Language (UML) was designed as part of the requirements planning. A package diagram was made to depict import and access dependencies between packages, classes, components, and other named elements within the proposed system. Figure 4.4 reflected four (4) components namely pre-processing, topic model, topic label, and custom settings. The pre-processing component included the elements like uploading files, cleaning data, and using/adding stop words. This component was connected with a Topic model which was another component with elements such as uploading the cleaned file, generating the topic model, and visualization. The topic model had access to the component of the Topic Label that was capable of classifying the models according to SDG or CHED thematic areas. In addition, a custom setting component was also made with elements like adding stop words, adding labels, and adding users. Each dependency was rendered as a connecting line with an arrow representing the type of relationship between the two or more elements.

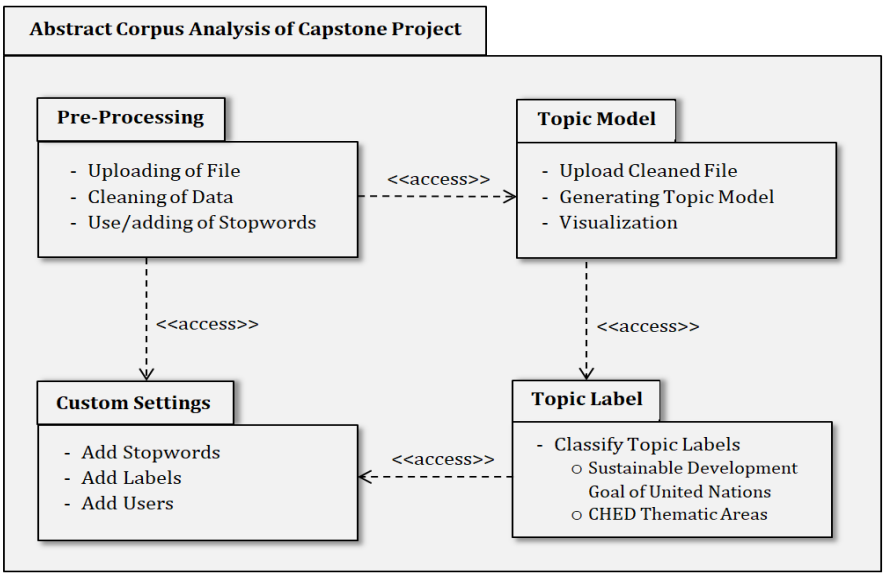

Figure 4.4. - Package Diagram

Figure 4.5 was the deployment diagram of the proposed system. This provided visualization of the hardware processors such as the client machines, web server, application server, and database server as an option. The devices to be considered in the proposed deployment set-up were a computer, web service, and web browser. The nodes included the web-based system, configuration files, and application installer. An artifact of the MySQL server was also included as an optional design in the proposed design.

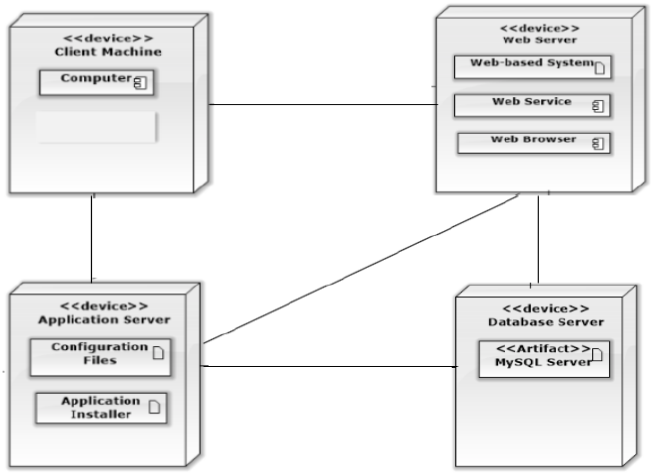

Figure 4.5. - Deployment Diagram

User Design (UD)

In this section, the User Design showed in detail the business activities associated with the proposed system. This included the proposed screen layouts for the most important automated functions and the appropriate construction approach for the system.

