Proceedings of the ASME 2020 International Manufacturing Science and Engineering Conference, MSEC 2020

June 22-26, 2020, Cincinnati, OH, USA

MSEC2020-12885

MSEC: A QUANTITATIVE RETROSPECTIVE

Thurston Sexton
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Michael P Brundage[*]
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Alden Dima
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Michael Sharp
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

ABSTRACT

The ASME 2020 Manufacturing Science and Engineering Conference (MSEC) is the 15th annual meeting organized by the Manufacturing Engineering Division (MED) of ASME. MED and ASME MSEC focuses on manufacturing sciences, technology, and applications, including machining, materials processing, sensing, robotics, manufacturing system dynamics, and production optimization. As the conference has grown and evolved from its inception, it can be difficult to intuitively visualize and discuss the broad range of research topics covered by the MSEC community or to intuitively ascertain their evolution throughout time. This paper discusses a methodology to quantitatively model research communities within bodies of literature—specifically, the relative change of relevant topics within AMSE MSEC conference papers through time, from 2006 through 2018. The goal of this work is to not only present how research in MSEC has shifted over time, but in a broader sense to provide a discussion on how others can interpret results so that similar analysis can be produced within other research communities. This methodology can be used to identify overlap of communities, monitor growth or stagnation within the communities, to aid in developing new symposiums and communities of interest, or even to dictate future standards needs by looking at research trends and subsequent standard development.

1 INTRODUCTION

In 2020, the Manufacturing Engineering Division (MED) in ASME (The American Society of Mechanical Engineers) will observe its 100 year anniversary [1]. MED focuses on "the knowledge base of manufacturing sciences and technology and its applications for improved production performance that is economically viable and meets industrial health, safety, and resource conservation legislation" [2]. In conjunction with its 100 year anniversary, MED will hold the 15th annual Manufacturing Science and Engineering Conference (MSEC) in Cinncinati, OH. Traditionally, MED and MSEC have covered a wide array of manufacturing topics, including: machine tools, materials processing, sensors and controllers, computer integrated manufacturing and robotics, manufacturing systems management and optimization, and emerging areas of manufacturing engineering [2].

As published research in manufacturing evolves, a method is needed to quantitatively evaluate the research topics within communities and interpret their evolution over time. This paper uses established statistical modeling and natural language processing (NLP) techniques, such as topic modeling and document cluster tracking, with papers from ASME MSEC on 1) the topics of discussion within ASME MSEC, and 2) how related research efforts have evolved over time within those topics. The goal of this paper is to not only discuss interesting trends in ASME MSEC, but more importantly to provide a detailed process for obtaining and interpreting these results. The larger goal is to ensure that the developed procedure can be replicated for other research collec-

[*] Corresponding Author: mpb1@nist.gov

tions, providing needed insights for other communities and domains. This quantitative method for evaluating and interpreting the evolution of research communities and their primary topics of interest can be used to identify success stories in areas of rapid growth or movement, new and developing thrusts that may be worthy of additional attention, stagnant topics in need of revitalization, as well as predicting future needs and next steps. Information gained through these types of analysis within a specific domain could even be used to help describe future conference tracks or symposiums, highlight standards needs, and provide justification for potential future research funding areas.

The rest of the paper is structured as follows. Section 2 discusses background on topic modeling, document cluster tracking, and other literature discussing research paper trends. Section 3 illustrates the methodology used to create the topics and document shift results. Lastly, conclusions and future work are presented in Section 4.

2 PROBLEM CONTEXT

With the rapidly expanding volume of available documents in a given research area, it is nearly impossible for a single person to manually survey and synthesize a full community of research. Evolution of language usage as well as changes in interest and focus through time can make identifying and evaluating groups of common research difficult. This paper presents a method to quantitatively assess and follow groups of semantically similar documents to create a timeline of research thrusts that identifies progressive changes in interest, and the relative interplay between thrust areas. Additionally, results are presented to gain a more global view of the intrinsic topic areas that arise through time and how different ideas will come in and out of vogue across the corpus of documents.

2.1 Data Acquisition

The quality of any analysis is directly related to the quality of the data used in the analysis, in this case the presented data relates a sampling of the available MSEC documents from 2006 to 2018. Information contained in the ASME Digital Collection web pages was used to identify and download MSEC conference papers as individual PDF files using the workflow created by the authors, shown in Fig.1. This step produced the data set used in our analysis by extracting and processing their text from the pdfs. However, due to a lack of a standardized storage and access format, the number of papers available for this research between 2006 and 2012 is significantly lower that the full number of papers identified by citation, as shown in Fig. 2. The work presented in Section 3 still includes these years for completeness and for demonstration of the method, but any conclusions drawn from this region must be regarded as highly uncertain due to the low number of full text papers that could be obtained for this work.

The top-level conference-related pages list conference papers by track and symposium. Starting from a list of manually collected top-level URLs, the data gathering workflow used a combination of common Unix command-line tools and established Python libraries to download the top-level conference papers (Fig. 1: Web Page Downloader). The downloaded pages were processed to extract the article metadata directly from the page HTML (Fig. 1: Metadata Extraction) including: the conference name, the PDF URL, and the article ID, DOI, Date, and URL. Each article's track and symposium was then extracted from the structural information of the conference proceedings webpage (Fig. 1: Track & Symposium Extraction). Using the extracted data, we identified a total of 2267 articles and were able to download 1457 MSEC-related PDFs from the ASME Digital Collection (see Fig. 2 and Fig. 1: PDF Downloader). We extracted their text, including titles, table and figure captions, and metadata, via the pdftotext command-line tool (Fig. 1: Text Extraction).

Though there are software tools that can be used to automate portions of the data acquisition, the general form of the task of obtaining and collating a collection of documents will still require human supervision and intervention at each stage. Various factors like the form and location of the publications, possibly Web site structure, document layout, and file naming conventions all affect the ability to automate the discovery and extraction of articles. There is also strong possibilities of desirable tacit information being encoded that is incredibly difficult for an automated algorithm to address without special considerations being made.

