ELEKTROTEHNIŠKI VESTNIK 89(1-2): 64–72, 2022
ORIGINAL SCIENTIFIC PAPER

Topic extraction by clustering word embeddings on short online texts

David Nabergoj[1,†], **Alessandro D'Alconzo**[2], **Danilo Valerio**[2], **Erik Štrumbelj**[1]

[1]*Faculty of Computer and Information Science, University of Ljubljana, Slovenia*
[2]*Siemens AG Austria*

[†] *E-mail: david.nabergoj@student.uni-lj.si*

Abstract. We demonstrate our topic extraction method in which topics are treated as clusters of word embeddings. The OPTICS algorithm is used to find small and arbitrarily-shaped clusters of embeddings, produced by a fastText model. The result is a set of dominant and non-dominant domain-specific topics. The focus of the method is on short online posts which are difficult to analyze with traditional topic extraction approaches because of the word collocation scarcity. The method is tested on dataset of posts from Twitter, LinkedIn and company blogs related to industrial automation. The method significantly outperforms traditional topic extraction approaches by finding relevant and understandable topics with related tokens.

Keywords: topic extraction, industrial automation, text mining

Iskanje tem v kratkih spletnih besedilih z gručenjem vložitev besed

Predstavimo novo metodo za iskanje tem v besedilih, ki teme obravnava kot gruče vložitev besed, dobljenih z modelom fastText. Naša metoda z algoritmom OPTICS najde gruče poljubnih oblik, ki predstavljajo prevladujoče in tudi manj opazne domensko specifične teme. Metoda je primerna za kratka spletna besedila, ki jih je težko analizirati s klasičnimi pristopi za iskanje tem zaradi majhnega števila kolokacij. Metodo testiramo na podatkovni množici objav s strani Twitter, LinkedIn in blogov različnih podjetij, povezanih z industrijsko avtomatizacijo. Naša metoda najde relevantne in razumljive teme s smiselnimi besedami ter deluje bistveno boljše od klasičnih pristopov.

1 INTRODUCTION

When doing market research, companies are interested in what their customers, suppliers, and competitors are doing and talking about. Gathering relevant information is a resource intensive task, especially if done manually. Instead, our work is aimed at automatically extracting topics from large collections of social media and blog posts. In particular, we propose an approach that can also identify less dominant and domain-specific topics that are often missed by standard approaches to topic extraction.

Topic extraction is typically performed using latent Dirichlet allocation (LDA) [1] and latent semantic analysis (LSA) [2]. Another commonly used probabilistic model is the hierarchical Dirichlet process (HDP) [3]. However, it has been found that such techniques achieve poor performance on shorter texts [4], which appear on social media and are an important source of information. Recent research in this area has focused on improving existing probabilistic models for Twitter posts [5], [6] and general documents [7].

Other approaches find topics by clustering vector representations of documents or tokens. Early research describes clustering documents, represented by TF-IDF [8] or bag-of-words features [9], [10]. More recent research focuses on clustering word embeddings [11]–[14]. Language models which produce such embeddings have gained a lot of traction because of their ability to capture contextual information better than bag-of-words-based approaches. The original proposed method was word2vec [15], but many improvements have been made with models like fastText [16].

Topic extraction methods typically find dominant topics – those that are easily noticeable and popular. This task becomes difficult with short documents because of scarcity in word collocations, so additional modeling considerations and domain knowledge may be required. It would be useful to not only find dominant, but also non-dominant topics in such datasets. No listed work specifically addresses the issue of finding non-dominant topics within a set of short documents.

In this paper, we present a topic extraction method based on clustering fastText word embeddings with the OPTICS algorithm [17]. Each cluster corresponds to a topic, represented by a set of related tokens. The method proposes many logical topics which the user can

Received 8 March 2022
Accepted 21 March 2022

reasonably quickly to determine their relevance. It can handle short and medium-length texts with misspellings, which are very common on microblogging websites like Twitter and LinkedIn. We are able to discover dominant and non-dominant topics without using domain knowledge in the modeling process.

In Section 2, we describe the processes for generating embeddings with fastText and clustering them with OPTICS. We describe the evaluation and results in Section 3. We discuss the method's behavior, potential improvements, and future work in Section 4.

2 METHODS

The two main stages in our approach are generating token embeddings and clustering them. We first describe how fastText is used for token generation and then how OPTICS can propose potentially relevant topics by clustering token embeddings.

