Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 4, 2023
A Regional Conference of the Society for Industrial and Systems Engineering

Rapidly Unlocking Insights from Army Experimental Data: A Topic Modeling and Natural Language Processing Approach

Anders Grau, Jenifer McClary, and Nicholas Reisweber

Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996

Corresponding Author's Email: andersgrau8@gmail.com

Author Note: CDT Anders Grau is studying for a Bachelor of Science in Operation Research at the United States Military Academy. MAJ Jenifer McClary and MAJ Nicholas Reisweber are instructors in the Department of Mathematical Sciences at the United States Military Academy. The authors would like to express their thanks to the sponsors of this research, TRAC-Monterey, for their support and guidance throughout the research process, with particular thanks to MAJ Daniel Ruiz and LTC Matthew Smith. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: Researchers across the U.S. Army are conducting experiments on the implementation of emerging technologies on the battlefield. Key data points from these experiments include text comments on the technologies' performances. Researchers use a range of Natural Language Processing (NLP) tasks to analyze such comments, including topic modeling. This research is dedicated to developing a methodology to analyze text comments from Army experiments and field tests. The methodology is tested on experimental data agglomerated in the Forge database, an Army Futures Command (AFC) initiative to provide researchers with a common operating picture of AFC research. As a result, this research offers an improved framework for analysis with topic models for researchers across the U.S. Army.

Keywords: Topic Modeling, N-Grams, Army Experimental Data, Text Comments

1. Introduction

Currently, researchers across the U.S. Army are conducting experiments on the implementation of emerging technologies on the battlefield. Among the variety of performance measures collected are text comments from experimenters that provide useful insights into these technologies' performances and weaknesses. Due to the large quantity of text comments collected, it is infeasible for researchers to read them all in a timely manner. Moreover, researchers' interpretations of the comments can be subject to bias. To help address these issues, a goal of Army data analysts is to unlock the insights from these comments using natural language processing (NLP). This will increase the speed that analysts can process data and get that information to senior leaders. This paper aims to outline a methodology with topic modeling that analysts can use to quickly and efficiently analyze Army experimental data.

1.1 Project Convergence and TRAC-Monterey

Project Convergence is an annual Army experiment testing its ability to integrate new technologies into the force at the tactical level. Project Convergence focuses its experiments specifically on artificial intelligence, cloud technologies, and autonomous systems (Smith and Ruiz, 2022). The experiment was conducted by Army Futures Command (AFC) and specifically aims to inform Army leaders on future concepts and capability requirements while showcasing future readiness and modernization (Smith and Ruiz, 2022).

The Research and Analysis Center (TRAC) is an Army data analytics center that directly reports to AFC and works to provide the Army and Department of Defense (DOD) with information on the challenges they face (*Army Futures Command*). TRAC-Monterey has been tasked to support Project Convergence 22 by providing data analysis on the experiments. One potentially useful data source for their analysis is text comments. Currently, experimenters involved in Project Convergence collect classified text comments on technologies' performances (Smith and Ruiz, 2022). Should there be an effective way to analyze these comments, the insights gained could be useful in TRAC-Monterey's analyses in Project Convergence and in other Army experiments.

1.2 Forge Database

To assist TRAC-Monterey's NLP analysis, this research uses data from the Forge database, developed by AFC. AFC is currently leading the Army's modernization and aims to develop capabilities for future soldiers to be successful on the battlefield (*Army Futures Command*). Forge is a unified data environment that provides data from AFC experiments, demonstrations and exercises to support common operating picture platforms and databases (Henderson 2022). Amongst the data collected in the database are text comments collected from experimenters during exercises. Analyzing them with topic models provides the opportunity to assess capability gaps in AFC efforts, a key task the Forge database was designed to accomplish (Henderson 2022). Moreover, being agglomerated from previous AFC experiments, the text comments are similarly structured to the classified comments collected during Project Convergence. Therefore, methods tested on data from Forge can be applied to similar data in other Army and AFC experiments.

1.3 Topic Modeling and N-grams

Breakthroughs in NLP modeling can potentially improve the way we analyze large amounts of natural language. For example, developments like BERT have improved the accuracy of NLP models (Devlin et. al., 2019). These developments have prompted greater application of NLP across many domains. Therefore, this research aims to develop NLP topic modeling techniques for military contexts in support of TRAC-Monterey's mission. These models are trained to analyze comments collected in the Army Forge database. The generated comments are then further analyzed through an n-gram analysis to verify the interpretation of the generated topic and gain further understanding of common themes in the clustered comments. Since topic modeling ultimately relies on human interpretation, combining topic modeling with topical trigrams ensured provided analysts with more information. Overall, this research aims to develop a methodology for topic modeling that integrates military domain-specific language models and applies this to the Forge database. Doing so not only provides analysis of existing text comments in the Forge database but also creates a framework for Army researchers to utilize in future work.