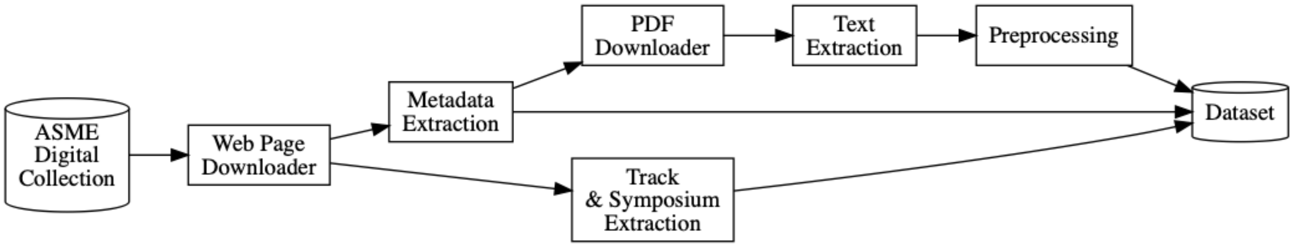

FIGURE 1. *Workflow used to acquire and process ASME MSEC articles.*

During our study, important information about the relationship between articles and symposia were encoded structurally in the Web pages and required human inspection and analysis to write the custom code for its extraction. Though each instance of acquiring data to process will have unique eccentricities, the most important aspects of collection for our pipeline are to ensure that each document with computer interpretable text is accessible, and their associated publication date are linked to them. Any additional meta data about locations, tracts, authorship, etc. can be useful for additional analyses, or may facilitate the data acquisition, but are not critically needed for the described analysis.

2.2 Data Preparation

Even in digital form, text is not amenable to direct analysis, especially by mathematical techniques based on linear algebra, which are typical of many NLP algorithms. Preprocessing is necessary to convert the text into a more suitable representation. We will discuss text preprocessing in the following two sections. Section 2.2.1 will give an overview of the different facets of text preprocessing. Section 2.2.2 will describe how we prepared our text for the analyses described in Section 3.

2.2.1 Overview of Text Preprocessing

Text is often preprocessed before analysis, with one goal being to simplify the data for analysis without losing necessary information [3]. This simplification can be thought of as sharpening the desired "signal" while reducing "noise". Another simultaneous goal is to structure the data in a manner that facilitates the analysis [4]. This restructuring is often designed to emphasize certain features that are important to the analysis and algorithms at the expense of others that may be less important. Examples of typical preprocessing beneficial to these types of analysis include:

Extraneous Metadata Removal: The removal of unwanted template text [5] that add "noise" to analyses. For example, each page contained text such as:

> Downloaded from https://asmedigitalcollection.asme.org /MSEC/proceedings-pdf/MSEC2008/48517/1/2715771 /1_1.pdf by NIST user on 18 September 2019

This unwanted text can be identified for removal using technologies such as regular expressions, a mathematical language for describing text patterns used by text processing software to identify specific text [6].

Stop Word Removal: The elimination of many common words such as *this, that,* and *the* that convey little semantic meaning to an analysis [7]. Preassembled *Stop Lists* can be used to identify and remove them from text prior to analysis [7, 8].

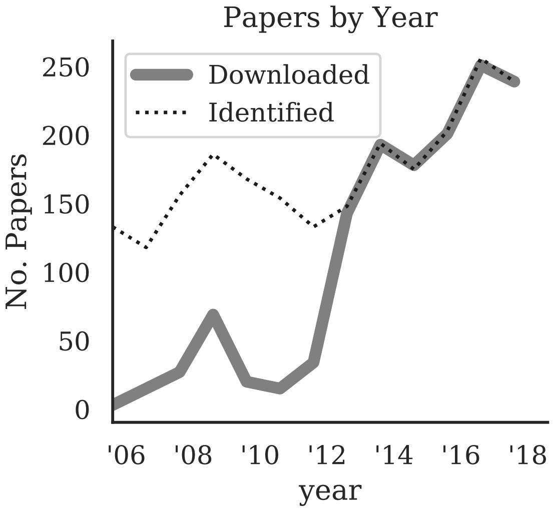

FIGURE 2. *Papers identified and downloaded from the ASME Digital Collection by year using our workflow*

Lemmatization: The reduction of inflected words via linguistic techniques into their dictionary forms [9], for example:

$$\{\text{boat, boats, boat's, boats'}\} \Rightarrow \text{boat} \quad (1)$$

By mapping all of the variants of a word into its base form, lemmatization can be seen as a way to increase the "signal" associated with that word.

Cleaning: The removal of characters and tokens, such as punctuation, numbers, and email addresses as well as low-frequency words, from the text to reduce noise features in the data that can negatively impact analysis [3, 9]

Tokenization: The process of separating text into its smallest meaningful units such as words and numbers [4, 7]. Because it identifies individual words, tokenization is a fundamental process that serves other downstream preprocessing and analysis tasks.

Sentence Segmentation: The identification of sentence boundaries in a text [4, 8]. Sentence segmentation plays a key role in supporting lemmatization by providing the input for the linguistic analyses required for lemmatization.

Bag of Words (BOW): A method for representing the contents of documents by identifying their words and counts without any ordering information [9]. For example:

$$\text{“Jack has a blue hat and Jill has a red hat”} \Rightarrow \\ \{(\text{“Jack”},1),(\text{“Jill”},1),(\text{“a”},2),(\text{“and”},1), \\ (\text{“blue”},1),(\text{“has”},2),(\text{“hat”},2),(\text{“red”},1)\} \quad (2)$$

Vectorization: In the presence of a dictionary to assign unique integers to each word, BOW can also be represented as vectors. For example, assuming that the previously used sentence was the first sentence in a text collection then the dictionary would contain the following information:

$$\text{dict} = \{(0,\text{“Jack”}),(1,\text{“has”}),(2,\text{“a”}),(3,\text{“blue”}), \\ (4,\text{“hat”}),(5,\text{“and”}),(6,\text{“Jill”}),(7,\text{“red”})\} \quad (3)$$

The integers serve as word ids and as subscripts to the vector representations for the BOW. The first example now becomes:

$$\text{“Jack has a blue hat and Jill has a red hat”} \Rightarrow \\ (1,2,2,1,2,1,1,1,0,\ldots) \quad (4)$$

where $0,\ldots$ represents positions associated with other words in the text collection not found in the first sentence. In practice, these vectors are sparse; for a typical text collection, zeros account for about 98% of the vector elements [10].

Considerations Text preprocessing is a form of analysis on its own, where typically the actual analysis is hidden from the user and performed in preexisting libraries and frameworks. For example, lemmatization requires the identification of the parts of speech that also depend on the identification of individual sentences. These operations are implemented by the NLP framework (Section 2.2.2) and involve commonly available statistical and neural-network language models whose full descriptions are outside the scope of this paper. However, interested readers can refer to books such as Krohn et al [11] and Rao and McMahon [12] and articles such as those by Young et al [13] and Akbik et al [14]. The ultimate text analysis performed thus rests upon the these hidden analyses.