2.1 Generating token embeddings

A word embedding is a method for mapping a word from a set of documents to a real-valued vector. Because some words frequently appear together, we treat them more generally as tokens, which are sequences of characters. Similarity between tokens can be expressed as the Euclidean distance between their embeddings. This reveals which tokens are semantically or syntactically related.

We use fastText to generate token embeddings. FastText is a library for text classification and representation. It contains algorithms which transform text into continuous vectors that can later be used on language related tasks. We refer to models, built using this library, as fastText models.

These models can be trained in a matter of seconds compared to other models, which require several hours or even days.* The embedding quality is only slightly worse than that of state-of-the-art approaches. The fast training time facilitates experimentation and lets us analyze large datasets quickly. The models are trained using subword information about all character n-grams of length two or more. Such information is helpful, because our text originates from microblogging services, which are known to often contain misspellings. A correct and incorrect spelling of a word would otherwise be treated as entirely different, even though a person would understand they refer to the same concept. Similarly, this information also helps relate words with the same lemma or stem, which would otherwise need to be grouped using additional lemmatization or stemming steps. Considering subword information can hence alleviate collocation scarcity issues by closely relating similarly spelled tokens.

*https://research.fb.com/blog/2016/08/fasttext/, accessed March 1, 2022

2.2 Clustering token embeddings

We need to make some important considerations when choosing a clustering algorithm. The number of topics in a set of documents is usually unknown, so we either need to guess it or determine it automatically. It is likely that some embeddings do not belong to any topic. These may be treated as noise. The embedding space may be difficult to understand and algorithms which rely on finding clusters of specific shapes may be inappropriate. We assume that embeddings form dense regions if they refer to similar tokens, but the shape of the regions may be arbitrary. With these considerations, we choose the family of density-based clustering algorithms as an appropriate choice for embedding clustering. In particular, we choose OPTICS as the clustering algorithm because of its ability to find clusters of varying density and shape, as well as its ability to detect noisy embeddings.

OPTICS can be understood as an extended DBSCAN algorithm [18]. In DBSCAN, density-based clusters are defined as sets of density-connected objects. Put simply, density-connected objects can be reached from one another through a chain of objects which are less than ϵ apart. Each cluster contains core objects and border objects. Core objects have at least *MinPts* neighbors in their ϵ-neighborhood and border objects have fewer. The idea of OPTICS is to extend this concept so that several ϵ parameters are used simultaneously which means finding clusters of different densities. We can find such varying-density clusters consistently if we obey a specific processing order. OPTICS generates this order, but does not explicitly assign cluster memberships. The authors proposed an algorithm to automatically assign these using the ξ parameter. By plotting *reachability-distances* of ordered objects, we see clusters as dents in the plot as visualized in Figure 1. The ξ parameter determines the necessary steepness of a dent at its beginning and end so its objects can be treated as a cluster. The *noise* set are those objects, which do not appear in any detected dent. Higher ξ-values can find only the most significant clusters, whereas lower ξ-values can find less significant clusters at the expense of also finding more noisy ones.

Our method works by iteratively clustering token embeddings at different parameter settings of the OPTICS algorithm. We start with an initial value of *MinPts*, which helps us identify topics with the most tokens. In each iteration, we decrease this value by 1 to find topics with fewer tokens. We set the hyperparameter $\xi > 0$ and choose a threshold for the minimal cluster size. Topics with fewer tokens than this threshold will not be proposed. Selecting a small minimum cluster size and a small initial value of *MinPts* can reveal non-dominant topics, because they may contain fewer tokens than dominant topics.

The entire procedure is described in Algorithm 1. We

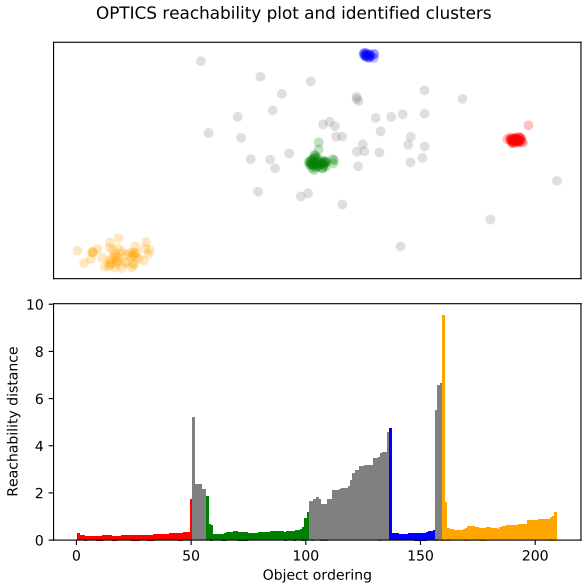

Figure 1.: Toy visualization of the OPTICS reachability plot. The colored points were identified as clusters, whereas the gray points represent noise.

