Emerald 110k:
A Multidisciplinary Dataset for Abstract Sentence Classification

Connor Stead [1] Stephen Smith [1] Peter Busch [1] Savanid Vatanasakdakul [2]

[1] Department of Computing, Macquarie University, Australia
[2] Carnegie Mellon University, Qatar

connor.stead@hdr.mq.edu.au stephen.smith@mq.edu.au
peter.busch@mq.edu.au savanid@cmu.edu

Abstract

Background: Datasets available for abstract sentence classification modelling are predominately comprised of abstracts sourced from biomedical research.
Aims: To contribute a large non-biomedical multidisciplinary dataset for abstract sentence classification model research.
Method: Bulk extract and transformation of Emerald Group Publishing structured abstracts indexed on Scopus.
Results: We present the largest multidisciplinary dataset for abstract sentence classification modelling, consisting of 1,050,397 sentences from 103,457 abstracts.

1 Introduction

Abstracts enable researchers to efficiently determine the relevance of literature to their research (Rowley, 1982, Collision, 1971, Cleveland and Cleveland, 2013). The desire to optimise this efficiency has resulted in the adoption of structured abstracts, which feature explicit headings reflecting key characteristics of a study. Examples of these headings include: aim, method, results and contributions. The alternative to structured abstracts are those where sentences addressing such characteristics are not specified.

Compared to unstructured alternatives, structured abstracts are perceived to offer greater value for researchers (Sharma and Harrison, 2006, Taddio et al., 1994 and Guimarães, 2006); permit advanced access to research findings (Mosteller et al., 2004), contain more relevant information (Budgen et al., 2008) and are easier to read (Kitchenham et al., 2008 and Budgen et al., 2008). Structured abstracts also increase the likelihood that relevant research is discovered (Eldredge, 2006, Mulrow, 1987, Haynes et al., 1990, Hartley, 1997, Bayley et al., 2002 and Bayley and Eldredge, 2003).

Natural language processing (NLP) has been used to automate the structuring of unstructured abstracts (Gonçalves et al., 2018; Jin and Szolovits, 2018, Dernoncourt et al., 2016); which is achieved through the development of Abstract Sentence Classification Models (ASCM), capable of classifying sentences sourced from unstructured abstracts into structured abstract headings.

This paper presents a novel dataset to advance ASCM research. The dataset introduced is unlike those already leveraged in ASCM development, primarily as it is comprised of abstracts originating from disciplines not yet explored in current research. The adoption of our dataset in future model development will enable the benchmarking of ASCM capability in new disciplines.

2 Related Work

There are numerous datasets available to researchers seeking to develop ASCM. These are outlined in table 1, an extension of the table presented by Dernoncourt and Lee (2017, p. 3). We extend their table by identifying the abstract's disciplinary domain. The size represents the number of abstracts reflected in the dataset. The 'manual' flag identifies if sentences were manually classified into structured abstract headings by the authors (Y) or were pre-structured in the original abstract (N).

The number of datasets available for ASCM does not directly correspond to the number of ASCM studies, as researchers re-use datasets to benchmark performance and to test novel algorithms. Further, studies may develop a dataset for model development without contributing the dataset as an artefact. Dernoncourt and Lee (2017) also presented a dataset in a standalone paper,

much like this body of work. Table 2 provides a summary of ASCM development efforts, along with the dataset used in model development.

Dataset	Size	Manual	Domain
Hara et al. (2007)	200	Y	BM (RCT)
Hirohata et al. (2008)	104k	N	BM
Chung (2009)	327	Y	BM (RCT)
Boudin et al. (2010)	29k	N	BM
Kim et al. (2011)	1k	Y	BM
Huang et al. (2011)	23k	N	BM
Robinson (2012)	1k	N	BM (RCT)
Zhao et al. (2012)	20k	Y	BM
Davis and Mollá (2012)	194	N	BM (RCT)
Huang et al. (2013)	20k	N	BM (RCT)
Dernoncourt and Lee (2017)	196k	N	BM (RCT)

Table 1: Existing ASCM datasets, BM = Biomedicine RCT = Randomised Controlled Trials.

It is evident the dataset contributed by Kim et al. (2011) and Dernoncourt and Lee (2017) enjoys significant adoption in ASCM development. This dataset represents in practice the concern that the almost exclusive benefactor of advancements in ASCM studies are researchers in the biomedical discipline, and that the abstracts of non-biomedical disciplines have predominately not been included in model development.