While it is convenient to visualize text preprocessing as a pipeline in which raw text flows in and is replaced with refined text for analysis, it is better to initially consider it as a series of stages in an converging iterative process. Ideally, the entire output of each stage should be available for inspection and analysis; “sanity checks” are crucial. The pipeline perspective makes the most sense at the end when all of the issues discovered during development have been resolved.

2.2.2 Text Preprocessing Used

Our final data set contained the following preprocessing steps. First, filenames were standardized to contain both the year and a unique paper ID for easy of storage, indexing, and retrieval. We then filtered the extracted text from each document using Python’s regular expression library to remove headers and extraneous metadata that can negatively impact the subsequent text analysis. For example, each PDF file contained text identifying its source URL, the account that downloaded it, and the date and time that it was downloaded; none of which is relavant to the subject of the document. The text was then segmented into individual sentences and tokenized words. *Stop words* were discarded and the remaining tokens were lemmatized using established procedures found in `spaCy`[1]. Special word/number occurrences were replaced with unique text-only aliases prior to the removal numeric entities. For example:“1D”, “2D”, and “3D” were replaced with “oned”, “twod”, and “threed”, respectively. The text was then scrubbed using the `clean-text` library[2] to reduce it to lower case and to remove URLs, email addresses, phone numbers, currency symbols, punctuation, and numbers. At the end of this phase of data preparation, each downloaded PDF document had a corresponding text file that contained a single processed sentence per line.

We then used gensim[3] to build a look up table style dictionary that relates individual tokens to a unique numerical identifier and converted each text file into a BOW represented by gensim’s sparse vector format [15]. These processes have allowed the data to now be ingested and analyzed by modern mathematical analysis techniques.

The presence of semantically interesting non-word tokens (such as “3D” and its replacement with “threed”) highlights that text preprocessing is transformative in nature. While it may be intuitive that some text should be removed as being not relevant, it may be less obvious that text analysis relies on the translation of the original text, for example the use of lemmatization, substitutions, and bags of words. Text preprocessing should not be seen as the rote application of recipes to the original text. Instead analysts must inspect the intermediate results while considering their ultimate goals. For a discussion of how preprocessing can alter analysis, see Denny and Spirling [3].

3 ANALYSIS WORKFLOW & RESULTS

To streamline analysis and interpretation of collections of research documents, this paper presents a workflow for deriving useful insights from raw text, not only from the collections gathered at MSEC, but in other domains as well. Many options exist for the mechanical implementation of each of the steps in the process, but the choice of specific algorithms or combinations of algorithms has the potential to greatly influence the quality and

[1] `https://spacy.io/`

[2] `https://github.com/jfilter/clean-text`

[3] https://radimrehurek.com/gensim/

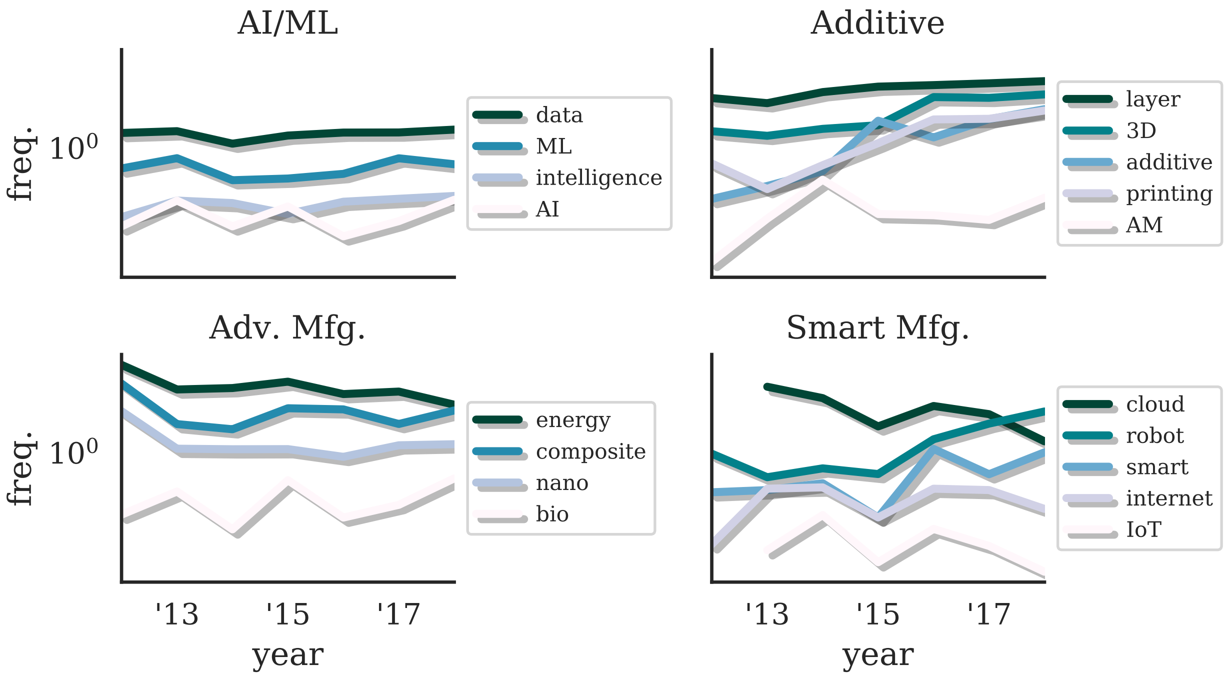

FIGURE 3. Keyword occurrences per document (log-scale) over time, for documents submitted after 2012. Each figure shows the top most-frequent terms occurring within the author-determined research areas. Darker color represents higher all-time occurrence.

outcome of the end analysis. This section discusses the benefits and difficulties of choices made at each step of the process to develop a guideline and philosophical approach for readers wishing to apply the procedure to better understand their own domain.

3.1 Keyword Trends

One *direct* way to access trends in a domain over time is to observe how keyword usage evolves within the documents that make up the domain. Given a frame of reference, such as a defined area of interest (*e.g.* , guiding documents from steering committees) we can categorize keywords by their relevant topic and observe the relative frequency of these keywords within.

For example, ASME defined several key research areas[4] that may define the coming decade in engineering research. Several of these are highly relevant to manufacturing specifically. By selecting keywords typically associated with each area, whether from anecdotal experience or expertise, we can observe high-level trends in the focus each area has over time.

Figure 3 illustrates an example selection of these key research areas identified from ASME with several terms the authors deem as relevant to those areas. Each term is reported for the years in which a significant fraction of total papers were recovered to avoid bias. Importance is compared through total expected occurrences *per paper*.