associate tokens to their embeddings, so it is possible to represent a topic as a set of tokens in a cluster. We first cluster embeddings at an initial value of *MinPts*. The identified clusters are stored as sets of topic tokens. We remove embeddings which belong to such clusters and only keep noisy ones. We then decrease the value of *MinPts* by 1 and repeat the clustering process on noisy embeddings. The algorithm stops once the value of *MinPts* falls below 2. We remove clustered embeddings to make sure they do not appear again in future iterations. The user then determines which of the proposed topics are relevant.

Algorithm 1: Topic extraction with OPTICS.

Input: Set of embeddings E, initial *MinPts* value m.
Result: Set of proposed topics P.
Let P be the set of proposed topics.
while $m \geq 2$ *and* $E \neq \emptyset$ **do**
  Cluster E with OPTICS(*MinPts* $= m$).
  Move found topics from E into P.
  $m \leftarrow m - 1$
end

3 Evaluation and Results

In this section, we describe our evaluation and the dataset we used to test our method. Our dataset consists of online posts about industrial automation. This choice was motivated by a business task from Siemens, where the goal was to create an overview of trending topics for the industrial automation sector using open media data. We use our approach to find topics within the posts and compare the results to some baselines. Finally, we list some interesting identified topics and provide a visualization of these.

3.1 Evaluation dataset

We obtained a dataset of text documents relating to industrial automation. These are posts from technological companies on Twitter, LinkedIn, and their blogs. The list of these companies was provided by industrial automation experts at Siemens. The content mainly consists of current and upcoming events and fields, technological and financial progress updates, and internal company events. Most posts were made between the years 2010 and 2020. By manually checking some randomly chosen documents, the most noticeable topics are sustainability, safety, maintenance, climate, smart technology, robotics, and artificial intelligence.

We use standard text pre-processing techniques to transform the documents into a more suitable form for model training and analysis. Each sentence in a document is represented as a sequence of tokens, a convention for use in many text analysis algorithm implementations. We summarize the pre-processing procedure in Figure 2.

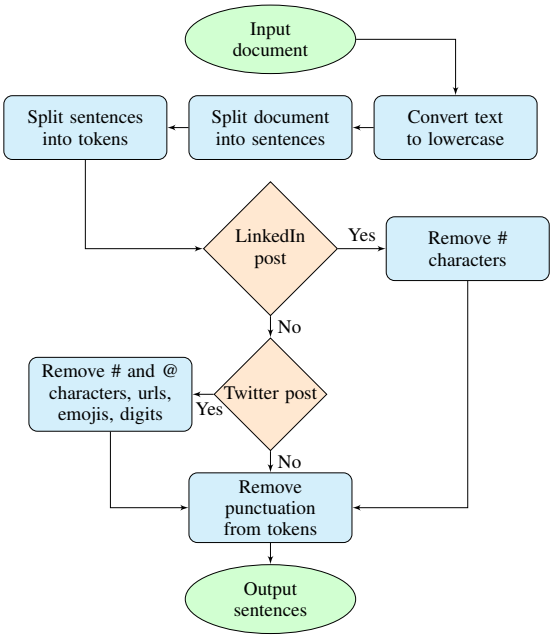

Figure 2.: Document pre-processing flowchart for blog posts, Twitter posts, and LinkedIn posts.. The input is a text document, the output is a list of sentences where each sentence is a list of tokens.

We first transform documents to lowercase. We perform tokenization with the *TweetTokenizer*, *PunktSentenceTokenizer*, and *TreebankWordTokenizer* classes

from the *nltk* library [19]. We split Twitter posts into tokens using the *TweetTokenizer* class. We split other documents into sentences with the *PunktSentenceTokenizer* class and transform them into tokens using the *TreebankWordTokenizer* class. Each Twitter and LinkedIn post corresponds to a single sentence, but blog posts may consist of several sentences. We do not retain any punctuation-only tokens. We remove "#" characters from LinkedIn and Twitter posts. We also remove mentions (tokens that start with "@"), url tokens, emojis, and digits. We remove all remaining punctuation within tokens and also remove zero-length tokens that are generated as a result of these transformations.