There are a few studies representing exceptions to the biomedical exclusive trend. These are identified in table 2 with an asterisk (*). The first example is Teufel and Moens (1998), who developed a Naive Bayes classifier using sentences retrieved from 201 computational linguistics and cognitive science abstracts, achieving 68.6% precision (p. 24). Further non-biomedical examples include Wu et al. (2006) who used the computer and information science academic index Citeseer as an abstract source and Liu et al. (2013) who used ScienceDirect, a primarily scientific and health science academic literature index. These datasets are not available for researcher utilisation.

In response to the lack of disciplinary diversity, we are exploring greater non-biomedical grounded ASCM development. We desire to increase the likelihood that ASCM capability can become a viable inter-disciplinary mechanism to increase research discovery and accessibility. As part of our research, we have created a novel multi-disciplinary abstract sentence dataset for future ASCM development. The dataset development process is outlined in the following section.

Study	Dataset
Teufel and Moens (1998) *	Study developed (Computation and language archive)
McKnight and Srinivasan (2003)	Study developed (Medline)
Shimbo et al. (2003)	Study developed (Medline)
Ito et al. (2004)	Study developed (Medline)
Yamamoto and Takagi (2005)	Study developed (Medline)
Wu et al. (2006) *	Study developed (Citeseer)
Lin et al. (2006)	Study developed
Xu et al. (2006)	Study developed (RCT – source unknown)
Ruch et al. (2007)	Study developed (Medline)
Hirohata et al. (2008)	Study developed (Medline)
Chung (2009)	Study developed (Medline)
Kim et al. (2011)	Study developed (Medline)
Lui (2012)	Kim et al., 2011
Verbeke et al. (2012)	Kim et al., 2011
Liu et al. (2013) *	Study developed (Science Direct)
Hassanzadeh et al. (2014)	Kim et al., 2011
Dernoncourt et al. (2016)	Kim et al., 2011 Dernoncourt et al., 2016
Nam et al. (2016)	Study developed (PubMed)
Jin and Szolovits (2018)	Kim et al., 2011 Dernoncourt and Lee, 2017
Gonçalves et al. (2018)	Dernoncourt and Lee, 2017

Table 2: Existing ASCM studies.

3 Dataset Development

We present a novel abstract sentence dataset for ASCM research. The dataset contains sentences retrieved from multi-disciplinary non-biomedical journal abstracts. Each sentence is classified as belonging to one of the following heading classes:

- Purpose
- Design/methodology/approach
- Findings
- Originality/value
- Social implications
- Practical implications
- Research limitations/implications

3.1 Abstract Identification

As of 2019, Emerald Group Publishing (henceforth: Emerald) publishes over 300 double-blind peer reviewed journals (Emerald Group Publishing, 2019). Emerald journals publish research from management, information science and engineering disciplines. This includes fields such as aerospace technology, management information systems, corporate governance, marketing, computing, accounting, public health, supply chain management and tourism (Emerald Group Publishing, 2019).

In 2005 Emerald began mandating the use of structured abstracts in their journal publications (Emerald Group Publishing Limited, 2005). The multidisciplinary nature of Emerald's journal portfolio combined with their mandated structured abstract adoption policy has resulted in a unique opportunity for ASCM development. However, existing ASCM research has failed to leverage Emerald journal abstracts for model development.

3.2 Abstract Extract

The Scopus academic literature index was utilised to obtain Emerald journal abstracts. This was due to the availability of an API to access Scopus content, as well as the reach and scope of the index. An initial examination of Scopus identified 336 Emerald journals available where research was published between 2005 and 2019. This count indicated that the Emerald portfolio was widely available through Scopus.

After determining the availability of Emerald journals on Scopus, we developed a Python program capable of autonomously querying Scopus for Emerald journal records, downloading results and storing them on a local machine. This was made possible by Elsevier's Scopus API (https://dev.elsevier.com/) and the Python package Pybliometrics (https://github.com/pybliometrics-dev/pybliometrics).

The program processed a CSV file containing a list of Emerald journal ISSN codes. The program iterated over each observation in the CSV, querying Scopus for all publications from the journal between 2004 and 2019. The year 2004 was chosen as it was possible that some journals adopted structured abstracts prior to 2005, the time in which Emerald mandated the use of structured abstracts across their publications (Emerald Group Publishing Limited, 2005). The downloaded observations did not include the full text of the article, only metadata such as: DOI, article title, authors, publication date and the abstract.

There were 138,613 journal article metadata observations retrieved from the Scopus queries. These were exported into a Microsoft Excel workbook for manual unstructured/structured abstract classification. An abstract was deemed to be structured if it featured the Emerald structured abstract headings and these headings were used to separate components of what would otherwise have been free text abstracts. As a result, 109,608 abstracts were classified as structured, with the remaining abstracts discarded.