Note that this analysis only shows general trends over time– whether a word is used more or less frequently at a global scale. In *AI/ML* we see a slight increase in "data" over time with "Machine Learning" displaying an earlier surge than "Artificial Intelligence", which surged recently after a slight down-turn in use in 2015-2016. *Additive* sees a steady increase across all major keywords, likely due to an overall increase of total interest (paper submissions) in the topic. Although a fairly broad topic scope, *Advanced Manufacturing* is seeing continued research in nano-scale and composites fabrication. A long-term decline in "energy" interest is being met with a rise in "bio-manufacturing" interest. Finally, *Smart Manufacturing* is showing a marked increase in "robot" use, while words like "Internet of Things" and "cloud" seem to be declining from a peak near 2014.

While this may present an excellent first-look into the trends within a community-of-interest, the actual "dynamics" here can be difficult to parse. Despite the apparent correlation between certain keywords, the same could be said about terms across topics. This is problematic if one wishes to make quantitative generalizations about topics overall, or even how well individual terms reflect the state of a topic. Do these "areas of interest" constitute an efficient snapshot of the community as a whole, and the trends hidden within?

[4] https://www.asme.org/wwwasmeorg/media/resourcefiles/aboutasme/asme_strategy.pdf

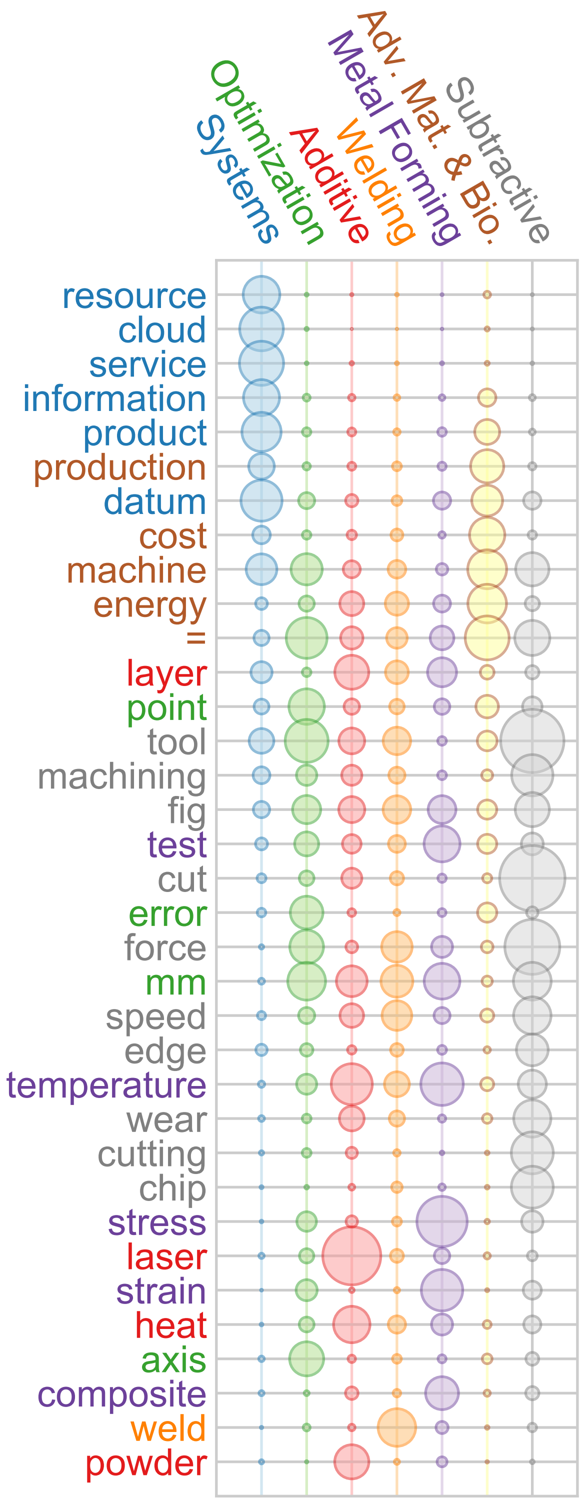

FIGURE 4. Term probabilities within each of 7 topics, as named by the authors. Showing top 35 terms having the highest probability in any topic. Rows are sorted to group terms occurring within similar contexts together. [16, 17] For instance, the top rows are all tightly related to the "Systems" topic, though they are increasingly linked to the "Advanced Materials and Bio-manufacturing" topic.

3.2 Topic Model

One mechanism to answer questions such as "what topics are contained within my documents," or, "what terms are most associated with a topic" is to infer patterns from the computational distribution of the documents. Doing this without directly labeling or classifying the documents beforehand constitutes *unsupervised learning* and is called topic modeling.

Topic modeling is a group of unsupervised natural language processing methods that can identify topics in large text collections [18, 19]. These topics are weighted sets of words that imply the overall semantics of the collection and its documents. The words weights in a topic tend to represent important concepts for that topic. The overall set of topics computed for a collection of documents can serve as an interpretable decomposition that summarizes its overall content.

Latent Dirichlet Analysis (LDA) is a topic modeling approach that uses a Bag of Words representation of each document (Section 2.2.1) to infer topics from a corpus. Being a generative technique, it assumes that documents are generated from two distributions: a per-document topic distribution that assigns topic weights to documents and a per-topic word distribution that assigns word weights to topics [15]. The goal of LDA then is to recover these distributions from the data in the form of the most likely allocation of term probabilities into discrete distributions of terms, called *topics*. So, if we observe terms as collections of words within documents, using an LDA model directly assumes that each observed document is generated from some set of topics each of which have a likelihood of "emitting" any of the terms in our corpus. Estimating those probabilities and the topic mixtures that make up each document involves training on a Bag of Words representation of each document in the entire corpus simultaneously. Note that the number of topics is an important parameter that must be passed before training has begun. It determines the number of possible "types of things" any given paper can draw upon, but allows the algorithm to determine an optimal distribution of term probabilities among them.

While it is theoretically possible to determine an "optimal" number of topics given a measure of topic quality such as coherence [20], in practice this parameter should be selected carefully to aid in communication and decision making. In the case of this analysis, AMSE MSEC accepts approximately seven submission tracks per year. Perhaps as a result of this, seven topics corresponding to a seven-dimensional space was found to result in more stable results for later analytics (see below) than all other parameter values tested by the authors ($5 \leq n \leq 15$). Similarly, as LDA merely defines n distributions over terms, it is necessary for *the analyst* to interpret each distribution of terms, estimating whatever latent groupings are being detected – i.e., to *name* the topics. This is a highly subjective process, even commonly called *"reading tea leaves"* in natural language processing [21]. Naming each topic, therefore, should constitute an iterative process of design, preferably among multiple stakeholders that carefully balances interpretations with defensible data-driven justifications.