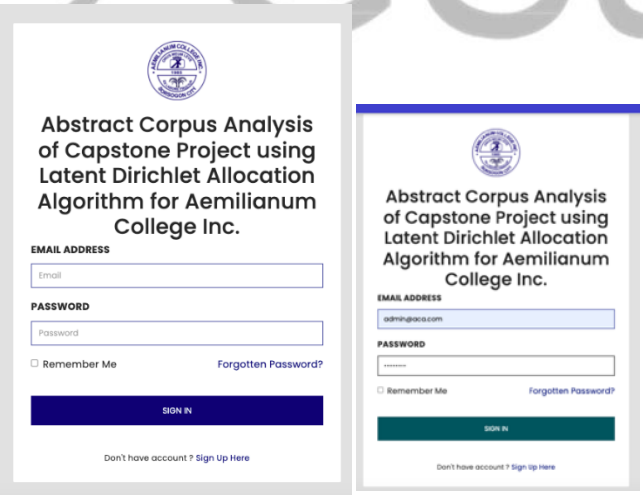

Figure 4.6. - Log-In Page

Figure 4.6 showed the Log-In Page of the proposed system. A login page was a web page or an entry page to a website that requires user identification and authentication, regularly performed by entering a username and password combination. These Logins shall provide access to the analysis and management of the page.

Summary of Findings

The following findings were obtained from the study:

1. Design and develop a corpus analysis system for the Capstone projects for Aemilianum College Inc.
    a. The Capstone abstract corpus undergoes pre-processing by removing the number of words, symbols, numbers, and stop words on each given dataset. Out of the 60 Capstone types of research from years 2007 to 2021, School Year (SY) 2021 has the highest frequency of study conducted although it was shown that SY 2018 obtained a large frequency of words, numbers, and stop words.
    b. In generating the thematic analysis using Latent Dirichlet Allocation (LDA), the number of topics and words were both set to five (5), the optimization interval is 18 and the iteration is 100.
    c. Word cloud is the visualization made on the given Capstone abstract datasets.
2. As to the 17 Sustainable Development Goals, the majority of the Capstone projects conducted for the years 2007 to 2021 focused more on Industry, Innovation & Infrastructure goals. Other studies focused on quality education, and Peace, Justice, and Strong Institutions. Further, it is good to note that almost all of the Capstone projects met the thematic areas set by CHED System Integration (SI), Application Development (AD), and Software Engineering (SE).
3. Using the industry software quality model – the ISO 25010 evaluation tool
    a. The experts evaluated the proposed system as far more than what is expected with the following ratings: reliability (4.85), maintainability (4.84), and portability (4.80) far more than what is expected 4.80.
    b. The User's Evaluation also obtained far more than what is expected with the weighted scores of Functional suitability (4.75), reliability (4.60), usability (4.51), performance efficiency/speed (4.00), maintainability (4.80), portability (4.20), and security (4.60), compatibility (4.50).

Conclusions

Based on the findings of this study the following conclusions are formulated:

1. The proposed corpus analysis system of the Capstone projects for Aemilianum College Inc. includes features like data cleaning, thematic analysis showing the word coherence and automatic labels, and visualization to translate information into a visual context.
2. The Capstone researches were conducted within the Sustainable Development Goals and met the research standard sets for Master in Information Technology (MIT) students.
3. The evaluation made by the experts and users obtained an overall weighted average rating of 4.80 and 4.50 respectively which are both interpreted as far more than what is expected.

Recommendations

Based on the conclusions, the following recommendation is hereby offered:

1. To utilize another algorithm aside from LDA in generating a thematic analysis.
2. To explore another classifier tool aside from Gensim.
3. To consider other word intrusion methods in evaluating the topic labels generated.