3.3 Abstract Sentence Transformation

Existing datasets utilised in ASCM research are presented as sentence level observations, featuring a sentence string with its corresponding structured abstract class. To ensure easy adoption in model development, it was necessary to deconstruct the abstracts into sentences, whilst maintaining the structured abstract class they reflected.

A program was developed which processed each abstract, identifying the locations of the structured headings and treating them as delimiters. This segmented the base abstract string into heading level substrings. We then used a tokenizer to split these into sentence strings, which were reviewed to identify data quality issues such as: sentences incorrectly split from the tokenizer (for example, seeing i.e. as an end of sentence condition), presence of a copyright indicator as the last sentence observation and invalid heading classes.

Any data quality issues identified were managed either through sentence modification or removal of the base abstract; which ensured the dataset contained all sentences from base abstracts.

3.4 Resulting Dataset

Post sentence transformation, we formed a dataset consisting of 1,050,397 sentences originating from 103,457 abstracts. A heading level summary of the sentence abstract count is provided in table 3. Sentence per abstract and token per sentence frequency as well as descriptive statistics are provided in figures 1 and 2. We note the low frequency for the 'Social implications' class. Table 4 identifies the sentence and abstract counts for the top 15 (of 406) journals featuring abstracts. This demonstrates its multidisciplinary nature.

We named our dataset Emerald 110k, following the ASCM dataset naming convention set by

Dernoncourt and Lee (2017) with their biomedical dataset PubMed 200k. The 110k reflects the 103,457 Emerald abstracts from which sentences originate. Our dataset is available via GitHub (https://github.com/connorstead/emerald_ascm) in .CSV, .SAS7BDAT and Python .PKL to enable cross platform utilisation.

Heading	Sentences	Abstracts
Purpose	198,277	103,394
Design/methodology/approach	223,312	101,328
Findings	269,321	103,268
Originality/value	187,986	102,559
Social implications	26	15
Practical implications	92,243	48,689
Research limitations /implications	79,232	40,544

Table 3: Heading level summary of resulting dataset

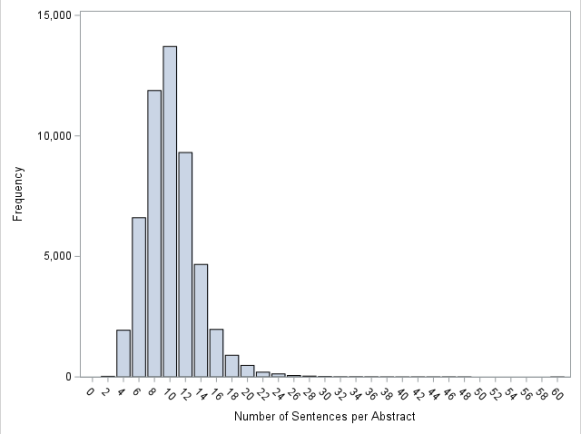

Figure 1: Sentence per Abstract Frequency. Minimum: 1 Maximum: 60 Mean: 10.1530 Standard Deviation: 3.4253 Skewness: 1.2720 Kurtosis: 5.2848

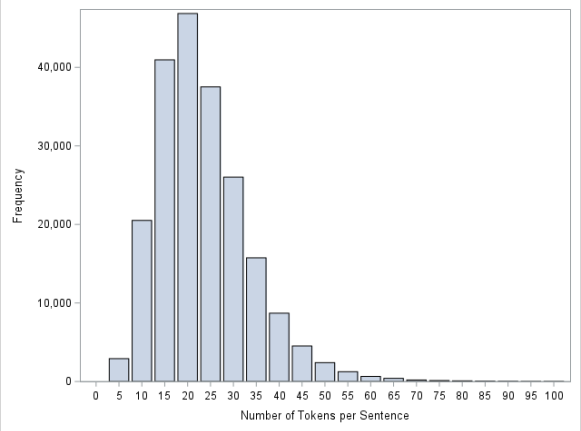

Figure 2: Tokens per Sentence Frequency. Minimum: 1 Maximum: 309 Mean: 23.34 Standard Deviation: 10.4790 Skewness: 1.3217 Kurtosis: 5.5276