Figure 4 illustrates how a topic model allocates term probabilities from the conference to each of seven latent topics. This is trained across the entirety of the corpus (all the papers from Fig. 2). In this plot, called a "termite" visualization[5] [16], rows correspond to terms and columns to topics that have been named through an iterative process by the authors. The probability that a term is generated by a topic corresponds to the size of each circle. Rows have been sorted by the method of *spectral seriation*, such that similar rows are grouped together as much as possible. [17]

This analysis provides some interesting insights into the topics of discussion at ASME MSEC. For example, the majority of submissions to MSEC have been historically related to manufacturing *process* research. This trend can be seen in the four topics: 1) *Subtractive*, 2) *Metal Forming*, 3) *Welding*, and more recently 4) *Additive Manufacturing*. Manufacturing Systems research is also a relevant topic as shown in the *Systems* topic. The *Advanced Materials and Bio-Manufacturing* topic captures another topic of research that has recently become prevalent at the conference. Lastly, the *Optimization* topic is often cross cutting, as deals with research relevant in each of the other topics as well. This type of trend can be seen in Fig. 4 as the *Optimization* terms of importance also appear in other topics (e.g., tool is more prevalent in *Subtractive*, despite being a core term in *Optimization*).

Figure 4 provides some other interesting results allowing experts to see terms relevant across multiple topics versus terms relevant only in one topic. For example, "cloud" is highly prevalent in *Systems*, but no where else, while "=" and "mm" are cross cutting in all topics except *Systems*[6]. Terms such as "cut" or "cutting" or "chip" are most important in *Subtractive*, while "temperature" is important to both *Additive* and *Metal Forming*. While these types of insights might seem obvious to some experts, this analysis and subsequent visualization provides a quantitative look at the conference and can be used to confirm or reject inferred perspectives to further more informed discussions. It provides a quick overview of the entire conference and the topics and related terms of discussion.

Although Figure 4 only tells a *static* story, we would also like to quantify the dynamic importance of terms, similar to what was done in our keyword frequency analysis (Fig. 3) while incorporating knowledge found within a topic modeling framework. For example, "cloud" was not discussed in retrieved documents prior to 2013, as shown in Fig. 3, and yet is a major term within the *Systems* topic. Surely that topic of interest had related terms that organized around the same themes prior to the introduction of cloud computing. As such, it would be helpful to see the evolution of each *topic* over time.

3.3 Topic Term Evolution

The naïve approach to temporal topic modelling would be to partition a corpus by date before creating separate LDA models for each piece and analyzing them each individually. Unfortunately, this drastically reduces the amount of data available to train each LDA instance. In addition, the process of "reading tea leaves" implies a complete lack of consistency from year-to-year as to which topics correspond to which preceding or following year's topics, let alone global-level topics as in Fig. 4.

Instead of creating topic models for each individual year, we can instead directly model the term evolution in each topic over time. This is called Dynamic Topic Modeling, as proposed by Blei et al.[7] [22]. Analyzing terms in each topic provides the necessary context that is missing in the analysis for Fig. 3, while enabling analytics and decisions based on trends over time — a dimension missing from Fig. 4. This topic term evolution analysis can help discover the most important terms in each topic *as they age*: how they ebb and flow over time.

As opposed to the general term trend analysis in Fig. 3, Figure 5 gives insights into term usage of importance within each topic individually. For instance, an expert can identify the trends from individual terms within a topic to better discover research trends. In the topic referred to as *Advanced Materials & Biomanufacturing*, a large overarching trend in consistent use of the word "energy" in the years 2011-2016 likely reflects strong incentive in the community to investigate energy consumption and sustainability. As of 2018, the topic is now dominated by composites and fiberous materials. This indicates what would otherwise be hidden dynamics between the two, having been aligned with the same topic.

Similarly, term coupling can provide useful insights. Using "cloud" as a continuing example, the same peak occurs in 2013 as in Fig. 3, but now it is possible to analyze the other important terms that are tightly coupled within that same topic, i.e., namely, "service" and "resource". These appear to rise and fall together. However, other terms are becoming more indicative of this topic in the past few years, with "datum" and "robot" growing significantly of late.

Finally, identifying how terms are distributed across topics over time can, for example, indicate the context where a technology is most "in vogue" at a given time. "Laser" is a key defining term in the *Additive* topic until 2014-15 where laser-based *Welding* appears to be not only important but dominant in that topic. Instead, keywords like "3D printing" are more indicative of *Additive* research with lasers serving as an enabling technology,

[5] Seriation as suggested by Chuang [16] achieved via spectral method proposed by Fogel [17], as implemented in the package `Textacy` (`https://github.com/chartbeat-labs/textacy`).

[6] The term "=" was left in the analysis to provide an approximation of the importance of equations in papers, while "mm" remains to illustrate the importance of dimensions/measurements.

[7] This is accomplished by assuming that the topic distributions for documents are sampled directly from the parameters of corresponding topic distributions in the previous time slice, thus achieving smooth, contiguous topic evolution.