References

1) Cruz, N. Oco & R. E. Roxas; A classifier module for analyzing community responses on disaster preparedness: *2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM),* doi:10.1109/hnicem.2017.8269532; Published 2017.
2) Arnelle C. Balane and Kurt Junshean P. Espinosa; NLPrep: A Streamable Preprocessing and Warehousing Platform for Natural Language Processing: Proceedings of the 11th National Natural Language Processing Research Symposium; Published April 24-25, 2015.
3) Breck, E., et. al.; NRRC summer workshop on multi-perspective question answering final report: Technical Report; Published 2003.
4) Cardie, C., et al.; annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2 – 3); Published 2015.
5) Carmen Bedia and Joaquim Jaumot; Data Analysis for Omic Sciences, Methods and Applications: Comprehensive Analytical Chemistry; Published 2018.
6) Chuchi Montenegro, et al; Using Latent Dirichlet Allocation for Topic Modeling and Document Clustering of Dumaguete City Twitter Dataset. ICCDE 2018: Proceedings of the 2018 International Conference on Computing and Data Engineering; Published May 2018; from https://doi.org/10.1145/3219788.3219799.
7) Y. Sy; Topic Modelling Disaster-Related Responses Using The Latent Dirichlet Allocation (LDA) Approach: University of the Cordilleras; Published 2018.
8) M. Blei; Probabilistic topic models: Commun ACM 55; Published 2012.
9) S. McNamara; Computational methods to extract meaning from text and advance theories of human cognition: Top. Cogn. Sci. 3; 3–17; Published 2011.
10) Cambria and B. White; Jumping NLP curves: a review of natural language processing research: IEEE Computat Intell Mag; Published 2014.
11) Esposito, Fabrizio and Cutugno, Anna Corazza Francesco; Topic Modelling with Word Embeddings. Proceedings of the Third Italian Conference on Computational Linguistics: Napoli Accademia University Press; Published 5-6 December 2016.
12) Heather Froehlich; Corpus analysis with AntConc. Programming Historian. ISSN 2397-2068; Published 2015.
13) J. D. M. Valencia et al.; Understanding Anonymous Social Media Posts using Topic Modeling: 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM) doi: 10.1109/HNICEM48295.2019.9072791; Published 2019.
14) J. E. Guiao et al.; Discovering topics from qualitative responses of a disaster preparedness e-participation system: TENCON 2017 - 2017 IEEE Region 10 Conference, doi: 10.1109/TENCON.2017.8228287; Published 2017.
15) J. Grimmer, B. M. Stewart; Text as data: The promise and pitfalls of automatic content analysis methods for political texts; Published 2013.
16) J. M. Imperial, A. De La Cruz, E. Malaay, R. E. Roxas; Cross-Textual Analysis of COVID-19 Tweets: On Themes and Trends Over Time. In: Yang XS., Sherratt S., Dey N., Joshi A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology: Lecture Notes in Networks and Systems, Vol 236, Springer, Singapore, https://doi.org/10.1007/978-981-16-2380-6_71; Published 2022.
17) J. R. Ancheta, C. Sy, L. Maceda, N. Oco and R. Roxas; Computer-assisted thematic analysis of Typhoon Fung-Wong tweets: TENCON 2017 - 2017 IEEE Region 10 Conference, doi: 10.1109/TENCON.2017.8227955; Published 2017.

