Analysis of INCOSE Systems Engineering Journal and international symposium research topics

Rudolph Oosthuizen [1,2] and Leon Pretorius[2]

[1]Council for Scientific and Industrial Research, Defence and Safety, Pretoria, South Africa

[2]

Graduate School of Technology Management, University of Pretoria, Pretoria, South Africa

*Correspondence

Rudolph Oosthuizen,Graduate School of Technology Management, University of Pretoria, Pretoria, South Africa.

Email: rudolph.oosthuizen@up.ac.za

Abstract

The pressure on systems engineering is ever-increasing to support the development and implementation of systems that meet a complex environment's demands. As a growing discipline, systems engineering requires insight into past research to identify opportunities for future growth. Analyzing the bibliometric data on published research provides valuable information on a scientific discipline's past progress and future prospects. Therefore, this paper extracts the research topics published in INCOSE's journal *Systems Engineering* and the annual international symposium proceedings to analyze their composition and allocation to papers. The implemented process applies natural language processing and topic modeling to extract the main topics from these papers' titles and abstracts. Analyzing these research topics' composition and mapping them to processed articles helps to understand their relative importance. The analysis's output confirms the importance of modeling in systems engineering, as it is the most popular topic. The additional focus of research papers on the systems engineering process, practice, and methodologies also indicates that the field is still growing and evolving. Some important topics to systems engineering, which were not found as prominent topics, are humans' roles in systems, verification and validation, and other specialty fields. This new knowledge about the structure of research into systems engineering can identify future research project opportunities to continue growing the field.

Keywords

Systems engineering, Research, Bibliometrics, Natural language processing, Topic modeling

1 Introduction

The International Council on Systems Engineering (INCOSE) Handbook defines systems engineering as an interdisciplinary approach and means to realize successful systems [1]. Systems engineering is a multidisciplinary profession, process, and perspective that considers the whole system rather than its parts. It employs a process that defines customer needs and requirements before designing and validating a solution system. The systems engineering effort integrates multiple specialty disciplines that proceed from a system's concept solution to production and operation. The solution system also has a life cycle, function, structure, behavior, and performance characteristics [1].

Although systems principles and systems thinking have existed for many years, the field of modern systems engineering only became formalized after World War 2. Engineers required an approach to support the development of weapon systems that continually increased in complexity [2]. Since then, the concepts of system design, analysis, and development have evolved into the systems engineering discipline. A knowledge base of

systems engineering methodologies, tools, and management techniques is currently formalized in a series of handbooks, standards, and other guides to provide a foundation for the discipline [3]. However, systems engineering is still relatively young and growing compared with other engineering fields. Only in the early 1990s was systems engineering formally defined as a discipline [4]. It has to continue developing processes and tools, supported by theories, through focussed research to cope with the growing complexity of engineering projects [5].

Research is an activity that aims to create knowledge by understanding, explaining, and predicting phenomena [6]. The typical starting point for a research project in systems engineering is a need for improvement, often triggered by an industrial problem [7]. Valerdi and Davidz [4] noted in 2009 that the tendency to rely on intuition and revelation in systems engineering research, instead of on a rigorous scientific process, hampers progress in the field. In the past, researchers focused on convincing the audience of systems engineering's potential rather than building theories, based on empirical results [8]. Researchers should instead focus on the long-term intellectual establishment of the field. Antons, Kleer, and Salge [9] found that evaluating the topic landscape of research publications in a research field may help researchers detect meaningful research opportunities [10].

This paper aims to extract and analyze the research topics in articles from official INCOSE academic publications. In the next two sections of this paper, bibliometric analysis is presented as a useful and valid method to trace the development in a research field. Data mining, implemented through natural language processing (NLP), is also discussed as an analytical tool for generating data on systems engineering research topics as part of bibliometric analysis. The NLP and topic modeling process steps are then presented in more detail before its implementation.

Research topics were extracted from the papers of the journal *Systems Engineering* and the proceedings of INCOSE's annual international symposium, using the proposed automated NLP techniques. Several validation steps are performed to ensure that the extracted topics accurately represent systems engineering research. These topics were explored for the structure and allocation of these topics to papers to understand the current state and identify future research opportunities. The allocation of these topics to papers helped to reveal the number of publications for each topic. Finally, these results are briefly discussed to highlight key findings.

2 Bibliometrics

Scientometrics, the "science of science," provides quantitative and statistical techniques to measure progress in the development of a research field through analysis of the published literature. Scientists tend to codify their findings in publications, which are the building blocks of science. Peer reviews and expert-based judgment validate the knowledge captured in published papers [11,12]. Therefore, a bibliometric analysis of these publications should quantify the state of research in the field of systems engineering.

Bibliometrics is a common systematic analysis approach to scientometrics. It is suited to most behavioral, engineering, and other scientific research fields. The bibliometric approach relies on the assumption that scientific research produces knowledge published in the scientific literature. The mapping of bibliometric outputs can visualize patterns and trends in scientific data, thus assisting researchers in identifying gaps and opportunities to focus their scholarly work [13,14]. Research projects based on these outputs should better advance knowledge within the field. Many scientific fields already apply the quantitative analysis of publication and citation data to map research growth and maturity. Published results of the approach focus on leading authors, publication performance, and citation trends [15,16].

Researchers may apply bibliometrics to analyze a single journal, a set of journals, or a whole domain to assess a research field [12]. Performance (citations) or science mapping (conceptual structure) can describe the research progress in a scientific field. Typical bibliometric indicators include citations, author statistics, paper publications numbers, keyword trend analysis, relational indicators (co-publication and co-citations), and research topics [10,12,16,17]. This paper applies topic modeling to identify the research trends in systems engineering over recent decades. Analysis of the topics may assist researchers in identifying new issues and concepts in a research field [18]. Researchers may perform manual or automated topic modeling to extract core topics from published research.

Traditionally, researchers manually assigned papers to a predetermined topic list based on subjective judgment and subject-matter experts' (SMEs) input. With the increasingly large number of papers available from the journals and symposium proceedings, reading and sorting each into manually identified topic fields is time-consuming and challenging, with a high risk of bias. This approach is costly in time and resources. Manual topic modeling may

also miss the latent and emerging topics in a large text corpus. Also, a single article may contain multiple topics. Implementing automated machine-learning methods should improve topic modelling of large text corpora [19]. This paper's research applied text-mining techniques to analyze paper titles and abstracts to identify the main topics from each of the publications [10].

3 Natural language processing

Most of the data available on the internet and other databases are unstructured free text, which software analytics struggles to understand and analyze. Text mining is the process of structuring unstructured text to derive information and meaning. NLP helps machines to interpret human language. The process usually includes converting text into numerical values for machine-learning algorithms to process. This process's outputs include key phrases, relationships, and patterns for researchers' evaluation and interpretation. NLP can also categorize and cluster text for extracting information to classify and summarise documents. Other uses of NLP include speech recognition and language translation [20,21].

Lately, text-mining approaches, such as topic modeling, have become more useful to researchers, due to the accessibility of software and information along with the availability of increasing computer processing power. Topic modeling is an unsupervised text classification method that extracts semantic information from text using quantitative statistical algorithms. Topic modeling methods use computational algorithms to automatically organize, understand, search, and summarise a large text corpus to discover latent themes and structures for researchers to interpret. The process is also independent of any prior understanding of the corpus documents to extract the thematic concepts [13,21].

Topic modeling assumes that the mixture of words in a document constitutes a set of latent topics. The topic modeling algorithm analyses these words' occurrence and hidden relationships in the documents to define each topic as a probability distribution of selected terms [13,22,23]. Latent Dirichlet Allocation (LDA) is currently a popular topic modeling algorithm. It provides a generative statistical model in which unobserved groups explain the similarity of the data. The model learns the distribution of topics in each document with their associated word probabilities to identify major thematic clusters from an extensive corpus of text documents, usually beyond human capacity [24,25,26]. This field continues to grow in popularity for the scientometric analysis of research disciplines to discover latent semantic topics and structures [13,23].

LDA requires the researcher to define the required number of topics for extraction as input to the process. The algorithm then probabilistically forms document-topic and topic-word pairs. A distribution of extracted keywords represents each of the topics. The algorithm evaluates each publication in the text corpus for association with a set of topics using a probability value. Interpreting, naming, and describing each topic still requires expert and domain knowledge [10,21,23].

The topic results are not deterministic because of the built-in stochastic processes and are influenced by the input parameters. Also, random seed values for each run of the algorithm may result in a different set of topics. Researchers can limit the inaccuracies by selecting suitable parameters for the algorithm and validation of the outputs [21,22]. This places a premium on evaluation and validation of the output topics. In essence, a useful output requires a cleaned and prepared text corpus, a suitable set of parameters followed by thorough inspection and naming of the topics [26].