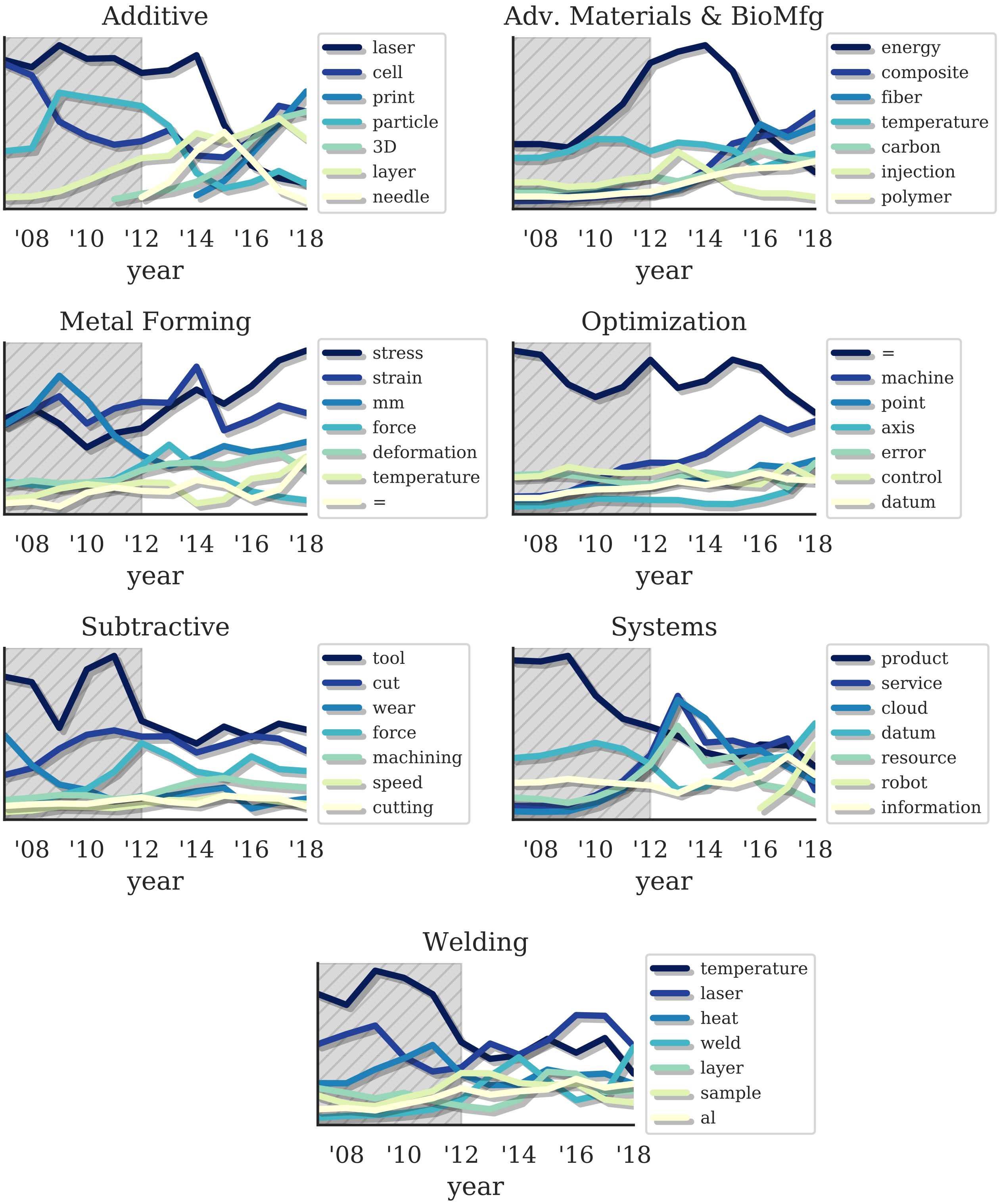

FIGURE 5. Dynamic term *importance*, i.e., relative term probabilities within dynamic topics, normalized by total topic occurrence in each year. Darker color means higher peak importance since 2012. Shaded region indicates years 2012 and before, for which downloaded document collection was incomplete.

These patterns let us quantify the degree to which specific topics of interest shift from one key idea to the next while maintaining an underlying connection and a more consistent relationship with the rest of the topics as a whole.

While this analysis provides a glimpse at how relevant research areas evolve over time, it is important to note that it says very little about how groups of individuals and/or papers *coalesce* around these areas through time. "Cloud" may be falling out-of-vogue within the topic overall, but ostensibly the community of researchers contributing to that body of work moved on to other relevant terms. Although terms like "robot" and "datum" are growing within their respective topic (*Systems*), this does not tell us whether they arising from the same community of interest; what are the original "cloud" researchers presenting on now? The next subsection presents a method to utilize topic modeling in conjunction with particle swarm tracking, to find and track clusters of actual documents–as opposed to topic mixtures – as they occur and migrate within the space of topics. This analysis has potential to provide a glimpse into the mixture of topics that individual communities of researchers are discussing, and how this mixture evolves over time.

3.4 Document Cluster Evolution

The idea of particle swarm tracking is not a novel one. The basic concept is to define the center of a group of similar entities (or entities that share a local region of some data space) as a swarm and track its movement through time. The number of particles that make up a group may shift over time, but so long as the local region defined by the swarm maintains a defined region and sensible movement criteria, then this local cluster can be monitored as a single entity.

As applied to this work, once a vector representation of a series of documents—i.e., conference papers—has been made, those documents can be thought to exist together somewhere in this semantic space. Using these locations, collections of proximity- or density-based clusters can be established that each contain a minimum number of documents. Labeling this group of documents as a single entity with volume and location, any number of particle tracking algorithms may then be applied to derive relative movement and interactivity of naturally occurring research efforts within the semantic document space.

In this work, a hierarchical density-based cluster approach was chosen to create the research document group instances. The specific tool used in this work is HDBSCAN from Rahman et al. [23]. This tool allowed for natural development of differing numbers of document groups, intuitively leading to new research thrust arrivals or merging or splitting of existing thrusts through time. Other clustering methods such as k-means, which specify a required number of clusters, makes such dynamic generation of topics less intuitive and forces prior assumptions upon the data.

The final step in the document group tracking is to connect the document groups to corresponding groups through time, thus making temporal traces of research thrusts. While there are many methods to accomplish this connection, for simplicity the intuitive method of requiring significant overlap across time of the document groups was selected. This method required the minimum number of assumptions and allowed for the method to be generalized to any selected characterization of the document-group space. In this work, the groups were represented as Gaussian-based N-dimensional fuzzy hyper-cubes. This allowed for rapid calculation of overlap in infinite space. With this, groups of different years can be checked for sufficient enough overlap between them to justify inclusion as a temporal trace. Although overlap was chosen as the connectivity metric here, any metric which captures significant semantic commonality between the clusters can be used to connect a cohesive trace through time.

Building each trace was performed as a single directional pass starting at 2018 and creating or extending traces one year at a time towards the beginning of the document. This single backwards stepping approach ensures that each termination point is unique for every particle trace. At each step backwards through time, the earliest point of each existing trace checks for the the document group with the most overlap (if any exist) to add as the new earliest entry. After all traces have claimed groups, any unclaimed particles are then defined as the terminal point of a new trace. Note that a single document group could be claimed by multiple traces as during their initial creation (while moving backwards in time). Subsequently, this will be shown as a split in research thrusts after construction when interpreting them forwards through time. This comparatively simple process of extending traces back through time accounts for expected behaviors of research efforts to die, merge, split, or even jump years if, for example, some new technology revitalizes some dormant research topics from the past.

There are other possibilities for creating particle traces of document clusters that may provide comparable results. The selections for this process were made to minimize the number of a priori assumptions and complex calculations to allow this process to be easily extended to higher-volume processing.