2. Literature Review

Of the range of NLP tasks, the one most suited to the comments in the Forge database is topic modeling for two primary reasons: it is an unsupervised method and can detect capability gaps. Topic modeling does this by generating abstract topics from a given set of documents. Given that the Forge database does not contain an assessment of the text comments, analyzing them requires an unsupervised method, such as topic modeling. Additionally, one of the tasks of Army Futures Command's (AFC) Top-Down Futures Development Process (TDFDP) is to assess capability gaps across its experimentation efforts (Henderson, 2022). Topic modeling is ideally situated to achieve such a task as it can potentially identify large-scale trends across experiments, to include areas of success and concern.

2.1 Topic Modeling for Comment Analysis

Topic modeling has been used in adjacent domains in identifying trends and classifying by topic. For example, Carreno and Winbladh (2013) used topic modeling to analyze user comments on mobile applications. This research looked at three applications: a calorie tracker, Mint.com, and Facebook (Carreno and Winbladh 2013). This work showed that topic could be applied to user comments and would generate topics that could be used by software engineers to make changes in requirements (Carreno and Winbladh, 2013). This is very closely related to TRAC-Monterey's goals in analyzing experimental results to recommend changes. One of the weaknesses of this study is that it takes its text sources from disparate applications. Though the data in the Forge database is from a range of experiments from across the AFC, the comments are not likely to cover as many topics as a database sourced on topics that include fitness, personal finance, and social media. Using a specific dataset like Forge would likely yield better insights than topic models trained on a more general dataset.

Similar research was also conducted on user comments specifically on diet-tracking mobile applications (Zecevic et. al., 2021). Like Carreno and Winbladh, this research used topic modeling to analyze user comments to identify areas for improvement, but it also utilized topical trigrams in its methodology (Zecevic et. al., 2021). Trigrams are a form of n-grams, a sequence of n words from a document, with $n = 3$ in the case of trigrams (Jurafsky and Martin, 2023). The topical trigrams acted as proxies for sentiment (Zecevic et. al., 2021). Though the model used by Carreno and Winbladh used a topic modeling method that incorporated sentiment analysis, this was not a Bidirectional Encoder Representations from Transformers (BERT)-based topic modeling technique (Zecevic et. al., 2021). BERT is a pretrained language model that uses bidirectional encoders to create word embeddings, leading to improved accuracy in language models (Devlin et. al., 2019). In contrast to Carreno and

Winbladh, the techniques used by Zecevic et. al. in their methodology are better suited to the goals of this research. Therefore, this methodology could be adapted to the requirements of this research.

2.2 Challenges Utilizing NLP Methods in Military Contexts

One of the central problems TRAC-Monterey is facing regarding the analysis of text comments from Project Convergence is that models trained on general English corpora are not always best suited for the language used in Army test comments. Previous studies have shown that training NLP models on domain-specific corpora can yield greater accuracy. For example, a study that created a parts-of-speech parser for medical texts achieved greater accuracy when trained on a combined corpus of an existing general English corpus and a smaller medical domain-specific corpus (Coden et al., 2005). Given that such a corpus is not yet available for military NLP, the methodology took a combined qualitative and quantitative approach.

3. Methodology

To implement topic modeling in an army experimental context, this research adapts the methodology proposed by Zecevic et. al. and utilizes the BERTopic Python topic modeling program. The process consisted of data processing, topic generation and analysis, and n-gram extraction and analysis.

3.1 Data Preprocessing

Text comments were obtained as part of a data frame of all events in the Forge database. This data was transferred into Python 3 for data cleaning and topic model analysis. After isolating the comments, they were cleaned in preparation for analysis. To clean the text, all capital letters were made lower-case and all punctuation was removed. With the comments prepared for processing, spam comments also had to be removed from the dataset. In the Forge database, many comments were nonsensical or tested the comment functionality, limiting the quality of the generated topics. These spam comments tended to form their own topics, so we ran BERTopic on the dataset. BERTopic is a topic modeling technique that aims to cluster documents relying on BERT as a pretrained model to generate document embeddings (Grootendorst, 2022). These embeddings are then clustered using HDBSCAN, a hierarchical clustering algorithm, and the key words are identified using c-TF-IDF, a measure that multiplies the frequency of a term in a document (TF) by the inverse document frequency (IDF) (Grootendorst, 2022). BERTopic was run on all the comments in the data frame and all the comments that classified into topics clearly identifiable as spam were removed from the data frame. From there, the data analysis could be conducted with the spam comments removed.