Despite these challenges, topic modeling may be more comprehensive and faster than other manual methods for performing an exploratory literature review, especially when analyzing large amounts of literary publications. Topic modeling algorithms also record all the data from processing steps and output files to support validation and post-processing. The recorded processing outputs enable researchers to expeditiously navigate the papers' to focus on a more in-depth analysis of critical elements within the literature. Lastly, the automated process provides information in a digital and structured format for improved analysis and reporting [27].

4 Method

4.1 Data sources

The purpose of this paper is to extract research topics from literature published on systems engineering. Systems engineering research outputs are published on a wide range of academic publication platforms due to the field's multidisciplinary nature. However, only a few sources focus primarily on systems engineering. Of these, the two primary and accessible platforms are affiliated to INCOSE:

1. *Systems Engineering* journal. *Systems Engineering* is INCOSE's journal for publishing research articles on systems engineering technologies, processes, and management approaches. The journal's stated goals are integrating and disseminating systems engineering knowledge, promoting collaboration between major role-players, and supporting appropriate professional standards. The journals' readership includes systems engineers, systems programmers, and computer system engineers [5]. Since 1998, *Systems Engineering* has published at least four issues per year.
2. INCOSE international symposium proceedings. Since 1991, the INCOSE international symposium has been the leading annual gathering of systems engineers to attend presentations, case studies, workshops, tutorials, and panel discussions. The program usually attracts a mix of international systems engineering professionals at all levels, including practitioners in government and industry, educators, and researchers.

A prerequisite for the automated topic modeling of published research is the accessibility of an article's bibliometric data (e.g., authors, year, affiliation, title, abstract, keywords, and citations) in digital format. Research databases (such as Scopus) or the publisher's website provide access to bibliometric data. Automated applications have to be able to access these repositories to extract the raw data. When thousands of papers need to be analyzed, manually downloading each article process becomes too time-consuming.

The two sources from INCOSE provide a representative sample of published systems engineering research for this research. Other sources that could be included in future research are the *Conference on Systems Engineering Research* (CSER), the *Institute of Electrical and Electronics Engineers (IEEE) Systems Journal*, the *American Society of Mechanical Engineers* (ASME), etc. However, many of these journals may not have a pure systems engineering focus. When using these data sources for the current study, the relevant systems engineering articles must be filtered and extracted first.

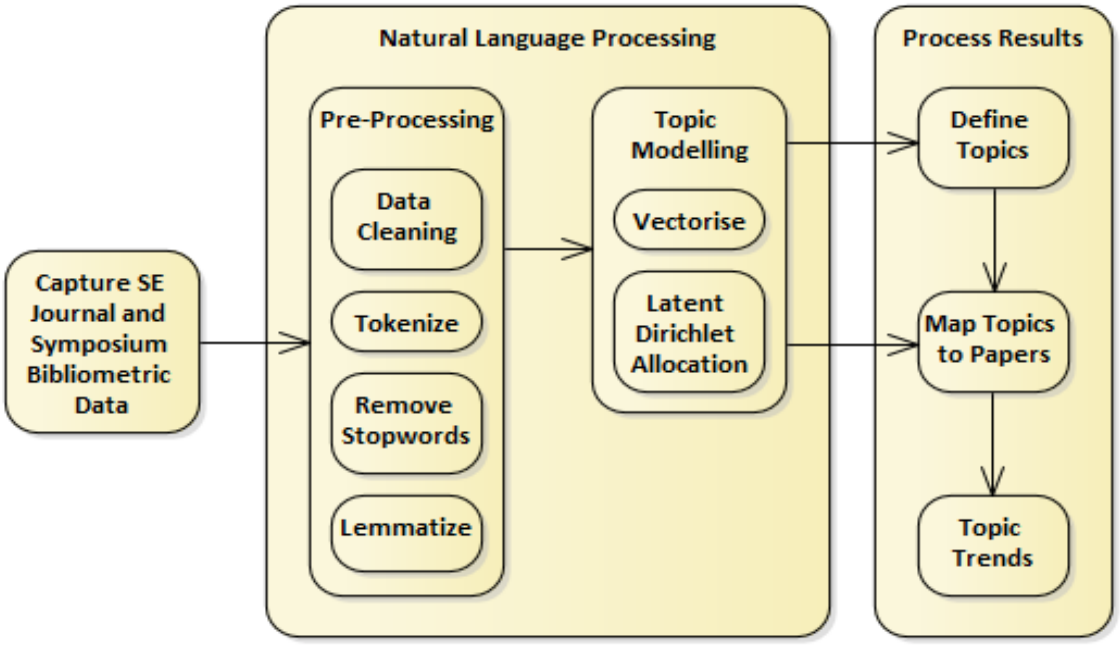

Figure 1: Bibliometric analysis process

4.2 Research process

The research process applied in this paper, as seen in Figure 1, was implemented to evaluate and analyze the titles and abstracts from the journal *Systems Engineering* and the proceedings of the annual INCOSE international symposium. The research process implemented Python-based algorithms and NLP tools. The authors also applied this process previously to assess the research trends for technology management [28].

4.2.1 Capture *Systems Engineering* journal and symposium bibliometric data

Usually, an NLP implementation starts with an extensive collection of documents. A library in Python, called Pyblometrics, was used to extract bibliometric information on the *Systems Engineering* journal from the Scopus database. Rose and Kitchin [29] developed Pybliometrics to access Scopus data through a consistent and straightforward interface that can integrate with Python's data science ecosystem for machine learning and visualization tools. Since Scopus does not list the INCOSE international symposium proceedings, the authors had to apply web-scraping with the Beautiful Soup library in Python for capturing the publisher's website's bibliometric data. The web scraper accessed all the pages on *Wiley.com*, where the symposium proceedings are published. The algorithms exported the data in a spreadsheet format with the following main fields:

1. Year of publication.
2. Authors.
3. Paper title.
4. Abstract.

These four fields represent only a small set of the bibliometric information types available from Pyblometrics. The outputs of the two systems engineering literature sources from INCOSE, mentioned in section 4.2, were combined into a single file for processing in the remainder of the steps. The combined data set provides a more extensive collection of text data for improved topic modeling. However, the data points were labeled according to the sources to enable separation during post-processing.

4.2.2 Natural language processing

4.2.2.1 Pre-processing

The SpaCy library in Python was applied to perform NLP on the captured systems engineering bibliometric text data. The text required pre-processing to be prepared and structured for analysis. Since abstracts may be relatively short, the titles and abstracts were combined to increase each document's size for improved topic extraction. The article's title tends to contain the most representative words. Simultaneously, the abstract summarizes the problem and results from research [21]. The list of author keywords was not included because it was not present in the earlier papers.

The text was then cleaned using Excel and Python functions. To ensure that the semantic meaning of key systems engineering terms or phrases would not be lost, they were converted to their abbreviated forms. These abbreviations tend not to be common words of the English language, which may be transformed by NLP processing. Typical examples of this step include changing "systems engineering" to "SE," "system of systems" to "SoS," "system modeling language" to "SysML," etc. This step also addressed multiple subtle variations in the phrases' spelling to be replaced by one abbreviation. Also, some spelling differences of keywords between the United Kingdom and the United States English (e.g., "modelling" vs "modeling," "behaviour" vs "behavior") were changed to the United States format.

The input text also contained punctuation (i.e., periods, commas, exclamation points, ampersands), whitespaces, letters, numbers, and special characters that could affect the NLP tokenization algorithms, which needed removal. Despite punctuation adding meaning to the text for a human reader, it is undesirable and uninformative for the NLP algorithm. The remainder of the text was then converted to lowercase for term unification. Tokenisation uses the spaces in the text to extract the linguistic units that represent the building blocks for sentences or paragraphs [30,31].

The next step removes the stop words from the text. Stop words are everyday words in a language that do not add real meaning to the text, such as "and," "the," "if," "a," etc. These words are not specific enough to represent document content for topic modeling. Stop words also have a high frequency that adds noise during text

vectorization [26]. Therefore, removing stop words reduces the text's dimensionality. Lemmatization also reduces the dimensionality of the text by calculating the lemma of each term. This is preferred to stemming, which linguistically normalizes text by reducing words to their root through truncation without considering the word's context. Lemmatization applies a vocabulary with a morphological analysis to transform the base word that is more linguistically correct and preserves the meaning of the word [10,21,26,30].

Table 1: Words excluded from the input text

Achieve	Characteristic	Document	Goal	Know	Order	Represent	Term
Address	Common	Early	Good	Large	Overall	Research	Test
Aim	Consider	eg, etc, ie	Great	Lead	Paper	Result	Time
Allow	Current	Evaluation	Help	Learn	Particular	Scale	Type
Analyze	Datum	Examine	High	Lesson	Point	Select	Use
Analysis	Deal	Example	Identify	Level	Possible	Set	Valuate
Apply	Define	Experience	Illustrate	Literature	Present	Share	Way
Area	Definition	Explore	Important	Long	Propose	Show	Well
Article	Demonstrate	Finally	Improve	Major	Provide	Significant	Wide
Aspect	Describe	Find	Include	Meet	Purpose	Single	Work
Associate	Determine	Focus	Increase	Multiple	Real	Specific	World
Author	Different	Follow	Introduce	Necessary	Reduce	Start	Year
Call	Discuss	Future	Involve	New	Relate	Study	Suggest
Achieve	Discussion	General	Issue	Number	Number	Suggest	
Case	Document	Give	Key	Offer	Report	Take	

Due to the academic nature of the published articles, many common words also do not provide information on systems engineering's specific research topics. Table 1 provides a sample of the academic terms removed from the text sample to improve the validity of systems engineering-specific topics.