As presented in Figure 6, a total of ten different research thrusts were discovered and trended as document swarms through time. Labeled generically A through J, Fig. 6 shows that while each of the major topic areas (the colored ribbons) have one research thrust that mostly centered in it over time, there also exist several thrusts that cross cut these seven topics. Also, the focus of some of the long running thrusts seems to shift over time.

Trace A captures the primary research thrust concerning *Advanced Materials and Bio-Manufacturing*. We can see that there is a constant contribution of optimization to this research thrust, which is intuitive given that part of advanced materials research is to optimize some need via the materials or to opti-

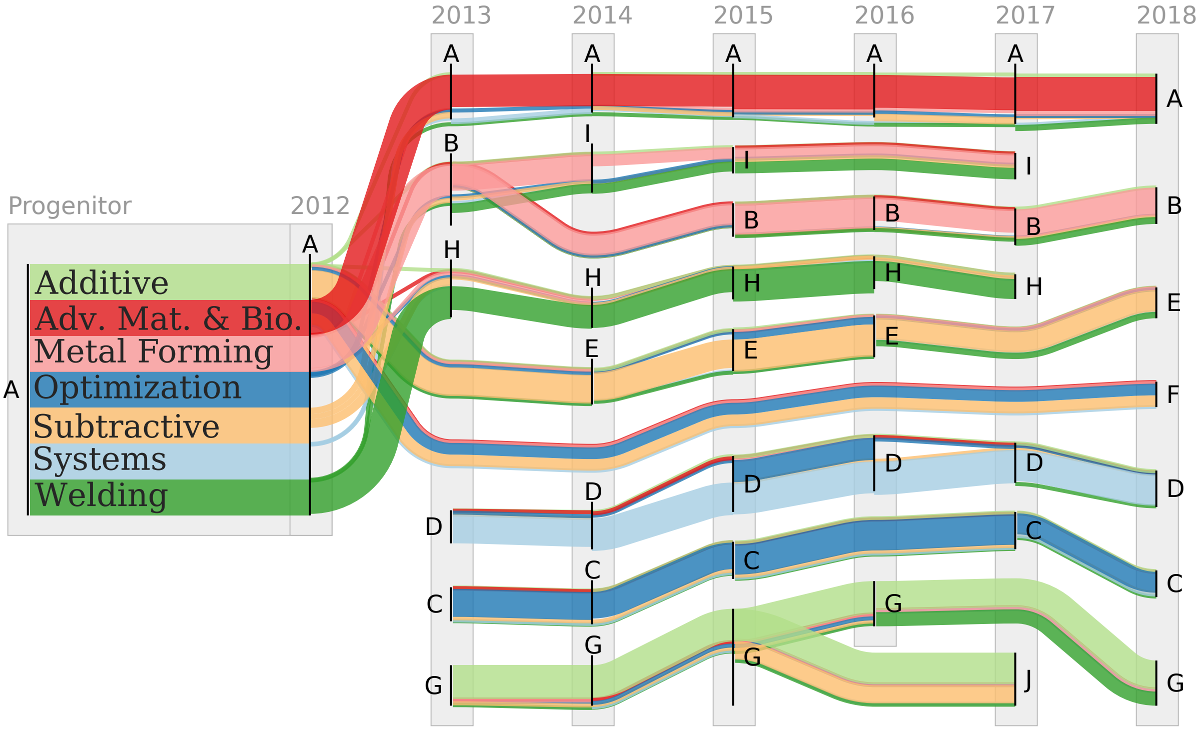

FIGURE 6. Sankey Diagram of document cluster alignment with LDA topics. Color indicates topic, while band thickness indicates proportion of that topic's global occurrence happening solely within each cluster. A trace is a presumed evolution of a persistent cluster. For example, if you look at the *Additive* topic, it is predominately in trace G until 2015 when it splits into a mixture of trace G and trace J. The key difference is the contributions from the *Subtractive* versus the *Welding* topic areas. This can be interpreted by thinking of G as a persistent research thrust that dominates the *Additive* topic until a new area splits apart form it in the form of trace J that focuses more on research related to the *Subtractive* domain while G moves into *Welding*.

mize construction methods for said material. Additionally, there are occasional spikes of interest in types of processing for this topic, namely *Subtractive* and *Welding*, which can relate again to both the use and creation of such materials. This research thrust remains the most focused through time as no other thrust has a major contribution from *Advanced Materials and Bio-Manufacturing* and no other topics contribute strongly to this thrust.

Although not as pure as A, traces B, C, D, E, H, and G each represent the central or dominate research thrust for one of the seven semantic topics. With rare exception, each of these are easy to interpret as analogs to the main topic with only slight leanings towards other topics through time. Notably, there seemed to be a brief, but strong shift towards *Optimization* in the *Systems* dominating research thrust (D) during 2015. This temporary shift may have precipitated from a shift in directives from industry drivers or some new technology hype that ultimately did not remain relevant in this research thrust community.

Some notable splits in related research can also be seen in Figure 6. For example, the divergence of I in 2014 shows that unique interest area revolving specifically around *Metal Forming* and *Welding*. This reflects a growing interest in general production methods as opposed to specific individual methods. Soon after, this area picks up a noticeable contribution from *Advanced Materials and Bio-Manufacturing*, indicating that the growing interest in use of advanced materials.

The research thrust J splits off from the *Additive* dominated thrust G in 2015 by gaining significant contributions from the *Subtractive* topic area. This research thrust would seem to again be a leaning towards more general manufacturing techniques by investigating complementary methods. Even the minor contribution from *Welding* seems to confirm this. Both major splits from the single topic dominated thrusts seems to be a step in interest towards higher-level investigations.

The only thrust area to have significant contributions from multiple topics from inception is F. Relating to mostly to *Sub-*

tractive, *Optimization*, and to a smaller degree *Advanced Materials and Bio-Manufacturing*, this thrust is likely characterizing research on the manufacturing process itself. Interest in optimizing the manufacturing process has understandably been a consistent research area in this community.

Obviously the labels and insights related to each of the thrusts are framed through the lenses of the established semantic topics. Were different numbers of semantic topics chosen, or different methods for creating these topics used, this may have lead to capturing slightly different research-thrust traces over time. With any methods of data collapse and visualization some nuance and specific information will be lost. Even so, the basic trends and results from this analysis can shed interesting insight into both trends and overall focus of research efforts.

