FINDING A COMMON GROUND IN HUMAN AND MACHINE-BASED TEXT PROCESSING

Roman Taraban
roman.taraban@ttu.edu
Lakshmojee Koduru
lakshmojee.koduru@ttu.edu
Mark LaCour
mark.lacour@ttu.edu
Philip Marshall
philip.marshall@ttu.edu
Texas Tech University, USA

Received June 12, 2018; Revised June 13, 2018; Accepted June 20, 2018

Abstract. Language makes human communication possible. Apart from everyday applications, language can provide insights into individuals' thinking and reasoning. Machine-based analyses of text are becoming widespread in business applications, but their utility in learning contexts are a neglected area of research. Therefore, the goal of the present work is to explore machine-assisted approaches to aid in the analysis of students' written compositions. A method for extracting common topics from written text is applied to 78 student papers on technology and ethics. The primary tool for analysis is the Latent Dirichlet Allocation algorithm. The results suggest that this machine-based topic extraction method is effective and supports a promising prospect for enhancing classroom learning and instruction. The method may also prove beneficial in other applied applications, like those in clinical and counseling practice.

Keywords: *natural language processing, machine-analysis, latent Dirichlet allocation, text analysis, classroom learning, clinical and counseling practice.*

Тарабань Роман, Кодуру Лакшмоджі, ЛаКур Марк, Маршалл Філіп. Пошук спільних рис під час обробки текстів людиною та машиною.

Анотація. Мова уможливлює людське спілкування. Крім повсякденних застосувань, мова може забезпечити розуміння думок та міркувань людей. Машинний аналіз тексту набуває великої популярності у сфері ведення бізнесу, проте його корисність у навчальному процесі залишається досі недослідженою темою. Тому мета статті – дослідити автоматизовані підходи, що можуть бути корисними під час аналізу писемної продукції студентів. 78 студентських робіт із галузей технології та етики було піддано аналізу з використанням методу вилучення загальних тем із письмового тексту. Основним інструментом для аналізу став алгоритм латентного розташування Дирихле. Результати свідчать про те, що цей автоматизований інструмент виокремлення теми є ефективним й перспективним у плані підвищення рівня навчання в аудиторії та викладання. Метод також може бути застосованим в інших прикладних програмах, наприклад, у тих, якими користуються під час клінічної практики та консультування.

Ключові слова: *обробка природної мови, машинний аналіз, латентне розташування Дирихле, аналіз тексту, навчання в класі, клінічна практика та консультування.*

East European Journal of Psycholinguistics, 5(1), 83–91. https://doi.org/10.5281/zenodo.1436358

1. Introduction

Human communication depends on language. Everyday tasks, at home and at work, are accomplished through language. Our first premise in the present work is that apart from the practical applications of language in everyday interactions, the language that an individual uses may reveal deeper aspects of the person. Wilhelm von Humboldt (1767-1835), a well-known German diplomat and scholar wrote: "Language is the outward manifestation of the spirit of people: their language is their spirit, and their spirit is their language; it is difficult to imagine any two things more identical" (in Salzmann, 2004:42). The linguist, Edward Sapir, believed that "language and our thought-grooves are inextricably interwoven, [and] are, in a sense, one and the same" (in Salzmann, 2004:43). In psychological research, Pennebaker and King (1999) proposed that "the way people talk about themselves reveals important information about them" (p. 1297). Chung & Pennebaker (2008) analyzed college students' narratives in order to gain insight into students' self-concepts and personality traits. In that work, as well as other work (Pennebaker et al., 2014, 2015), Pennebaker and colleagues showed that expressive writing could be effectively analyzed and interpreted through the aid of automated computer methods.

In the last ten years, there has been an upsurge in the use of machine systems to analyze natural language. The origin of the recent exponential upturn in machine-based language processing capacity can be attributed in large part to two factors: an increase in the physical storage capacity and processing speeds of computing systems; and significant advances in Bayesian and other analytic methods. Machine tools for analyzing language operate according to a fundamental computational principle: *There are probabilistic markers (cues, features) in the input (e.g., student essays) that characterize key constructs in the input*. We take this principle as the second premise in the present study. The theoretical position taken in some models regarding the fundamental role of probabilistic features in language acquisition and processing (see Taraban & Marshall, 2017) is also consistent with this second premise. Thus, the reliance on probabilistic representations of linguistic features forms a common ground in human and machine-based language processing.