Lastly, the terms "SE" and "system" were also removed as they are too common throughout the text sample with a high occurrence frequency. These words are likely to have a high conditional probability in many topics but do not contribute to identifying the specific topics within systems engineering. These terms may skew the definition and naming of topics 26. However, when naming the extracted topics, these terms could be reintroduced to improve the topic's description. The next step processes the remaining words to extract the main topics from the corpus of titles and abstracts from the sample papers.

4.2.2.2 Topic Modeling

Topic modeling is the critical bibliographic analysis process step used in this research. The Scikit-learn library in Python vectorizes the text and performs topic modeling to provide topics with their classifying keywords. Python's Scikit-learn applies the CountVectorizer function to transform all the words remaining from the paper's titles and abstracts into a document term matrix. The algorithm can also extract phrases that have a meaning independent of the individual words. The Scikit-learn library in Python also provides the LDA function to process the data term matrix to extract the topics. The outputs of the LDA function are the following:

1. The predefined number of topics with their describing terms (words and phrases).
2. The importance of the terms to each topic.
3. The probability of each topic to be present in each document.
4. The perplexity of the topic model.

Because the LDA algorithm implements a random number generator to initiate its training, different runs with the same text input may result in unique output topics. The selection of LDA parameters must ensure the stability of the LDA model [25]. Determining an optimal parameter set is not easy with such an unsupervised, data-driven algorithm. There is not yet a standard approach to evaluate these models [26]. However, perplexity is an important measure to determine the statistical goodness of fit, or quality, of the topic model [32]. It is the inverse of the geometric mean per-word likelihood and estimates a probabilistic method's ability to predict a sample. A lower perplexity indicates the better, and perhaps more appropriate, LDA model [21,22]. Perplexity is used to identify the best-suited set of parameters for a valid topic model for this research [10,14,21,33]. Other related quantitative diagnostic metrics, not addressed in this paper include topic coherence and mutual information measures. The main topic model parameters are the following [26,27]:

1. Shape and size of the document term matrix. The algorithm to vectorize the text applies a minimum and maximum document frequencies setting (max_df and min_df), which allows the algorithm to ignore words that are too common or too scarce over the range of papers in the corpus. For example, a max_df of 0.8 means that the algorithm ignores terms that appear in more than 80 percent of the documents. A min_df of 0.1 implies that terms must appear in at least 10 percent of the papers before being considered for topic modeling. Min_df has a significant impact on the number and variety of terms available for topic modeling.
2. Number of topics. The topic number adjusts the granularity of the topic model. More topics provide more specific and narrowly defined descriptions. However, this may result in many similar topics that are difficult to distinguish and name meaningfully. A small number of topics may provide too broad entities that combine different aspects that could have value independently. The size of the corpus and variety in the topic terms affect selecting a suitable number of topics.
3. Prior parameters (α and β). These parameters are not considered to determine an optimal topic model in this research. The default values provided in the algorithm are used.

However, quantitative diagnostic metrics are only one measure of the accuracy and utility of an LDA model. Other heuristic-based characteristics should also be satisfied, such as the corpus size, the simplicity of the topic descriptions, an adequate number of topics, and the number of terms available for topic description. Determining the optimum set of parameter values involves running several combinations of the parameters and calculating the topic model's perplexity number. Using a single numerical parameter optimization procedure must be avoided. Instead, several different measures should be combined with subjective qualitative human judgment. The final set of parameters often tends to be "good enough," but seldom optimal. Cross-validation with the interpretability methods is required to improve the model [26,27,35]. These will be addressed in the next section.

4.2.2.3 Topic validation

The LDA model and the output topics from the process discussed above, using the selected parameters, still require evaluation. Since topic modeling does not have an external ground-truth, validation of topics and their meanings is not a trivial exercise. The probabilistic nature of topic modeling algorithms, using random seed values, may produce non-deterministic results. However, a good model with proper parameter values should result in relatively minor differences in topic identification and paper allocation that are good enough to support bibliometric research of an extensive publication corpus. The evaluation of topic models is still immature despite research in these approaches having accelerated over the past few years [26].

The evaluation should consider interpretability, replicability, external validity, and internal coherence between documents and topics. Statistical measures should not be the only method of validating a model but have to include semantic and predictive methods [35,33]. The domain-specific nature of the topics demands domain expertise for the assessment of validity. Validity depends on the extent to which a measuring procedure represents the intended concept [25,26,35]. Semantic validation can include SMEs to evaluate the interpretability and plausibility of the topics. External validation verifies that the topics reflect relevant information to the research without being uninformative, misleading, or wrong. The typical steps in a validation approach should include the following [26,27,35,33]:

1. Implement a reliable process for useful quality data, cleaning, and preparing. The correct order of the steps in text processing is essential. The pre-processing steps can have a significant impact on the validity of results. The process applied in this paper is in line with guidelines from published literature on topic modeling.

2. Ensure that the most suited set of initial parameters of the LDA process is selected for the research. This process is covered in section 5.2.
3. The topics' ability to model the phenomenon under investigation in the text corpus needs to be evaluated. For this, SMEs are utilized to interpret and name the extracted topics. Human interpretation and understanding of the LDA outputs provide a measure of intra-topic semantic validity. SMEs inspecting the model output can also determine if the output topics make sense in the research context. This interpretation and naming of topics will be discussed in 4.2.3.
4. The top topic terms are further investigated to determine the face validity of the model. Topic terms that are out of place in the context of the specific topic and the research field may indicate low validity.
5. Analysis of the input documents and their allocated topics also validates the model. A random set of documents per topic is evaluated to determine the accuracy of the allocation. The purpose of this evaluation is to detect any misplaced papers.
6. The final form of validation is to compare the outputs to known other published research on systems engineering topics. The systems engineering research topics identified using the topic modeling methodology in this paper have to be in line with previously published research. Where discrepancies occur, these should be subjected to more in-depth investigation for an explanation.

4.2.3 Process results

The Pandas and Numpy libraries in Python were applied in conjunction with Excel to process the NLP and topic modeling outputs to achieve this paper's objectives. The LDA algorithm could only cluster documents by their topics, without an indication of what they were. Manual analysis still has to interpret the bibliometrics results and assign a topic name to each cluster. Word clouds, tag clouds, or Wordles were implemented to support the topic description. A word cloud, or weighted list of words, represents text data, where the importance of each tag is depicted using different font sizes or colors. This format helps to form a quick visual image of the most prominent terms relative to others. Instead of frequency, the relative importance value of terms defining the topic can represent the significance of each word [34]. The word clouds were presented to a number of systems engineering SMEs through a focus group to help define each topic. The word cloud analysis was supported by a table containing the top 15 terms and their relevance scores.

5 Results

5.1 Pre-Processing

Accessing and extracting the publication bibliometric data, as described in section 4.2, resulted in 622 papers from the journal and 3 694 papers for the symposium proceedings. The authors manually removed the non-research articles, such as errata, correspondence, and editorials. Figure 2 shows the distribution of the publications over the period. The symposium proceedings have an average of 128 papers per year, while the journal only provides 29 papers per year.