3.2 Topic Modeling and Topical N-gram Extraction

Once spam comments were removed, topic modeling could begin in earnest. A variety of topic modeling methods exist, including BERT-based ones like BERTopic (Grootendorst 2022). BERTopic was used to conduct topic modeling twice. In the first iteration, no stop words (common words such as ‘is’, ‘as’, or ‘a’) were removed prior to topic modeling. BERTopic outputs a data frame of the identified topic numbers, the amount of documents in each, and their associated key words as identified by TF-IDF. We also created a data frame of each document with the topic number it was assigned. On the second iteration, the researchers removed stop words from the comments by dropping words in the comments that appeared on the stop word list built into the NLTK package. The BERTopic modeling process was then repeated in the same way on the comments without stop words. Ultimately, to assess the quality of the topic models and resultant bigrams (n-gram with $n = 2$) and trigrams, a list of expected topics was generated based off the descriptions for the Forge events. The gnernated topic models were compared with the expected to see how well they aligned.

Following the creation of the topic models, in line with the research by Zecevic et. al., bigrams and trigrams were created for each topic in each of the two sets of models. Before we could do this, we removed stop words from the data frame with documents and their associated topic. This was necessary to ensure that the generated n-grams would have a similar form to those from the second iteration and that the n-grams did not contain stop words that could limit insights. This was done by using code that adapted an n-gram function in the NLTK package (Boldenow, 2019). The resultant output was a data frame of the n-grams and their counts, ordered in descending order.

3.3 Analysis of Topic Modeling and Topical N-grams

To compare the two sets of topics, we did a qualitative comparison and assessment. First, the researchers looked at the five most common topics from the titles of the AFC experiments associated with the analyzed text comments. These were

then compared to the generated topics and their key words to see if any of them clearly aligned with the most common subjects of the experiments. Then, a qualitative assessment was made of how well the two sets of topic models aligned with the expected topics and which did so better. We were also interested in how similar the two topic models were and if one generated more insightful n-grams. To do this, we identified the topics whose key words and n-grams aligned with the expected topics. Then we compared the documents' assigned topic and identified how often a document fell into the same topic across topic models. This was only done for topics that clearly mapped to one another.

From here, we compared the bigrams and trigrams within each topic to see whether bigrams or trigrams generated more insights into the generated topics by identifying commonly repeated themes to validate whether the topics generated align with the expected topics in the data. We elected to create both bigrams and trigrams to see which was more applicable in an Army experimental context. We hypothesized that trigrams would be better for topics with higher document count, as trigrams contain more information and, in a larger set of comments, will appear more frequently. On smaller topics, trigrams would not perform well due to low counts. We also hypothesized that bigrams would best for smaller topics as certain trigrams may occur infrequently when compared to bigrams. Finally, we were interested in which topic modeling method produced more valuable n-grams for analysts. To do this, we calculated the ratios of useful n-grams in the 20 most common n-grams for each topic, where useful is defined as an n-gram that would provide insight into a technology's performance.

4. Results

In pursuit of an ideal method to conduct topic modeling on Army experimental data, this research investigated the results of multiple methods. Firstly, we conducted data preprocessing. This process resulted in 433 documents from 82 AFC events. After the spam removal outlined in the methodology, we were left with 284 documents.

4.1 Topic Modeling: With and Without Stop Words

Based off the experiment descriptions, the five most common topics were the following: field artillery (FA) systems, combined operations, theater sustainment, multi-domain operations, and cyber operations. These topics are the ones we most expected to see in the topics that we generated and will be what we use to assess the quality of the topic models and bigrams and trigrams. Shown in Table 1 are the results of the topic modeling with the stop words removed and without the stop words removed. Note that BERTopic groups documents considered to be noise into a '-1' topic.

Table 1. Topic Modeling Results: With and Without Stop Words

	Stop Words Removed		Stop Words Not Removed	
Topic	Key Words	Count	Key Words	Count
-1	support, operations, force, systems, joint	118	the, and, to of, in	77
0	theater, army, operations, support, capabilities	130	to, and, the, of, be	64
1	JTF, corps, HQ, formations, joint	29	theater, and, the, to, in	46
2	LADS, minutes, soldiers, GPS, aiming	27	the, and, to, army, for	37
3	network, dependencies, concept, RCS, discussion	22	JTF, corps, the, and, to	35
4	site, identification, fixed, detailed, also	11	fires, the, to, and, DFC	13
5			SESU, threat, of, wing, fixed	12

The key words associated with each topic are those that have the highest TF-IDF scores across all the documents within a given topic. When we do not remove stop words, the key words tend to be stop words, such as 'the', 'to', 'and'. TF-IDF would normally weight these words lower, but due to the relatively small size of the data set this was not done. With these stop words removed, the terms related to the Forge events received high TF-IDF scores, becoming the key words for the associated topics.