The present paper takes the first steps in exploring the possibility that by extracting and identifying key elements of texts, machine-based systems can mimic, in part, the classification and interpretation of students' written work by humans. From an applied perspective, machine processing of students' academic writing may afford educators automated aid in the analysis and evaluation of students' work. In the discussion, we consider similarities and differences between human and machine processing of written text, extensions of these methods to other areas, and the limitations of this methodology.

1.1. Natural Language Processing

Natural language processing (NLP) refers to the use of computers and artificial intelligence (AI) to process and analyze natural language in written or spoken form. NLP encompasses speech recognition, language comprehension, and language production. Counterparts to today's intelligent language interfaces

emerged decades earlier, including Weizenbaum's (1966) ELIZA, Winograd's (1972) blocks world (SHRDLU), and Schank and colleagues' (1975) MARGIE. As early as the 1950s, machine translation was showing some early success. By the 1980s, developers were creating machine-based conversational partners, typically to assist in practical situations.

Language varies by genre. Research papers, blogs, and Twitter, for instance, have different writing styles. Machine analysis has addressed these differences in roughly the same way, which is to apply machine-learning algorithms to a sample of the material of interest, after which the machine learning algorithm can automatically discover similar patterns in new materials. In so-called supervised learning, the human analyst knows in advance which patterns or topics should be learned and generalized by the machine to new materials. In unsupervised learning, the analyst depends on the machine algorithm to discover the patterns or topics in the training materials. Latent Dirichlet Allocation (LDA) is a topic modeling algorithm, and is one example of unsupervised machine learning. This is the algorithm of interest in the present study.

1.2. A Case Study

The context of this study is a sophomore-level course that is offered to engineering majors at Texas Tech University in the U.S. (Taraban et al., 2017, 2018). This course develops ethical reasoning through an introduction to ethical theories and contemporary ethical issues in engineering, technology and society. Course materials and assignments consider *intuitionism*, which is a person's intuitive reaction to ethical issues, three ethical theories – i.e., *utilitarianism, respect for persons,* and *virtue ethics* – and the National Society of Professional Engineers Code of Ethics, which is an accepted code of ethics for professional engineers in the U.S. Course activities require students to analyze and respond to ethical issues in contemporary social settings involving engineering dilemmas. A major course requirement is a capstone paper incorporating Social Impact Analysis (SIA). The general purpose of SIA is to identify and analyze the positive and negative social consequences of engineering plans and projects. In students' SIA papers, they identify and discuss a contemporary engineering technology (e.g., autonomous tractor trailers, fracking, drones, ethical hacking). They are required to incorporate knowledge from one or more of the ethical theories into their analyses.

The goal of the present study was to develop and test the application of the Latent Dirichlet Allocation (LDA) algorithm for the automatic extraction of topics in a random sample of capstone papers submitted by students in the ethics course to fulfill a course requirement.

Three empirical questions guided this analysis:

- Can LDA find topical differences that distinguish between Non-Ethics and Ethics documents?
- Do the respective topics make sense?
- Could a distribution of topics within each document be developed based on the LDA output?

2. Methods

The materials for this analysis were a random sample of 78 capstone papers that were submitted for course credit in the ethics course described earlier. In a previous study (Taraban et al., 2017), each paper was recompiled by the researchers into two parts. One part consisted of the text in the paper that discussed the engineering technology that was the subject of the paper; the other part consisted of the text that described the ethics associated with that technology. This resulted in 78 documents consisting of technical descriptions and 78 documents consisting of ethics discussions, referred to here as the Non-Ethics and Ethics texts, respectively. These two types of texts were analyzed separately, as described next.

2.1. Software Tools

Two computer applications, LDA and MEH, were applied in this study. MEH is available online at no cost (see https://meh.ryanb.cc/). The MEH website includes a link to R computer language code for LDA https://meh.ryanb.cc/understanding-output/ that we used with small modifications described below. The R code ran using R-Studio https://www.rstudio.com/products/rstudio/download2/ .

LDA (Latent Dirichlet Allocation). LDA is an unsupervised machine learning algorithm (Blei, Ng, & Jordan, 2003). LDA is based on the assumption that a person composing a document has a number of topics in mind and that these topics can be recovered from an analysis of the document (Ostrowski, 2015). LDA treats each document as a mixture of topics and every topic as a distribution over the words in the document. The goal of applying LDA is to identify latent topic information across a collection of documents. LDA assumes that all documents in the collection share the same set of topics, but each document exhibits those topics with different proportions. Thus, it should be possible to recover an estimate of the distribution of topics within a document.