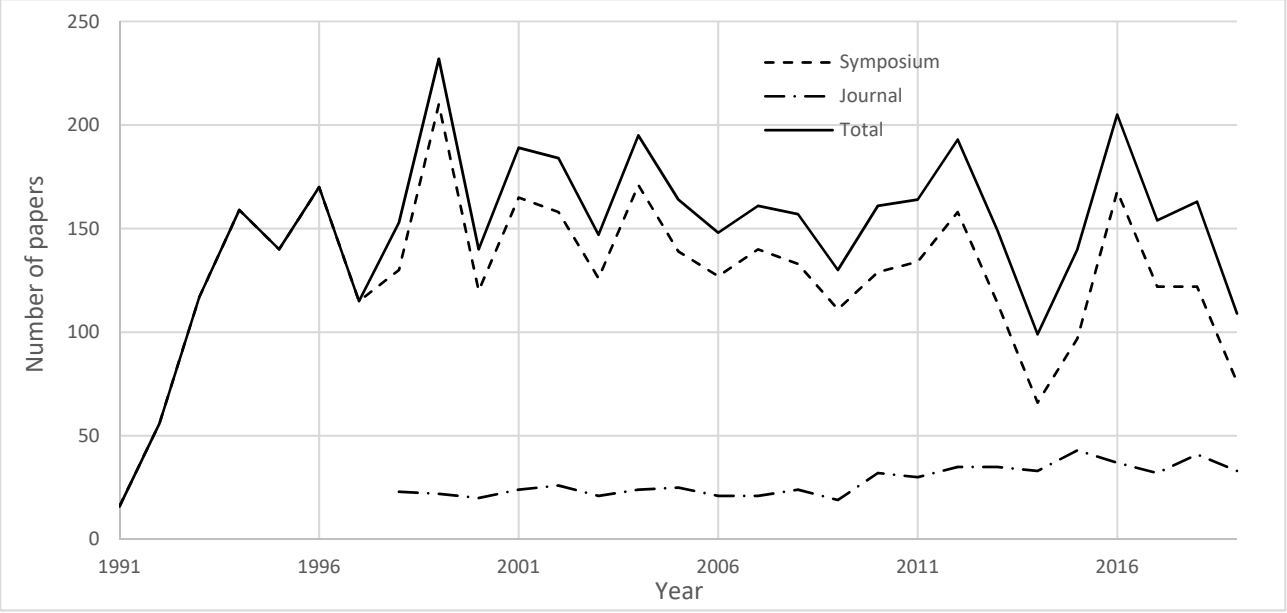

Figure 2: Papers published per year

Combining the two sources increased the input data for topic modeling to 4 316 documents for processing to extract the topics. As Isoaho [35] noted that effective topic modeling requires at least 1000–2000 documents of about 100–200 words in length (at the start of the process), the corpus in this research should be adequate for topic modeling. The histogram in Figure 3 shows the number of words per document (combined title and abstract per paper) before and after pre-processing the input text. The total number of words processed to extract topics was 297 541, with an average of 69 words per paper. The text cleaned and prepared for topic modeling represents 44 percent of the original text from the extracted documents.

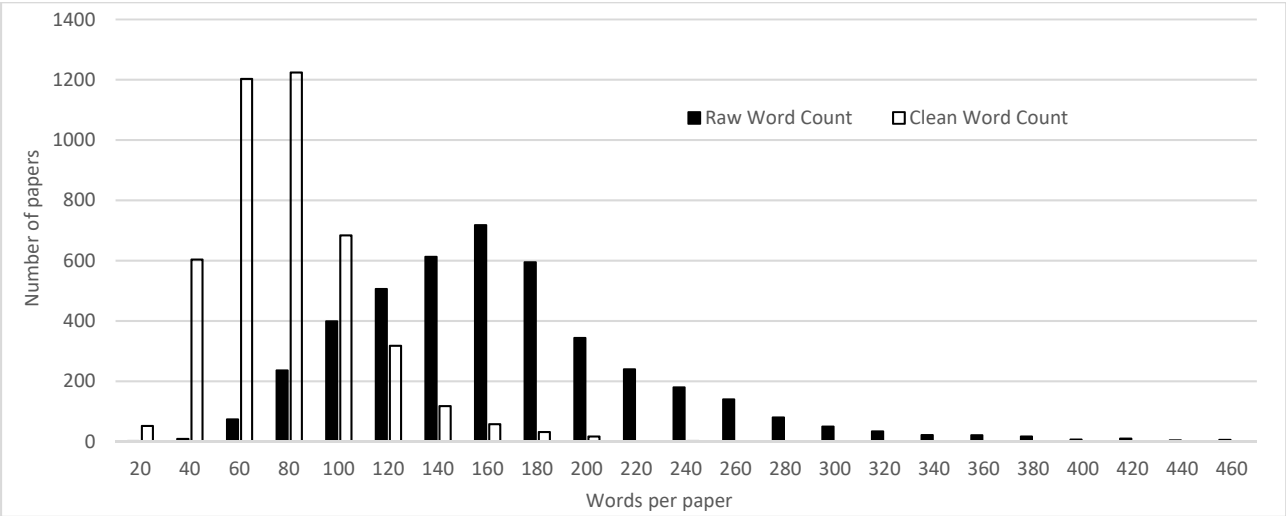

Figure 3: Document sizes

Figure 4 shows the frequency of the top 30 words that occur in the prepared text data. Note that the terms "systems engineering" and "system" were removed for the topic modeling. They add no additional meaning to the identification of topics within the field of systems engineering. However, these terms were considered and reintroduced when naming the extracted topics. In various combinations, these words form the basis for extracting the topics from the text corpus.

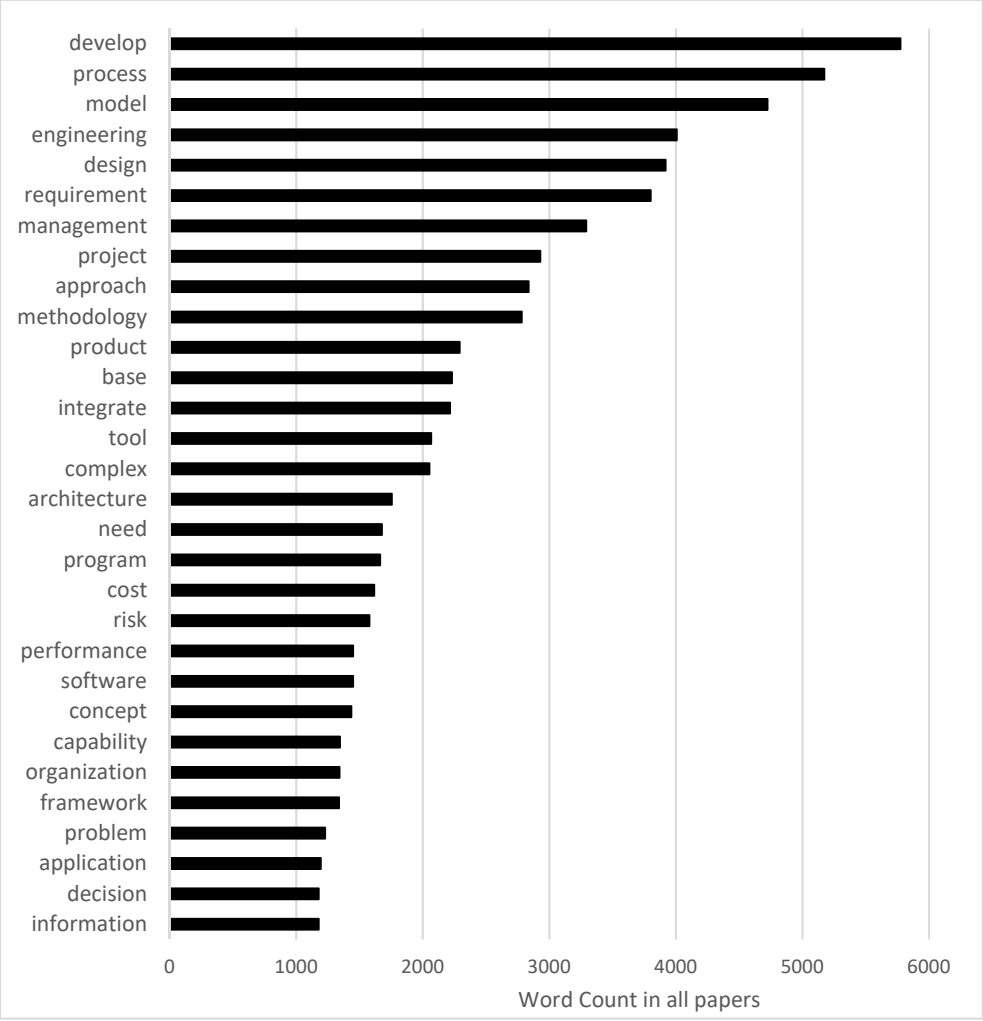

Figure 4: Word counts

5.2 Parameter Selection

The LDA model's perplexity is calculated to select the main input parameters, such as max_df, min_df, and the number of topics to extract (as per section 5.2). An iterative parameter analysis was performed for the values shown in Table 2. The boundaries of these values were determined using a limited number of preliminary exploratory runs. Min_df was found to be the most sensitive parameter as it determines the number of different words for inclusion into the analysis vocabulary. A larger size of the processed vocabulary results in more terms being available to define the extracted topics. If the pool of terms is too small, it may not be possible to extract enough unique topics, missing some more subtle or hidden topics.

Table 2: Iterative parameter list

Parameter	Start value	End value	Step size
Topic Number	7	30	1
Max_df	0.85	0.99	0.01
Min_df	0.04	0.2	0.01

Table 3 presents a sample of the output values from the iterative parameter analysis as an extract of the 5 520 iterative runs. From these results, it seems that a model with a small number of terms extracted from the texts results in a better (lower) perplexity score. The model has more confidence in producing a small number of reliable topics. Due to the high selected min_df value, the smaller diversity in the terms available results in a lower probability of errors for the smaller number of expected topics. This model is not quite suited for the research in this paper as we aim for a broader range of topics covered in systems engineering research. As the min_df values decrease, more terms become available for more topics, resulting in an increasing perplexity score. Therefore, the model requires a balance between the span of possible topics extracted and the accuracy thereof.