Comparing the generated topics to the expected topics, topic modeling conducted after stop word removal more clearly aligns with the expected topics. For the analysis without stop words, Topic 2 aligns with the expected topic dealing with FA systems, with 'Location and Azimuth Determining System (LADS)' and 'aiming' identified as key words. Topic 0 aligns with the expected topic on theater sustainment, Topic 1 with the expected topic on mult-domain operations, and Topic 3 with cyber operations. Thus, the topic modeling without stop words identified topics whose key words correspond with expected topics.

In contrast, the topics from the analysis with stop words do not align as clearly with the expected topics. In most of the topics the key words are stop words. This means that only Topic 4 could be clearly aligned with FA systems due to 'fires' being a key word and Topic 3 could be aligned with multi-domain operations due to 'Joint Task Force (JTF)' being a key word.

4.2 Topical N-grams

Shown in Table 2 are the most common bigrams and trigrams for two topics from the topic modeling without stop words that illustrate the strengths and weaknesses of bigrams and trigrams. Result quality of topical n-gram analysis varied by the number of documents in a given topic. For topics with a greater number of documents, trigrams were frequent and highly informative. For topics with a lower number of documents, trigrams were infrequent but bigrams still offered valuable context of the documents. When comparing the insightfulness and quality of the bigrams and trigrams identified based on whether stop words were removed before or after topic modeling, the bigrams and trigrams were consistently better if stop words were removed prior to topic modeling.

Table 2. Most Common N-grams for Topics 0 (Theater Sustainment) and 1 (Multi-Domain Operations) from Topic Modeling with Stop Words Removed

Topic 0				Topic 1			
Bigram	Count	Trigram	Count	Bigram	Count	Trigram	Count
high altitude	12	established sustainment distribution	6	force posture	7	calibrated force posture	7
forward stationed	9	distribution systems processes	5	calibrated force	7	multidomain formations corps	2
threat IADS	8	theater well established	5	multidomain formations	6	corps transition JTF	2
sustainment distribution	6	well established sustainment	5	senior leader	6	force posture corps	2
established sustainment	6	high altitude capabilities	5	JTF HQ	5	corps operating JTF	2

Looking at Topic 0, we can see that both the bigrams and trigrams can help verify whether the generated topics align well with the predicted topics. In Topic 0, bigrams and trigrams such as ‘sustainment distribution’ and ‘theater well established’ verify that documents relating to theater sustainment form important parts of Topic 0. They also show that other documents not related to sustainment are included in Topic 0, with ‘high altitude’, a term related to air defense systems testing, being the most common bigram. These n-grams provide a qualitative assessment of the generated topic models and can provide greater insight into their contents than the key words.

Comparing the trigrams in Topic 0 and Topic 1, we can see that the quality of trigrams in Topic 0 are greater than in Topic 1. The trigrams in Topic 0 occur more frequently than those in Topic 1 and contain a qualitative assessment of the events, as in the trigram ‘well established sustainment’. While the frequency of the trigrams in Topic 0 is likely owed to the larger topic size, the same sentiment was expressed by multiple experimenters, allowing analysts to draw limited conclusions. In comparison, the trigrams in Topic 1 mostly occurred infrequently, with only 2 occurrences across the topic for most of them. These smaller frequencies limit the ability of researchers to draw larger conclusions about the documents.

4.3 Topic Model Similarity

For topics dealing with theater sustainment operations, 41 documents were shared between the topics. 27 documents were shared for multi-domain operations topics, and none were shared for the cyber operations and FA systems topics. The number of shared documents was lower than predicted, especially for topics with no shared documents and we wanted to see how the n-grams differed from one another. Comparing the topics, the models generated topics about different elements of the expected topics. For example, the topic relating to FA systems in the topic model without stop words discusses LADS, a FA survey system while the topic model with stop words discussed ‘fires’ more generally. We show these ratios in Table 3.

Table 3. Ratio of Useful N-grams by Topic and Method

Topic	Topic Model with Stop Words Removed	Topic Model with Stop Words Included
Theater Sustainment Operations	0.15	0.25
Multi-domain Operations	0.0	0.05
FA Systems	0.1	0.2
Cyber Operations	0.15	0.0

ISBN: 97819384964-4-8

As the table indicates, the n-grams from the topic model trained with stop words outperformed the model without stop words. This suggests that researchers most interested in the resultant n-grams from topic models should conduct topic modeling with stop words.