MEH (Meaning Extraction Helper). MEH carries out a number of relevant functions related to text analysis prior to the application of LDA. Of interest here is the construction of a Document-Term Matrix (DTM), which LDA uses in order to identify the document topics. MEH constructs a DTM by first deleting *stop words* from the documents. Stop words are typically function words, including conjunctions, determiners, and prepositions, which carry little lexical meaning in a document. The remaining content words are converted to lemmas, that is, inflectional endings are removed, leaving only the base form of the word. Each document is represented as a vector of lemmas in the DTM.

2.2. Procedure

The 78 ethics and 78 non-ethics document sets were analyzed separately, by applying the following steps. The MEH software was opened and the application Wizard was chosen to guide the process. First the document files were uploaded to MEH, stop words were deleted from the documents, the remaining words were converted to lemmas, and a document-term matrix (DTM) was chosen as the output. R Studio was then opened, the LDA code described above was opened in R Studio, the number of requested topics and terms per topic were input in the R code, and three additional function calls were made in the code, which were for 1. document-

to-topic distribution, 2. topic-to-term probabilities, and 3. the probabilities of topics associated with each document. These additional functions provided the data for Figure 1 and Table 1 in this paper.

3. Results

When LDA was applied to the Non-Ethics content of the SIA papers, five prominent topics corresponded to the topics that students often chose to focus on in their papers: Topic 1: company organization and stakeholders; Topic 2: technical aspects of hydraulic fracking; Topic 3: technical aspects of solar energy roadways; Topic 4: artificial intelligence technology; Topic 5: electric vehicle technology.

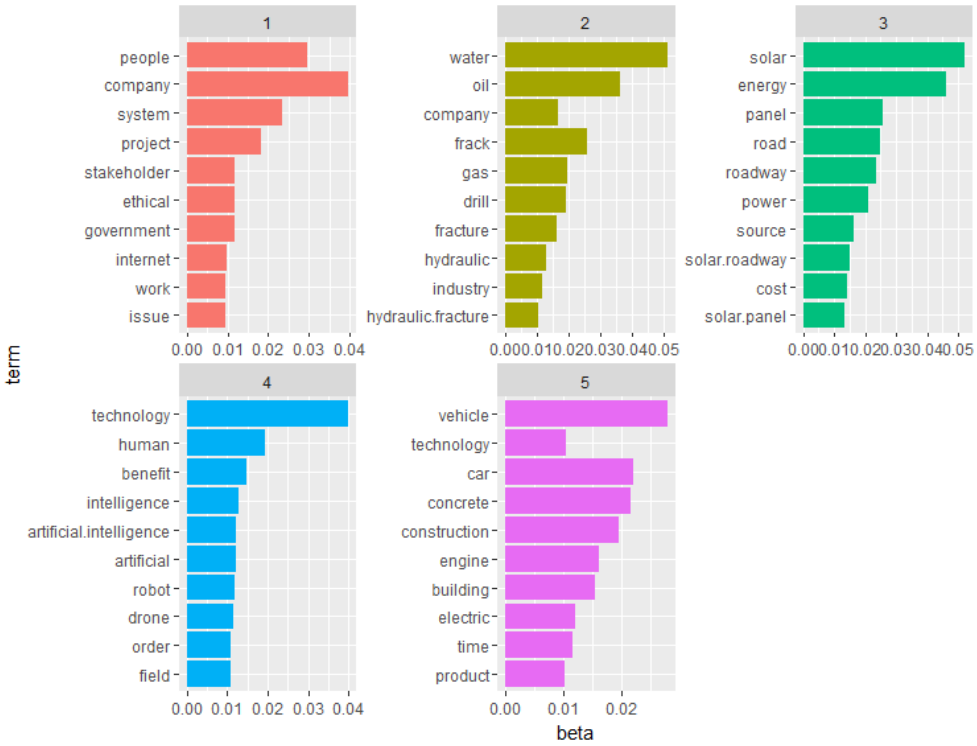

Fig. 1. *Topics and Weighted Lemmas Associated with Non-Ethical Document Topics*

The ten most frequent terms for each topic and the weights associated with the most critical lemmas (concepts) associated with the topics are shown in Figure 1. When LDA was applied to the Ethics content of the SIA papers, a visibly different set of topics emerged.