Table 3: Sample of parameter analysis output

Topic Number	Min df	Max df	Number Terms	Perplexity
9	0.19	0.88	19	16
10	0.16	0.87	28	23
13	0.14	0.87	37	29
16	0.13	0.97	42	33
17	0.12	0.92	47	36
20	0.11	0.9	54	41
22	0.09	0.93	65	49
26	0.07	0.87	96	69
28	0.06	0.87	115	79
28	0.04	0.87	183	114

A focussed literature search (which will be discussed in section 5.4.3) into research published about systems engineering topics provided a consolidated list of about 20 topics. This number of topics offered a good starting point to determine the LDA model's input parameter values. The authors also heuristically preferred having at least twice the number of topic terms than the required number of topics for improved topic identification. After manually investigating several topic sets that resulted from a sample of the parameters in Table 3, the shaded cells' values were selected for the detailed analysis in this paper. These input parameter values produce 20 topics from 54 terms with a perplexity score of 41, using a min_df of 0.11 and a max_df of 0.9.

5.3 Systems engineering research topics

Table 4 presents the topics extracted from the journal and conference proceedings. The topics were identified using the topic analysis process from section 4.2.3. Systems engineering SMEs participated in two online focus group discussions to identify the topics and analyze the terms associated with each topic. A sample of four of the word clouds used to identify and name the topics is shown in Figure 5. The full set of word clouds is not shown in this paper due to current paper length limitations.

The evaluation team consisted of 10 systems engineers with an average of 24 years' experience. Seven of the participants were members of INCOSE, while three were Certified Systems Engineering Professionals (CSEP). Recorded transcripts of the focus group discussions supported finalizing the description (name) for each topic in support of the title. Table 4 also provides the main topic terms and a short description of what the topic entails as a context to understanding the name. The authors manually analyzed the top papers associated with the topic by the highest allocation probability (provided by the LDA algorithm) and the SME discussion transcript to derive a description for each topic.

Table 4: Extracted topics

No	Topic name	Top topic terms	Description
1	Architecting	Architecture, develop, approach, framework, need, function, concept, design, base, solution	Defining, developing, using, synthesizing, optimizing, analyzing systems (functional and physical), and enterprise architectures. The implementation principles of architecture frameworks, service-oriented architectures, and reference architectures.
2	Modeling	Model, base, develop, approach, engineering, complex, process, application, concept, information	Application of modeling in systems engineering, including the roles, advantages, and implementation of model-based systems engineering. Modeling approaches, standards, frameworks, and languages for computational, conceptual (functional) modeling are included.
3	Education and Program Management	Program, technique, develop, success, structure, engineering, implement, require, activity, performance	Education (university degrees) and accreditation programs as well as a compilation of several large projects (especially in the government and defense domain) due to dual use of the word "program."
4	Risk management	Risk, management, cost, approach, performance, effective, process, planning, base, methodology	Processes for risk modeling, analysis, and management to handle uncertainty in systems engineering projects and ensure return on investment.
5	Integration	Integrate, develop, process, planning, management, team, tool, product, program, environment	System integration (processes and plan), integrated systems engineering and development environments, and integrated systems (of systems) supported by integrated capability maturity models.
6	Systems Engineering Methodology	Methodology, problem, approach, base, develop, complex, application, technique, model, performance	Methods applied as part of and supporting systems engineering for problem-solving, evaluation, and development. These also include systems approaches, systems thinking (soft systems), and philosophical issues.
7	System Operation	Operation, function, industry, environment, concept, need, capability, develop, understand, performance	Operational concepts with reference to scenarios, functions, modes, support, and environmental impact. The concept solution's ability to achieve the operational potential that satisfies stakeholder's needs for operation in the industry.
8	Systems Engineering Capability	Capability, organization, information, process, develop, model, performance, team, structure, environment	Establish, develop, and improve a systems engineering capability in different organizations and assessments through capability maturity models that consider information and infrastructure. This also includes capability engineering for systems.
9	Project management	Project, management, planning, approach, success, develop, implement, activity, team, engineering	Systems engineering projects, different project management approaches, and integrating systems engineering and project management to support each other.
10	Systems Engineering Practice	Engineering, team, practice, develop, effective, understand, need, problem, success, process	The profession of systems engineering in relation to other fields. The competencies, traits, and skills to practice as systems engineers in multidisciplinary teams.
11	Tools and Cost	Tool, cost, develop, performance, effective, engineering, need, process, design, methodology	Tools to support systems engineering, specifically, tools to estimate, model, and optimize system cost.

No	Topic name	Top topic terms	Description
12	Systems Engineering Management	Management, challenge, solution, need, complex, approach, practice, implement, effective, problem	Implementation systems engineering in organizations through various management approaches, systems thinking, enterprise engineering, change management, and knowledge management to solve problems.
13	Systems Engineering Processes	Process, standard, change, approach, implement, develop, need, activity, organization, industry	Systems engineering process implementation, standards, tailoring, development, management (improvement), and best practices.
14	Decision Support and Frameworks	Framework, decision, base, structure, approach, develop, process, complex, methodology, understand	Frameworks in systems engineering for decision making, trade-offs, validation and verification, competency, architecture, and systems engineering strategy.
15	System Life Cycle	Cycle, life, lifecycle, develop, process, engineering, environment, design, approach, activity	Systems engineering processes and other aspects throughout the life cycle stages.
16	Software Engineering	Software, technical, develop, performance, activity, application, implement, approach, environment, engineering	Development of software systems, implement software in systems and applications in support of systems engineering. It includes the technical aspect of systems such as integrity, quality, and social activity.
17	Complexity	Complex, technology, develop, environment, concept, application, challenge, change, approach, understand	Complex systems and systems of systems, along with the principles, types, and causes of complexity. Including methods to describe, analyze, and address complexity. Assessment and development of technology that contributes to complexity.
18	Requirements	Requirement, process, management, develop, need, engineering, tool, function, implement, problem	Requirement engineering and management, including process to capture and quality of all types of requirements.
19	Product development	Product, develop, process, approach, industry, need, methodology, practice, structure, success	Product and system development with the role of systems engineering processes. Including assessing and optimizing the process and product line.
20	Design	Design, concept, approach, performance, base, function, process, develop, methodology, engineering	Design of systems, including concepts, as part of systems engineering with design principles such as flexibility and reliability.

Figure 5: Sample of topic word clouds

Two methods were considered to allocate papers to the extracted topics. The first method assigns the paper only to the topic with the highest probability. Since systems engineering is a multidisciplinary field, there is a strong possibility that each article may address multiple topics. The second method does not allocate an article to a single topic but rather by the top five topics with the highest probability of allocation. This data is also available from the recorded LDA algorithm output. As seen in Figure 6, the average allocation probability of the first topic to a paper for the whole data set is 42 percent, 23 percent for the second topic, 14 percent for the third topic, 8 percent for the fourth topic, and 5 percent for the fifth topic. The average topic allocation per paper, using the percentages of only the top five topics, adds up to a 91 percent probability. Therefore, this allocation method should provide a more accurate indication of the popularity of topics instead of only the first topic, which turned out to be 42 percent. Figure 7 presents the histograms for the paper allocation percentages of the top five topics per paper.