5. Discussion

5.1 Key Takeaways

This research set out to identify an effective method for topic modeling with Army experimental data using BERTopic. When topic modeling with Army experimental data, this research identifies stop word removal prior to the process as a best practice for greater interpretability of the models. Due to the use of TF-IDF in identifying topic key words, smaller datasets will generate topics with stop words as key words, causing topic interpretability issues for analysts. Since TRAC-Monterey has identified small datasets as a common issue, this research recommends stop word removal prior to topic modeling when most interested in key words. Looking at n-grams, however, topic modeling with stop words removed tended to yield more insightful results. Therefore, analysts interested in n-grams should keep stop words when modeling.

When generating topical n-grams, trigrams tend to be preferrable to bigrams. Trigrams have the potential to provide qualitative assessments of experiments and exercises more frequently than bigrams. The main limitation to trigrams is that they can occur less frequently, especially in topics with fewer documents. This results in trigrams with low counts that analysts cannot draw larger conclusions from. In such cases bigrams can still provide some qualitative assessment and be used to verify analysts' interpretations of topic key words. Therefore, this research recommends that analysts focus on trigrams in larger topics but use bigrams in smaller topics.

5.2 Limitations and Future Research

One of the main limitations in this research was the small dataset size from the Forge database. Larger datasets with tens of thousands of observations tend to yield better results when topic modeling. This smaller dataset, however, more closely resembles datasets used in Army experiments, such as Project Convergence. Therefore, it was worthwhile to identify the challenges of topic modeling with small datasets.

Another limitation was the focus solely on BERTopic. While BERTopic was chosen due to the ongoing development of military domain-specific language models, other topic modeling methods may have yielded different results. Future research could compare multiple topic modeling methods to identify which is best suited for Army experimental data.

Future research should also focus on the implementation of military domain-specific language models into topic modeling. Efforts to create domain-specific language models have extended to the military. Currently, researchers at MIT-Lincoln Labs are developing a military, domain-specific BERT model called MilBERT (Hallapy et al., 2022). The development of MilBERT represents a unique opportunity as one of the first efforts to develop a military, domain-specific pretrained deep language representation model. Future research should focus on modifying existing topic modeling techniques to incorporate these military-specific models by leveraging existing BERT-based topic modeling techniques and replacing BERT with MilBERT. This will allow for a topic modeling technique that accounts for domain-specific language.

6. References

Army Futures Command. (n.d.-a). Retrieved September 18, 2022, from https://armyfuturescommand.com/trac/.

Army Futures Command. (n.d.-b). Retrieved December 8, 2022, from https://armyfuturescommand.com/supporting-commands/.

Boldenow, B. (2019). *Textual Data Exploration with N-Grams*. https://kaggle.com/code/boldy717/textual-data-exploration-with-n-grams

Carreño, L. V. G., & Winbladh, K. (2013). Analysis of user comments: An approach for software requirements evolution. *2013 35th International Conference on Software Engineering (ICSE)*, 582–591. https://doi.org/10.1109/ICSE.2013.6606604

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., & Chute, C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, *38*(6), 422–430. https://doi.org/10.1016/j.jbi.2005.02.009

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Grootendorst, M. (2022, March 11). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. ArXiv.Org. https://arxiv.org/abs/2203.05794v1

Hallapy, J., Hawkins, T., Kelley, T., O'Brien, C., & Zipkin, J. (2022). *MilGLUE: A Multi-Task Benchmark Platform for Natural Language Understanding in the Military Domain.*

Henderson, M. (2022). *An Independent Analysis of the Forge Database Design*. Naval Postgraduate School.

Jurafsky, D. & Martin, J. (2023). *Speech and Language Processing*. [Unpublished manuscript].

Ramamonjisoa, D. (2014). Topic modeling on users's comments. *2014 Third ICT International Student Project Conference (ICT-ISPC)*, 177–180. https://doi.org/10.1109/ICT-ISPC.2014.6923245

Smith, M. & Ruiz, D. (2022, Aug. 24). *Project Convergence – Overview and Potential Research Topics* [Brief]. Microsoft Teams meeting, online.

Zečević, M., Mijatović, D., Koklič, M. K., Žabkar, V., & Gidaković, P. (2021). User Perspectives of Diet-Tracking Apps: Reviews Content Analysis and Topic Modeling. *Journal of Medical Internet Research*, *23*(4), e25160. https://doi.org/10.2196/25160