Representative topics were as follows: Topic 1: environmental concerns associated with oil fracking; Topic 2: general ethical themes related to public health, safety, the environment, and engineering NSPE code; Topic 3: human benefits of technology and ethical theory of utilitarianism; Topic 4: human benefits associated with solar highways; Topic 5: ethical issues associated with autonomous vehicles.

Curiously, the lemma "ethical" appears in Figure 1, in Non-Ethics Topic 1. The likely reason is that some students chose to write on the topic of "ethical computer hacking" in industry and government. In the non-ethics portions of the papers, the

term "ethical" appeared in descriptions of the practice of ethical hacking and ethical hackers, but without connections to ethics per se. So here, LDA included the term "ethical," but without reference to ethics. This is one example of how LDA output could be somewhat confusing and could require closer examination of the source documents.

Fig. 2. *Topics and Weighted Terms Associated with Ethical Document Topics*

Overall, the differences in topics in Non-Ethics and Ethics documents recovered by LDA support a positive response to the first empirical question: *Can LDA find topical differences that distinguish between Non-Ethics and Ethics documents?* The second question will be addressed informally: *Do the respective topics make sense?* Based on careful reading of the associated SIA papers, there is a clear relationship between the technical and ethical topics recovered by LDA and the content in the SIA papers. A more formal statistical approach to this question will be undertaken in future work (cf., Chen & Wang, 2007). The third empirical question: *Could a distribution of topics within each document be developed based on the LDA output?* is addressed in Table 1, which shows the ethics-text outcomes for a sample of ten of the 78 participants in this study. The Table shows that each participant's document can be described as a mixture of topics, with probabilities associated with each topic. The magnitudes of the probabilities in Table 1 show that some participants in the sample focused on specific topics (e.g., participants 2, 3, 8), while others showed flatter distributions across multiple topics (e.g., participants 5, 7). Thus, the answer to the third empirical question is also affirmative. As is the case with the second question, this question also deserves a more formal statistical assessment approach to establish the validity of these probability distributions.

Table 1

Probability Distribution of Ethics Topics in Participants' Documents (Dominant topic in each document is bolded)

	Topics				
Participant	1	2	3	4	5
1	0.15	0.19	**0.42**	0.09	0.15
2	0.03	0.056	**0.75**	0.13	0.04
3	0.04	**0.75**	0.10	0.06	0.05
4	0.13	**0.35**	0.32	0.10	0.10
5	0.05	0.09	0.36	**0.38**	0.11
6	0.08	0.09	0.30	**0.46**	0.07
7	0.22	0.18	0.16	0.13	**0.32**
8	0.06	0.07	0.16	**0.66**	0.05
9	0.06	0.07	0.21	0.12	**0.53**
10	0.10	0.11	0.24	**0.42**	0.13

Note: Topic 1: environmental concerns associated with oil fracking; Topic 2: general ethical themes related to public health, safety, the environment, and engineering NSPE code; Topic 3: human benefits of technology and ethical theory of utilitarianism; Topic 4: human benefits associated with solar highways; Topic 5: ethical issues associated with autonomous vehicles.

4. Conclusions

Machine-based analyses of texts are becoming widespread in business applications and in the analysis of social media exchanges. The present study examines prospects for applications in learning contexts, which is presently a somewhat neglected area of scholarship.

A premise behind this work, stated in the introduction, was that *There are probabilistic markers (cues, features) in the input (e.g., student essays) that characterize key constructs in the input*. There is general agreement in how language is learned and processed, and as Feldman (1999) points out, part of learning involves finding the key patterns in linguistic utterances:

Children learn language by discovering patterns and templates. We learn how to express plural or singular and how to match those forms in verbs and nouns. We learn how to put together a sentence, a question, or a command. Natural Language Processing assumes that if we can define those patterns and describe them to a computer then we can teach a machine something of how we speak and understand each other. Much of this work is based on research in linguistics and cognitive science. p. 62

The premise that relates language processing to the probabilistic processing of key terms and patterns is consistent with the computational view that has been adopted across the domains of cognitive science, artificial intelligence, and cognitive neuroscience. Within this class of computational models, weighting and processing probabilistic cues forms the common ground of human and machine processing.