18) J. R. Ancheta, K. D. Gorro and M. A. D. Uy, "#Walangpasok on Twitter: Natural language processing as a method for analyzing tweets on class suspensions in the Philippines," 2020 12th International Conference on Knowledge and Smart Technology (KST), 2020, pp. 103-108, doi: 10.1109/KST48564.2020.9059411.
19) K. Gorro et al.; Qualitative data analysis of disaster risk reduction suggestions assisted by topic modeling and word2vec: 2017 International Conference on Asian Language Processing (IALP) doi: 10.1109/IALP.2017.8300601; Published 2017.
20) Kaji, N. and Kitsuregawa, M.; Automatic construction of polarity tagged corpus from HTML documents. Proceedings of the COLING / ACL on Main Conference Poster Sessions; Published 2006; Pak, A. and Paroubek; Twitter as a corpus for sentiment analysis and opinion mining: Proceedings of Seventh International Conference on Language Resources and Evaluation; Published 2010.
21) Ken D. Gorro, et al.; Exploring Natural Language Processing Techniques in Social Media Analysis during a Pandemic: Understanding a corpus of Facebook posts using Word2vec and LDA. ICIT 2020: 2020 The 8th International Conference on Information Technology: IoT and Smart City; Published December 2020; from: https://doi.org/10.1145/3446999.3447012.
22) Kong, T. E.; Natural Language Processing (NLP) based on the SSTC (Structured String-Tree Correspondences): Proceedings of the 11th National Natural Language Processing Research Symposium; April 24-25, 2015.
23) K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner; Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches: PLOS ONE 6; Published 2011.
24) Lady Angelica Buen Guerzo et al.; Topic Modelling and Clustering of Disaster-Related Tweets using Bilingual Latent Dirichlet Allocation and Incremental Clustering Algorithm with Support Vector Machines for Need Assessment: 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM); Published 2021.
25) L. D. Austero, C. Y. Sy and M. J. P. Canon; Discovering Themes from Online News Articles on the 2018 Mt. Mayon Eruption: 2018 International Symposium on Computer, Consumer and Control (IS3C) doi: 10.1109/IS3C.2018.00068; Published 2018.
26) L. Dietz, *Topic Model Evaluation: How Much Does it Help?* Universitat Mannheim WebSci; Published 2016.
27) Lany L. Maceda, Jennifer L. Llovido, and Thelma D. Palaoag. Corpus Analysis of Earthquake Related Tweets through Topic Modelling. International Journal of Machine Learning and Computing, Vol. 7, No. 6, December 2017.
28) Ligutom; Proceedings of the COLING / ACL on Main Conference Poster Sessions; Published 2016.
29) Masada, T. et al.; Dynamic Hyperparameter Optimization for Bayesian Topical Trend Analysis: Proceedings of the 18th ACM conference on Information and knowledge management; Published November 02 - 06, 2009.
30) Mohd Shamrie Sainin, Asni Tahir, Suraya Alias; Corpus Analysis: A Case Study on Kadazandusun. Newspaper Archive. Institute of Electrical and Electronics Engineers. Published January 2020.
31) Petra Storjohann; Colligational patterns in a corpus and their lexicographic documentation. Liverpool: University of Liverpool; Published 21 June 2016.
32) Tang Enya Kong; Natural Language Processing (NLP) based on the SSTC (Structured String-Tree Correspondences): Proceedings of the 11th National Natural Language Processing Research Symposium; Published April 24-25, 2015.
33) Thomas James Tiam-Lee and Solomon See; Development of a Sentiment Corpus from a Gamified Approach: Proceedings of the 11th National Natural Language Processing Research Symposium; Published April 24-25, 2015.
34) Tiam-Lee, T. J. and See, S.; Development of a Sentiment Corpus from a Gamified Approach. Proceedings of the 11th National Natural Language Processing Research Symposium; Published April 24-25, 2015.
35) V. A. Rohani, S. Shayaa and G. Babanejaddehaki; Topic modeling for social media content: A practical approach: 3rd International Conference on Computer and Information Sciences (ICCOINS); Published 2016.

36) William, et al.; Twitter in mass emergency: What NLP can contribute. In NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media; Association for Computational Linguistics; Published 2010.
37) Yang, T. and Lee, D.; On handling textual errors in latent document modeling: Proceedings of the 22nd ACM international conference on Information & Knowledge Management; Published 2013.
38) Yi, X. and Allan, J.; Evaluating topic models for information retrieval: Proceedings of the 17th ACM Conference on Information and knowledge management; Published 2008.
39) Z. Ghahramani; Probabilistic machine learning and artificial intelligence: Nature 521; Published 2015.