Other revealing metrics useful for characterizing the overall research efforts in a set of documents relate to the movement and size of each of the identified document swarms. Looking at relative movement can help to estimate progress and the existence of common driving forces or technologies between efforts. On the other hand, swarm volume or counts of included documents could be used to estimate interest and participation by communities, based on the intuition that more interest will produce more publications within the trace clusters (i.e. swarms). Although omitted for space in this paper, these metrics are simple to calculate and monitor for each identified trace via the process described above.

The major pitfall of this or any particle swarm tracking method is that it must be performed with a number of entities both large enough to successfully group and distributed in such a way to characterize the behavior you wish to capture. Due to the low fraction of retrieved documents prior to 2013, any trends or insights derived are hugely unreliable during those years. As shown in Figure 6, all discovered traces (A through J) derive from a single progenitor. This is strictly due to the low number of retrieved documents for that time frame only being able to produce a single cluster. With access to more of the published documents, it is expected that multiple progenitor or initiating research points would be discovered.

4 CONCLUSIONS & FUTURE WORK

This paper discusses a methodology to determine topic trends and evolution for bodies of research publications, with a specific case study of the AMSE MSEC conference. The paper provides the steps needed to repeat the process and interpret the results in the hopes that others will use this method for quantitatively analyzing new research areas throughout time.

One important takeaway from this paper is the need for multidisciplinary teams when analyzing these results. Merging domain expertise and NLP knowledge while interpreting these topics is key when using this analysis for decision making purposes. Another important takeaway is the necessity of consulting multiple visualizations when analyzing this data; no single metric or visualization can completely describe the complex interplay of research thrusts within a community. At various times while interpreting the results for this paper, the authors simultaneously consulted multiple visualizations to get the "big" picture. For example, by looking at Fig. 3, we noticed that that the term "cloud" started to appear in 2013. We were then able to consult Fig. 4 and further analyzed it place in the topic space. Once we discovered that "cloud" appears predominantly in the *Systems* topic, we could then use Fig. 5 to analyze other important terms from that topic. Finally, Fig. 6, allowed us to analyze the evolution of clusters of papers through time.

A key improvement for this analysis pipeline would be to better facilitate the iterative, collaborative process between research domain experts and the NLP algorithms or analysts (assuming they are not the same). For instance, allowing a mixture between domain expertise-driven topic definitions and latent-topic discovery would improve interpretability and understanding for the domain experts and ensure patterns recovered by the NLP algorithms and analysts are more likely to be useful. Future work should investigate various tools for continual human-NLP collaboration, taking cues from models like Anchored Correlation Explanation [24], which allow users to guide topics toward more human-readable distributions.

An application using the method presented in this paper could be adapted to automatically predict paper placement in symposiums or to help define research sub-communities of interest within a larger community, such as a conference. Other future work could be predicting research effort movement to anticipate future interests and create topics to allow conference planners to better organize current community interests. For example, given the trends in MSEC, optimization topics with advanced materials or additive manufacturing seem to be gaining momentum and may produce a unique research thrust centered in those topics soon. Lastly, this type of analysis could be used to better predict standards needs by studying the lag time between surges in academic publications in a space (e.g., additive manufacturing) and the first mentions and development of standards in the same space. This would allow standards organizations to better adapt to high velocity technical research and start analyzing standards needs more efficiently.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

[1] ASME, 2019. *"Manufacturing Engineering Division Newsletters"*. "Available at `https://community.asme.org/cfs-file.ashx/__key/communityserver-wikis-components-files/00-00-00-18-32/2158.ASME-MED-Fall-2018-Newsletter.pdf`. Accessed 11-13-19".

[2] ASME, 2019. *"Manufacturing Engineering Division"*. "Available at `https://community.asme.org/manufacturing_engineering_division/w/wiki/3639.about.aspx`. Accessed 11-13-19".

[3] Denny, M. J., and Spirling, A., 2018. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it". *Polit. Anal.,* ***26***(2), pp. 168–189.

[4] Palmer, D. D., 2000. "Tokenisation and sentence segmentation". *Handbook of natural language processing*.

[5] Kohlschütter, C., Fankhauser, P., and Nejdl, W., 2010. "Boilerplate detection using shallow text features". In Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, ACM, pp. 441–450.

[6] Friedl, J. E. F., 2002. *Mastering Regular Expressions*, second edition ed. "O'Reilly Media, Inc.", Sebastopol, California.

[7] Manning, C. D., and Schutze, H., 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

[8] Jurafsky, D., and Martin, J., 2000. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey.

[9] Manning, C. D., Raghavan, P., and Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[10] Isbell, C. L., and Viola, P., 1999. "Restructuring sparse high dimensional data for effective retrieval". In Advances in Neural Information Processing Systems, pp. 480–486.

[11] Krohn, J., Beyleveld, G., and Bassens, A., 2019. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Addison-Wesley Professional, Aug.

[12] Rao, D., and McMahan, B., 2019. *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning*. "O'Reilly Media, Inc.", Jan.

[13] Young, T., Hazarika, D., Poria, S., and Cambria, E., 2017. "Recent trends in deep learning based natural language processing".

[14] Akbik, A., Blythe, D., and Vollgraf, R., 2018. "Contextual string embeddings for sequence labeling". In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649.

[15] Řehůřek, R., and Sojka, P., 2010. "Software framework for topic modelling with large corpora". In LREC 2010 workshop New Challenges for NLP Frameworks., pp. 46–50.

[16] Chuang, J., Manning, C. D., and Heer, J. "Termite: Visualization techniques for assessing textual topic models". In Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12, ACM, pp. 74–77. event-place: Capri Island, Italy.

[17] Fogel, F., d'Aspremont, A., and Vojnovic, M. "Spectral ranking using seriation".

[18] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. "Latent dirichlet allocation". *J. Mach. Learn. Res.,* ***3***, pp. 993–1022.

[19] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M., 2009. "Reading tea leaves: How humans interpret topic models". In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds. Curran Associates, Inc., pp. 288–296.

[20] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A., 2011. "Optimizing semantic coherence in topic models". In Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp. 262–272.

[21] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. "Reading tea leaves: How humans interpret topic models". pp. 288–296.

[22] Blei, D. M., and Lafferty, J. D. "Dynamic topic models". In Proceedings of the 23rd international conference on Machine learning - ICML '06, ACM Press, pp. 113–120.

[23] Rahman, M. F., Liu, W., Suhaim, S. B., Thirumuruganathan, S., Zhang, N., and Das, G., 2016. "Hdbscan: Density based clustering over location based services". *arXiv preprint arXiv:1602.03730*.

[24] Gallagher, R. J., Reing, K., Kale, D., and Steeg, G. V. "Anchored correlation explanation: Topic modeling with minimal domain knowledge".