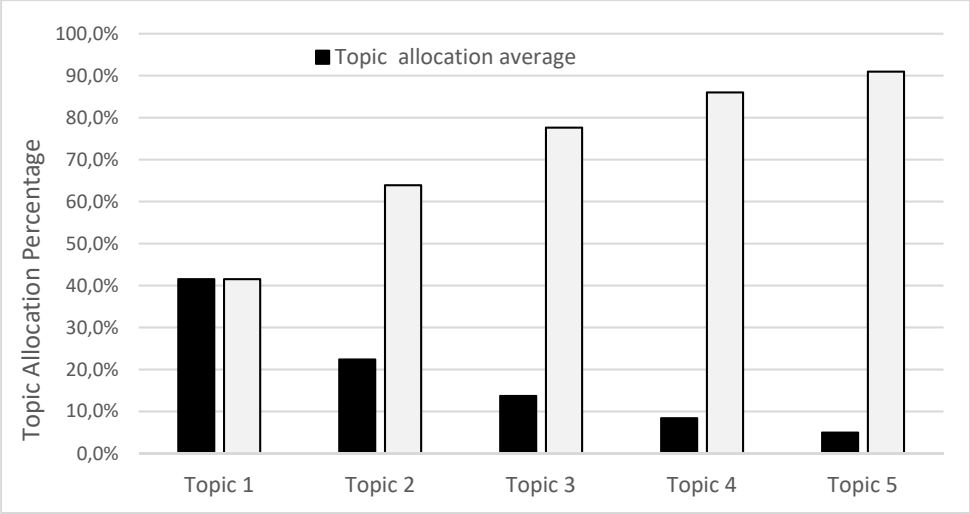

Figure 6: Average topic allocation probability per paper

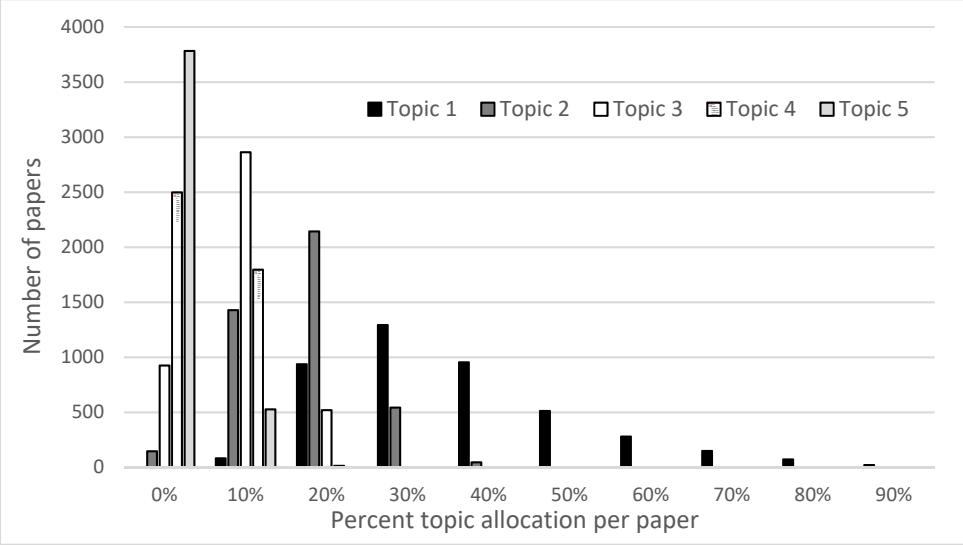

Figure 7: Allocation of top five topics per paper histogram

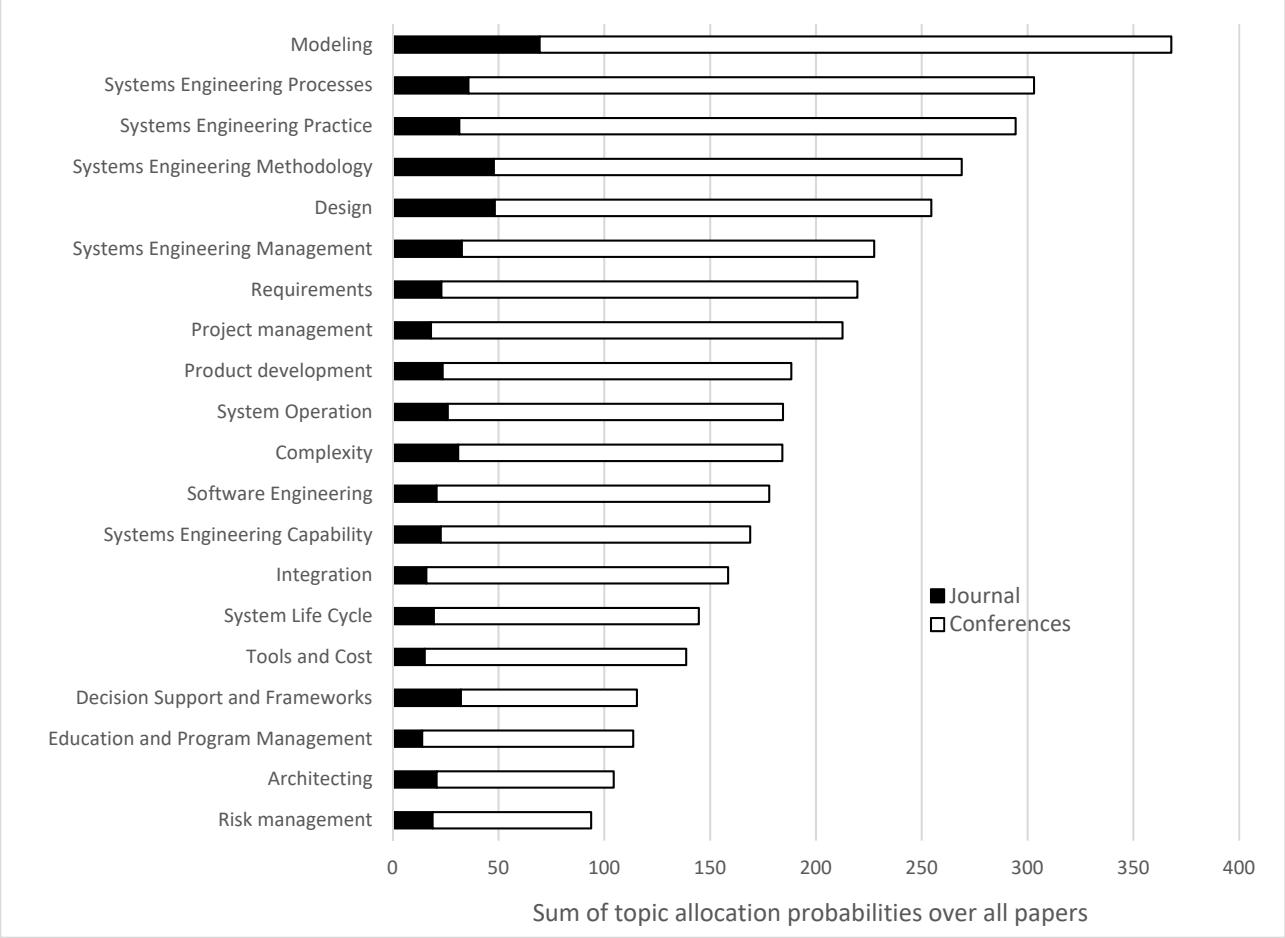

Figure 8: Topic Frequency of all the papers

Figure 8 shows the distribution of papers allocated to the topics in both the *Systems Engineering* journal and conference proceedings, as described in the previous paragraph. The graph also provides the contribution of the journal relative to the conference proceedings. There seems to be a difference in scope between the two sources

of data. Systems engineering processes and practices have a higher relative frequency at conferences, while methodological aspects and systems design are more prominent in the journal. This insight confirms that conferences provide practitioners with a platform to report on practical experience in applying and adapting systems engineering. The systems engineering journal tends to focus more research on methodological aspects of systems engineering.

Figure 8 also shows that modeling is the most popular topic of published systems engineering research in both the journal and conference proceedings. The role and importance of modeling in systems engineering are clear. The second most popular topic covers the processes and standards for systems engineering. Other popular topics are about systems engineering methodology and practice, as well as design.

5.4 Systems engineering research topic validation

The first two validation process steps, from section 4.2.2.3, have already been addressed in the motivation, execution, and description of the topic modeling process up to this stage. The remaining validation steps are now covered in this section.

5.4.1 Expert evaluation of extracted topics

The SMEs have also investigated the topic terms from the LDA algorithm, as part of the topic identification in section 5.3, to identify possible terms out of place. Viewing the word clouds and a list of the top 15 terms per topic, the SMEs could not find any term not relevant to the specific topic. However, this is understandable as the processed text corpus focuses on the field of systems engineering. The pre-processing and text cleaning process also removed most unnecessary words that could contribute to this problem.

Table 5: Topic allocation loading as validated manually

No	Topic	Allocation rating
1	Systems Engineering Processes	88.9%
2	Risk management	86.4%
3	Software Engineering	85.3%
4	Modeling	85.2%
5	Project management	84.8%
6	Systems Engineering Management	84.8%
7	Systems Engineering Practice	84.5%
8	Architecting	83.3%
9	Decision Support and Frameworks	83.3%
10	Systems Engineering Methodology	83.2%
11	Requirements	80.6%
12	Design	79.0%
13	Education and Program Management	78.3%
14	System Life Cycle	77.0%
15	Complexity	76.0%
16	Systems Engineering Capability	75.8%
17	Product development	75.3%
18	Tools and Cost	74.6%
19	Integration	73.8%
20	System Operation	68.4%

5.4.2 Topic allocation to papers

A random set of five documents (papers) per topic (assigned by the first topic) was identified as a subset (sample) of the corpus for evaluating the allocation of topics using a random number generator in Excel. The purpose of this evaluation is to detect any papers possibly misallocated to the topics manually. Each of the top five topics allocated to an article was evaluated to determine its relevance. The percentages of each of the five topics assigned to the paper, which are deemed relevant through visual inspection, were summed to derive a percentage value for the accuracy of that paper's topic allocation. The descriptions in Table 4 were used in the topic allocation evaluation to ensure consistency. The average allocation of the five papers per topic is shown in Table 5.

The average topic allocation accuracy over the entire sample is 80.4 percent, with a standard deviation of 12 percent. The maximum allocation percentage of a single paper is 98.5 percent, while the lowest value is 47 percent. Only four papers of the total sample (100) could not be reliably assigned to the proposed first topic, and 14 papers to the second topic. Statistically, the topic model allocates topics to each paper at an accuracy of 78.4 percent and a significance of $p < 0.05$ (T-score = 1.66 over the sample of 100 papers).