On the other hand, more empirical support is needed for the machine-based principles that the brain implements probabilistic information in near-optimal ways (Recchia et al., 2015: 13). Further, in spite of how valid "patterns and templates" may be (Feldman, 1999), they do not characterize a complete description of the nature of language. As Jerome Bruner (1990) and others, like John Searle (in Mishlove, 2010) have argued long before the more recent upsurge in machine analysis, language has a significant component of meaning. Current methods, like LDA, capture the patterns, i.e. the syntax, of the analyzed texts. However, extensions of this methodology, or new methods, are required in order to extract and represent the deeper meaning in these texts.

The strength of the LDA method is the ability to separate out topics in large corpora. It is often the case in large classroom sections that the instructor must resort to fixed assessment instruments, like multiple-choice tests, because of a shortage of human resources to evaluate open-ended questions related to course materials. The success of automated methods, like machine-based applications of LDA, will allow instructors to provide students with more open-ended opportunities to express themselves. This is because LDA and related applications may be able to extract the meaning being communicated by students instead of requiring instructors to figure out what students are trying to communicate. Further, these machine applications may be able to automate at least part of the feedback and grading process, providing students with helpful feedback and guidance.

The prospects for LDA and related applications are good, in areas of ethics and technology, as in the present example, but in other areas as well, for instance, in applied clinical and counseling settings, as in the seminal work of Pennebaker and colleagues (Chung & Pennebaker, 2008; Pennebaker, 2004; Pennebaker & King, 1999; Pennebaker et al., 2014, 2015). Clinicians, for instance, may be able to apply these methods to extract the central topics in therapeutic writing by clients on specific topics. This additional source of input could potentially improve the effectiveness of treatment.

There are limitations to immediate classroom and other applications of LDA. Primarily, there are currently only a few statistical methods available for estimating the most valid number of relevant topics or terms in a corpus of documents (Chen & Wang, 2007). The best method for assessing the validity of LDA analyses may still be manual inspection by human judges. The problem with human judgment, however, is that this is a time-consuming process. Ideally, methods will be developed to more directly and transparently assess the validity of LDA analyses.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993-1022.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Chen, K. Y. M., & Wang, Y. (2007). Latent dirichlet allocation. http://acsweb.ucsd.edu/~yuw176/report/lda.pdf.

Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, *42*(1), 96-132.

Feldman, S. (1999). NLP meets the Jabberwocky: Natural language processing in information retrieval. *Online Magazine*, *23*, 62-73. Retrieved from: http://www.onlinemag.net/OL1999/feldmann5.html

Mishlove, J. (2010). https://www.youtube.com/watch?v=0XTDLq34M18 (Accessed June 12, 2018).

Ostrowski, D. A. (2015). Using latent dirichlet allocation for topic modelling in twitter. In *Semantic Computing (ICSC), 2015 IEEE International Conference (*pp. 493-497). IEEE.

Pennebaker, J. W. (2004). Theories, therapies, and taxpayers: On the complexities of the expressive writing paradigm. *Clinical Psychology: Science and Practice*, *11*(2), 138-142.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC 2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, *9*(12), e115844.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296-1312.

Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation. *Computational intelligence and neuroscience*, *2015*, 1-18.

Salzmann, Z. (2004). *Language, Culture, and Society: An Introduction to Linguistic Anthropology (3rd ed)*. Westview Press.

Schank, R. C., Goldman, N. M., Rieger III, C. J., & Riesbeck, C. (1973). MARGIE: Memory analysis response generation, and inference on English. In *IJCAI*, *3*, 255-261.

Taraban, R., Marcy, W. M., LaCour Jr., M. S., & Burgess II, R. A. (2017). Developing machine-assisted analysis of engineering students' ethics course assignments. *Proceedings of the American Society of Engineering Education (ASEE) Annual Conference*, Columbus, OH. https://www.asee.org/public/conferences/78/papers/19234/view.

Taraban, R., Marcy, W. M., LaCour, M. S., Pashley, D., & Keim, K. (2018). Do engineering students learn ethics from an ethics course? *Proceedings of the American Society of Engineering Education – Gulf Southwest (ASEE-GSW) Annual Conference*, Austin, TX. http://www.aseegsw18.com/papers.html.

Taraban, R., & Marshall, P. H. (2017). Deep learning and competition in psycholinguistic research. *East European Journal of Psycholinguistics*, *4*(2), 67-74.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Winograd, T. (1972). Understanding natural language. New York: Academic Press.