Journal (ISSN)	Sentences	Abstracts
International Journal for Computation and Mathematics in Electrical and Electronic Engineering (03321649)	16,659	1,627
British Food Journal (0007070X)	15,821	1,477
Kybernetes (0368492X)	14,676	1,513
Management Decision (00251747)	14,518	1,455
International Journal of Numerical Methods for Heat and Fluid Flow (09615539)	12,365	1,217
European Journal of Marketing (03090566)	12,283	1,136
Industrial Management and Data Systems (02635577)	10,728	978
International Journal of Contemporary Hospitality Management (09596119)	10,629	1,046
Engineering Computations (02644401)	10,471	1,026
International Journal of Social Economics (03068293)	9,730	970
Industrial Lubrication and Tribology (00368792)	9,611	941
Rapid Prototyping Journal (13552546)	9,593	863
Strategic Direction (02580543)	9,355	1,267
Benchmarking (14635771)	9,343	815
Journal of Knowledge Management (13673270)	9,017	859

Table 4: Sentence and abstract frequency for the top 15 journals in the dataset (ordered by sentence count)

4 Conclusion and Ongoing Research

This paper explored the development of a novel dataset for ASCM research. The novelty of this dataset is primarily due to its composition of abstract sentences from a range of non-biomedical disciplinary literature. Our dataset is also the second largest dataset available. It offers a unique opportunity for ASCM researchers to explore the performance of their models outside of biomedical abstract datasets.

Our future research is concerned with expanding ASCM outside of biomedicine and providing associated advancements to new disciplines. Accordingly, we are utilizing this dataset in our own exploration of state of the art ASCM development. We also intend to update this dataset as additional Emerald structured abstracts are published each year, whilst seeking to identify new sources of structured abstracts for ASCM research.

5 Acknowledgements

The authors wish to acknowledge the Australian Government Research Training Program Scholarship which enabled this research to take place.

6 References

Bayley, L., & Eldredge, J. (2003). The structured abstract: an essential tool for researchers. Hypothesis, 17(1), 11-13.

Bayley, L., Wallace, A., & Brice, A. (2002). Evidence based librarianship implementation committee. research results, dissemination task force recommendations. Hypothesis, 16(1), 6-8.

Boudin, F., Nie, J.-Y., Bartlett, J. C., Grad, R., Pluye, P., & Dawes, M. (2010). Combining classifiers for robust PICO element detection. BMC Medical Informatics Decision Making, 10(1), 29.

Budgen, D., Kitchenham, B. A., Charters, S. M., Turner, M., Brereton, P., & Linkman, S. G. (2008). Presenting software engineering results using structured abstracts: a randomised experiment. Empirical Software Engineering, 13(4), 435-468.

Chung, G. Y. (2009). Sentence retrieval for abstracts of randomized controlled trials. BMC Medical Informatics Decision Making, 9(1), 10.

Cleveland, A. D., & Cleveland, D. B. (2013). Introduction to indexing and abstracting. Santa Barbara, California: ABC-CLIO.

Collison, R. L. (1971). Abstracts and abstracting services. Santa Barbara, California: ABC-CLIO.

Davis-Desmond, P., & Mollá, D. (2012). Detection of evidence in clinical research papers. Paper presented at the Proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management-Volume 129.

Dernoncourt, F., & Lee, J. Y. (2017). Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:.06071.

Dernoncourt, F., Lee, J. Y., & Szolovits, P. (2016). Neural networks for joint sentence classification in medical paper abstracts. arXiv preprint arXiv:.05251.

Eldredge, J. (2006). Evidence-based librarianship: the EBL process. Library hi tech, 24(3), 341-354.

Emerald Group Publishing Limited. (2005). Emerald structured abstracts have arrived! Journal of Managerial Psychology, 20(1).

Emerald Group Publishing Limited. (2019). Emerald | Product Information | Journals. Retrieved from https://www.emeraldgrouppublishing.com/products/journals/index.htm

Gonçalves, S., Cortez, P., & Moro, S. (2018). A Deep Learning Approach for Sentence Classification of Scientific Abstracts. Paper presented at the International Conference on Artificial Neural Networks.

Guimarães, C. A. (2006). Structured abstracts: narrative review. Acta cirurgica brasileira, 21(4), 263-268.

Hara, K., & Matsumoto, Y. (2007). Extracting clinical trial design information from MEDLINE abstracts. New Generation Computing, 25(3), 263-275.

Hartley, J. (1997). Is it appropriate to use structured abstracts in social science journals? Learned Publishing, 10(4), 313-317.

Hassanzadeh, H., Groza, T., & Hunter, J. (2014). Identifying scientific artefacts in biomedical literature: The Evidence Based Medicine use case. Journal of Biomedical Informatics, 49, 159-170.