Figure 9 shows a histogram of the total allocation percentages per paper for the random sample. These numbers reflect a reasonably accurate allocation of the topics to the papers for this study. Reasons for misallocation included the low difference in the probability for the allocated topic and different meanings of the topic terms. Inconsistent use of keywords in engineering by authors may cause some inaccurate allocations; typical example terms encountered include framework, program, and methodology. However, due to the corpus size and method of topic allocation, these inaccuracies tend to diminish.

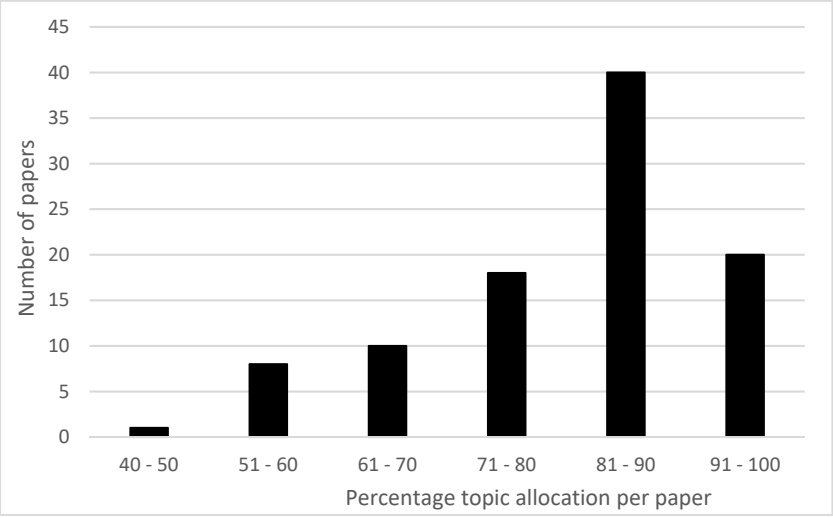

Figure 9: Topic allocation accuracy of a sample

5.4.3 Systems engineering research topic comparison

The final step of validation compares the output of the topic modeling process presented in this paper to systems engineering research topics proposed in the literature. A number of authors have proposed various lists, for different problems or contexts, since 2008 [2,35,37,38,39,40]. However, deriving a unified list of systems engineering research topics from literature is not easy. The topics identified in these papers have different contexts, objectives, structures, and levels of abstraction. These sources also tend to propose future research requirements as opposed to considering past trends. Regardless, the authors integrated these research topics into a list to compare the LDA extracted topics in this paper, as seen in Table 6. The descriptions from the systems engineering journal (INCOSE) and the international symposium proceedings website, referring to the topics covered, were also included.

In general, there seems to be a good fit between the topics from literature and the topic model; 19 of the 20 extracted topics can be related to the integrated set of topics proposed in the literature. Only the topic of Systems Engineering Practice is not so obvious to map the research topic list from literature. As seen from the topic description in Table 4Table , this topic covers a number of overlapping topics making a specific allocation difficult.

Table 6: Systems engineering research topic from literature

No	Topics from literature	Topics from LDA	Sahraoui (2008) [35]	Ferris (2009) [2]	Squires (2012) [37]	Cook (2013) [38]	Axelsson (2015) [39]	Verma (2018) [40]	Broniatowski (2018) [41]	Journal & Proceedings
1	Modeling and simulation (MBSE)	Modeling	X	X	X	X	X	X	X	X
2	Integration (communication)	Integration	X			X	X	X	X	X
3	Specialities (Sustainability, Safety, security)				X	X	X	X	X	X
4	Requirements management	Requirements	X		X	X	X			X
5	Architecture	Architecting	X	X	X	X	X		X	X
6	Systems engineering processes and standards	Systems engineering processes	X	X		X	X		X	X
7	Systems engineering tools	Tools and Cost	X	X			X			X
8	System development	Product development Software Engineering	X		X	X		X		X
9	Design	Design	X		X		X		X	X
10	Testing, verification, and validation		X		X	X		X		
11	Life cycle (cost and effectiveness)	System life cycle	X		X		X			X
12	Project management	Project management		X	X		X	X	X	
13	Risk management	Risk management		X	X		X	X	X	
14	Complex systems and systems of systems	Complexity		X	X			X	X	X
15	Organizational management	Systems engineering management	X		X	X		X	X	X
16	Decision making (trade-off)	Decision Support and Frameworks	X			X		X		
17	Humans in systems (sociotechnical issues)			X						X
18	System science and underlying principles	Systems engineering methodology		X		X			X	
19	Education	Education and Program Management		X		X			X	
20	Maturity models	Systems engineering capability			X	X				
21	Operational Concepts	System Operation	X			X		X		
22		Systems Engineering Practice								

This topic coverage may also be an example where topic modeling can extract underlying topics from the text corpus, which are not evident in manual processing. The two topics of Product Development (hardware) and Software Engineering are both about the development and implementation of systems. Therefore, these two topics can be allocated to the research literature proposed topic of System Development, representing a broad research area.

Some proposed topics from the literature search are also very particular, such as the systems engineering specialties, distinct life cycle stages (testing, verification, and validation), and sociotechnical issues. It would be possible to find these topics within modeled topics of systems engineering processes, -methodology, or –capability. However, if more topics were extracted with the automated process, these could also be allocated. As stated before, the downside of this increased number of topics is reducing the topic model's accuracy. The differences between the topic lists may also indicate areas in systems engineering that lack, for example, actual published research.

6 Discussion

The LDA based topic modelling process in this paper extracted a set of 20 topics from INCOSE's *Systems Engineering* journal and the proceedings of the annual international conferences. These topics were identified and validated through a number of steps derived from the literature and discussed in section 4.2.2.3. These topics also compare well with similar reviews in literature, as seen in Table 6. This paper's highlight is the in-depth validation process followed to analyze the topics derived from unsupervised machine learning. Topic modeling also provides an effective method to search a vast corpus of documents for information on a specific topic. Performing a simple search based on a selection of keywords may result in finding many irrelevant papers and not finding all the required documents in which the topic is hidden or implied in the text.

The relative popularity of the topics was also addressed. This paper presents a quantitative comparison between the research topics in systems engineering. The topic of "modeling" features prominently as part of systems engineering research, reaffirming its importance in the field. The other popular topics address the processes and standards, as well as the practice of systems engineering. These may indicate that the field is still growing as researchers continue to publish about conducting systems engineering. The topic of "Systems engineering methodology" is only ranked as the fourth most prominent topic. Here the methods and theories for systems engineering are covered for problem-solving, evaluation, and development. The remainder of the topics starts to cover systems engineering processes and life cycle stages.

The low popularity of some topics provides an opportunity for research projects. Topics such as "risk management," "architecting," "decision support," and "integration" are critical aspects of systems engineering. These topics deserve a larger footprint within systems engineering research. Another significant output of the topic modeling process is the missing topics. These were highlighted during the comparison with the research analysis of Table 4 and also present research opportunities.

Another concern is the seeming absence of common topics within the systems engineering fraternity, such as systems-of-systems and MBSE. This warrants a more in-depth analysis of the data. Firstly, MBSE is part of the topic of "modeling." When considering the input data, the specific phrases "MBSE," "model-based systems engineering," of "SysML" only occur in 40 of the total of 622 papers published in the journal. This is in contrast to the sum of 90 articles allocated to the topic of "modeling" by the first topic. The topic model algorithm extracted hidden and latent topics from documents independent of specific search terms. The same analysis can be performed on the symposium proceedings, where only 267 of the 3 695 symposium papers contain the same terms. However, the total number of papers extracted through topic modeling and allocated to "modeling" was 409.

The same analysis can be performed for systems-of-systems. As noted in section 4.2.2.1, the words "system of system" was replaced with the abbreviation "SOS" to protect it from the NLP. After processing the text in preparation for the topic modeling, the term "SOS" occurred only with a frequency of 591 times, in term frequency position 63, in the processed text for topic modeling. Therefore, the term did not make it into the extracted 54 terms for defining the topics in this analysis. Except for two terms, the topic terms all occurred in the top 54 words by

occurrence frequency. However, the topic modeling caused the terms' priority order to change to differ from the frequency order due to an occurrence pattern in the documents.

Therefore, the relative importance of the terms "MBSE" and" SOS" to other words in the text data is too low to generate unique topics. This may indicate a gap in systems engineering research as there seems to be a lack of papers focussed purely on these topics. As these are relatively recent concepts in systems engineering, a topic model of only the newer articles may, such as the last ten years, may result in a different set of topics with new relative importance values.

Unfortunately, despite the utility and value of the topic modeling approach, some deficiencies do exist. Firstly, this paper's topic modeling approach may also suffer from bias despite being unsupervised machine learning. Text cleaning, selection of parameters, and naming of the topics may introduce bias into the process. Another aspect that may cause some concern is the topic model's tendency to combine different concepts into one topic due to the text corpus' co-occurrence. One example is the dual use of the term "program" for education (training) and big government defense acquisition projects, as seen in Table 4. NLP and LDA cannot interpret the subtle contextual difference between these terms.