Haynes, R. B., Mulrow, C. D., Huth, E. J., Altman, D. G., & Gardner, M. J. (1990). More informative abstracts revisited. Annals of internal medicine, 113(1), 69-76.

Hirohata, K., Okazaki, N., Ananiadou, S., & Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. Paper presented at the Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I.

Huang, K.-C., Chiang, I.-J., Xiao, F., Liao, C.-C., Liu, C. C.-H., & Wong, J.-M. (2013). PICO element detection in medical text without metadata: Are first sentences enough? Journal of Biomedical Informatics, 46(5), 940-946.

Huang, K.-C., Liu, C. C.-H., Yang, S.-S., Xiao, F., Wong, J.-M., Liao, C.-C., & Chiang, I.-J. (2011). Classification of PICO elements by text features systematically extracted from PubMed abstracts. Paper presented at the 2011 IEEE International Conference on Granular Computing.

Ito, T., Shimbo, M., Yamasaki, T., & Matsumoto, Y. (2004). Semi-supervised sentence classification for medline documents. Methods, 138, 141-146.

Jin, D., & Szolovits, P. (2018). Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts. arXiv preprint arXiv:.06161.

Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. Paper presented at the BMC bioinformatics.

Kitchenham, B. A., Brereton, O. P., Owen, S., Butcher, J., & Jefferies, C. (2008). Length and readability of structured software engineering abstracts. IET software, 2(1), 37-45.

Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. Paper presented at the Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology.

Liu, Y., Wu, F., Liu, M., & Liu, B. (2013). Abstract sentence classification for scientific papers based on transductive SVM. Computer Information Science, 6(4), 125.

Lui, M. (2012). Feature stacking for sentence classification in evidence-based medicine. Paper presented at the Proceedings of the Australasian Language Technology Association Workshop 2012.

McKnight, L., & Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. Paper presented at the AMIA Annual Symposium Proceedings.

Mosteller, F., Nave, B., & Miech, E. J. (2004). Why we need a structured abstract in education research. Educational Researcher, 33(1), 29-34.

Mulrow, C. D. (1987). The medical review article: state of the science. Annals of internal medicine, 106(3), 485-488.

Nam, S., Kim, S.-K., Kim, H.-G., Ngo, V., & Zong, N. (2016). Structuralizing biomedical abstracts with discriminative linguistic features. Computers in Biology Medicine, 79, 276-285.

Robinson, K. A., Saldanha, I. J., & Mckoy, N. A. (2011). Development of a framework to identify research gaps from systematic reviews. Journal of Clinical Epidemiology, 64(12), 1325-1330.

Rowley, J. E. (1982). Abstracting and indexing. London: Clive Bingley.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., . . . Lovis, C. (2007). Using argumentation to extract key sentences from biomedical abstracts. International Journal of Medical Informatics, 76(2-3), 195-200.

Sharma, S., & Harrison, J. E. (2006). Structured abstracts: do they improve the quality of information in abstracts? American journal of orthodontics dentofacial orthopedics, 130(4), 523-530.

Shimbo, M., Yamasaki, T., & Matsumoto, Y. (2003). Using sectioning information for text retrieval: a case study with the medline abstracts. Paper presented at the Proceedings of Second International Workshop on Active Mining (AM'03).

Taddio, A., Pain, T., Fassos, F. F., Boon, H., Ilersich, A. L., & Einarson, T. R. (1994). Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. CMAJ: Canadian Medical Association Journal, 150(10), 1611.

Teufel, S., & Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. Paper presented at the AAAI Spring Symposium on Intelligent Text summarization.

Verbeke, M., Van Asch, V., Morante, R., Frasconi, P., Daelemans, W., & De Raedt, L. (2012). A statistical relational learning approach to identifying evidence based medicine categories. Paper presented at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Wu, J.-C., Chang, Y.-C., Liou, H.-C., & Chang, J. S. (2006). Computational analysis of move structures in academic abstracts. Paper presented at the Proceedings of the COLING/ACL on Interactive presentation sessions.

Xu, R., Supekar, K., Huang, Y., Das, A., & Garber, A. (2006). Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts. Paper presented at the AMIA Annual Symposium Proceedings.

Yamamoto, Y., & Takagi, T. (2005). A sentence classification system for multi biomedical literature summarization. Paper presented at the 21st International Conference on Data Engineering Workshops (ICDEW'05).

Zhao, J., Bysani, P., & Kan, M.-Y. (2012). Exploiting classification correlations for the extraction of evidence-based practice information. Paper presented at the AMIA Annual Symposium Proceedings.