Another example is the combination of two topics is with the terms of tools and cost. Although some papers address both cost and tools, the topic may also be assigned to articles where cost is discussed with risk and system life cycle. The same applies to the topic of decision support and frameworks. Some papers discussing frameworks have little to do with decision support. However, these topics are also the least prominent and marginal topics to be assigned to the least number of papers, as seen in Figure 8. However, these discrepancies should not detract from the conclusions made in this paper about systems engineering research.

Regardless of the limitations of automated unsupervised machine learning-based topic modeling, it can still provide valuable support for systems engineering researchers in literature studies. An in-depth analysis of the output topics could also support developing a comprehensive list of research topics in systems engineering to update the list in Table 6Table 4. As seen from the statistical in section 5.4.2, this method provides an 80 percent accurate answer for extracting and allocating topics to a corpus of 4 316 papers within a fraction of the time used for performing manual analysis. A manual process will also suffer from bias and errors.

7 Conclusion

This paper's primary contribution is a set of validated research topics extracted from papers in INCOSE's *Systems Engineering* journal and the annual international symposium proceedings. A rigorous approach was followed to generate and validate the systems engineering research topics. The output data provide rich information for further analyses to investigate specific questions about systems engineering. This validated set of topics can help researchers understand the current and historic research state in systems engineering research.

The methodology of using NLP with topic modeling presented in this paper is a valuable tool to analyze systems engineering research. NLP and machine learning were applied to systems engineering research published in the *Systems Engineering* journal and INCOSE's international symposium proceedings to explore the research topics' trends. This provides valuable bibliometric information to investigate the development of the systems engineering discipline through published research. The topics extracted from the two sources compared well, with a clear overlap in the topics extracted from selected articles. Implementation of the unsupervised machine learning process may not produce a perfect set of topics during each probabilistic algorithm execution. However, selecting the proper parameters and using quality text data are significant factors in this process. The errors should also diminish over a large corpus when evaluating trends. The ability to provide interesting and even counter-intuitive results has been demonstrated in the case of MBSE.

The authors aim to continue this approach to investigate different aspects of systems engineering research. The data set generated in this research has many future opportunities to support developing a research agenda for systems engineering. The topic modeling may also be improved by focussing on different word types in the text sample. The topics may further be subjected to social network analysis, due to the co-occurrence of topics per paper, to investigate systems engineering research structure. The growth or decline of a topic over time needs investigation to extract research trends. The articles may even be grouped by publication date and processed separately.

Future research opportunities also include expanding the input literature to other publication platforms and journals, even those not focusing solely on systems engineering research. Topic modeling may be utilized as a smart filter to extract the papers relevant to systems engineering from a broad keyword search over multiple databases. The text corpus size can also be increased by including all the article's content instead of only the titles and abstracts. The topic modeling algorithm can also be adapted to be more sensitive to newer articles to identify recent trends. Lastly, this paper's unsupervised machine learning process may be combined with supervised approaches to improve topic extraction accuracy and individual paper allocation.

8 References

1. Walden D, Roedler G, Forsberg K, Hamelin R, Shortell T. *INCOSE Systems Engineering Handbook*. 4th ed. Hoboken: Wiley; 2015.
2. Ferris, T.L.J. "9.1. 3 Development of a Framework of Research Topics in Systems Engineering." *In INCOSE International Symposium*; 2009; 19(1)"1378-1390.
3. Brill J. Systems engineering? A retrospective view. *Systems Engineering*; 1998;1(4):258-266.
4. Valerdi R, Davidz HL. Empirical Research in Systems Engineering: Challenges and Opportunities of a New Frontier. *Systems Engineering*, 2009;12(2):169-181.
5. Sage AP. Systems engineering: Purpose, function, and structure. *Systems Engineering*; 199;81(1):1-3.
6. Kuhn TS. *The Structure of Scientific Revolutions*. The University of Chicago Press, United States of America; 1962.
7. Muller G. Systems Engineering Research Methods. *Procedia Computer Science*; 2013;16:1092-1101.
8. Ferris TLJ. On the Methods of Research for Systems Engineering. In *Proceedings of 7th Annual Conference on Systems Engineering Research*, Loughborough University; 2009:20-23.
9. Antons D, Kleer R, & Salge T.O. Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories. *Journal of Product Innovation Management*; 2016; 33(6): 726-749.
10. Eker S, Rovenskaya E, Langan S, Obersteiner M. Model validation: A bibliometric analysis of the literature. *Environmental Modelling & Software*; 2019;117:43-54.
11. Hood WW, Wilson CS. The literature of bibliometrics, scientometrics, and informetrics. Scientometrics; 2001;52(2):291-314.
12. Keathley-Herring H, Bean A, Chen T, Vila K, Ye K, Gonzalez-Aleu F. Bibliometric analysis of author collaboration in engineering management research, Proceedings of the *International Annual Conference on American Society for Engineering Management,* S. Long, E-H. Ng, and A. Squires eds; 2015.
13. Jiang H, Qiang M. Lin P. A topic modeling based bibliometric exploration of hydropower research. *Renewable and Sustainable Energy Reviews*; 2016;57:226-237.
14. Jie L, Xiaohong G, Shifei S. Jovanovic A. Bibliometric Mapping of "International Symposium on Safety Science and Technology (1998-2012)". *Procedia Engineering*; 2014;84:pp.70-79.
15. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*; 2017;11(4):959-975.
16. Kalantari A, Kamsin A, Kamaruddin HS, Ebrahim NA, Gani A, Ebrahimi A, Shamshirband S. A bibliometric approach to tracking big data research trends. *Journal of Big Data*; 2017;4(1):30.
17. Jia Y, Wang W, Liang J, Liu L, Chen Z, Zhang J, Chen T, Lei J. Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences. *Computer methods and programs in biomedicine*; 2018;166:19-32.
18. Lamba M, Madhusudhan M. Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. *Scientometrics*; 2019:1-29.
19. Lee H, Kang P. Identifying core topics in technology and innovation management studies: A topic model approach. *The Journal of Technology Transfer*; 2018;43(5):1291-1317.
20. Banu GR, Chitra VK. A Survey of Text Mining Concepts. *International Journal of Innovations in Engineering and Technology*; 2015;5(2).
21. Agrawal A, Fu W, Menzies T. What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*; 2018;(98):74-88.

22. Kunc M, Mortenson MJ, Vidgen R. A computational literature review of the field of System Dynamics from 1974 to 2017. *Journal of Simulation*; 2018;12(2):115-127.
23. Tong Z, Zhang H. A Text Mining Research Based on LDA Topic Modelling. *In International Conference on Computer Science, Engineering and Information Technology*; 2016:201-210.
24. Suominen A, Toivanen H. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*; 2016;67(10):2464-2476.
25. Hecking, T and Leydesdorff, L. Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared with semantic maps. *Research Evaluation*; 2019; 28(3): 263-272.
26. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, Pfetsch B et al. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology." *Communication Methods and Measures*; 2018; 12 (2-3): 93-118.
27. Asmussen, CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data; 2019; 6(1): 93.
28. Oosthuizen R, Pretorius L. Bibliometric Analysis of Technology Management Research Topic Trends. *International Association for Management of Technology (IAMOT)*; 2020.
29. Rose ME, Kitchin, JR. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. SoftwareX; 2019;10:100263.
30. Patel FN, Soni, NR. Text mining: A Brief survey. *International Journal of Advanced Computer Research*; 2012;2(4):243.
31. Lin JR, Hu ZZ, Zhang JP, Yu, FQ. A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM. *Computer-Aided Civil and Infrastructure Engineering*; 2016;31(1):18-33.
32. Blei D M, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research*; 2003; 3; 993-1022.
33. Hagen L. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?. *Information Processing & Management*; 2018; 54(6):1292-1307.
34. Bashri MF, Kusumaningrum R. Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization. *In 5th International Conference on Information and Communication Technology (ICoIC7)*; 2017 (pp. 1-5). IEEE.
35. Isoaho K, Gritsenko D, Mäkelä E. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*; 2019.
36. Sahraoui A, Buede DM, Sage AP. Systems engineering research. *Journal of Systems Science and Systems Engineering*; 2008; 17(3): 319-333.
37. Squires A, Olwell D, Roedler G, Ekstrom JJ. Gaps in the body of knowledge of systems engineering. *In INCOSE International Symposium*; 2012; 22(1): 1967-1976.
38. Cook SC, Ferris TLJ, Nowakowski S. A Framework for Visualizing Systems Engineering Research Coverage. *In INCOSE International Symposium*; 2013; 23(1): 933-945.
39. Axelsson J. A systematic mapping of the research literature on system-of-systems engineering." *In 2015 10th System of Systems Engineering Conference (SoSE); IEEE*; 2015: 18-23.
40. Verma D, Miller W. Systems Sciences and Systems Engineering Research Needs." *INSIGHT 21*; 2018; 1: 48-51.
41. Broniatowski DA. Building the tower without climbing it: Progress in engineering systems. *Systems Engineering*; 2018; 21(3):259-281.