The result is a collection of sentences where each sentence is a sequence of tokens. We treat these sentences as individual documents. The final dataset contains 32831 documents – 14077 Twitter posts, 14916 LinkedIn posts, and 3838 posts from company blogs.

3.2 Evaluation methodology

Our evaluation is qualitative, which is not uncommon in topic modeling and clustering. We justify this with the following observations:

- there is no labeled text dataset meant for finding non-dominant domain-specific topics,
- many tokens do not belong to any topic, so metrics requiring class balance are inappropriate,
- some sought topics are non-dominant, so metrics making use of frequencies are inappropriate,
- our method is not probabilistic, so metrics such as perplexity are inappropriate,
- the choice of relevant topics is subjective,
- the space of embeddings is not well-understood.

We will evaluate our method by manually checking the list of proposed topics. Most of the proposed topics should contain semantically or syntactically related tokens, but there may also be some unrelated ones. Depending on the dataset and the user, we also wish to see relevant topics which vary in terms of dominance. Besides qualitatively checking to what extent these criteria are satisfied, we will also propose some goal topics. These are topics which we expect are present in our dataset based on checking a random sample of documents. Our method should be able to identify these goal topics.

3.3 Settings and baseline for comparison

The trained fastText model produces 40-dimensional token embeddings. When using this dimensionality in combination with our dataset, we saw dominant and non-dominant topics which are relevant and have related tokens. Training is performed using the skip-gram model [20] and hierarchical softmax. These are beneficial, because they focus on infrequent words which may form domain-specific topics. We set the initial *MinPts* value to 10 and the minimum cluster size value to 3. We set the ξ parameter to 0.05 and use the Euclidean distance to determine core and boundary embeddings.

We use LDA, LSA, and HDP models as baselines for comparison. To achieve the best baseline performance, we first transform tokens with the *WordNetLemmatizer* class [21] from the *nltk* library. This transforms most words into their base form, which is desirable when dealing with bag-of-words-based representations. We remove stopwords for topic extraction with the baseline methods, but this is not necessary to achieve good performance with our method. We train LSA with TF-IDF features. We train LDA and HDP with standard bag-of-words features. We use the fastText wrapper, LDA, LSA, and HDP implementations from the *gensim* library [22] and the OPTICS implementation from the *scikit-learn* library [23].

3.4 Identified topics

We used our method to extract topics from the evaluation dataset. The identified relevant topics are listed in Table 1. We also list a sample of the proposed LDA topics in Table 2 for comparison. The full results are provided in the appendix. The LDA and LSA methods are able to identify a few relevant tokens, but almost no topics. HDP proposes topics with more interesting, but unrelated tokens. Examples of such tokens are "#PowerGrids", "#GenderEquality", "fossil-fueled", and "dewatering". We also compare methods with respect to their ability to identify selected goal topics in Table 3. Our method identifies dominant and non-dominant topics, which are more relevant than the baseline topics and contain tokens that are more closely related.

Some of the identified topics are immediately noticeable in the dataset, namely the ones relating to sustainable energy, robotics, and climate change. However, the e-voting, fish farming, and bioprocessing topics appear rarely in the dataset and a person needs to invest significantly more time into reading the documents to identify them manually.

We also observed the proposed topics in different iterations of the algorithm. We find that most of the relevant topics are identified in later iterations of the algorithm, when the value of *MinPts* is small. This is visualized in Figure 3. Earlier iterations proposed topics whose tokens were mostly links, numbers, financial terms and figures, and non-English words. Later iterations proposed more topics. Most of those were irrelevant or had unrelated tokens, but there were also relevant topics, whose tokens were related.

3.5 Topic visualization

We can present the identified topics in a way that reveals how they are related to each other. Since the underlying embeddings are high-dimensional, it is standard practice to visualize them by using dimensionality

EV, chargers, #EVCharging
electric, charging, stations, charger
hybrid, vehicles, buses, vehicle
gas, oil, natural, LNG
solar, PV, rooftop
wind, farm, turbine, turbines
water, treatment, wastewater
#WaterQuality, #SurfaceWater, #FishFarming
hydraulic, filtration, roller
drilling, drill, saws
thermostat, radiator, thermostats
bioprocesses, single-use, bioprocessing
Cobots, #MobileRobots, #CollaborativeRobots
IIoT, #EmpowerTheField, IoE
#SmartHome, #VoiceActivated, #IFA15
SJI, #PCBdesign, circuitboard, PCB
CX5230, #UltraCompact, #IndustrialPC, C6030
#EnergyTransition, #NetZeroCarbon
#ClimateAction, neutral, #ClimateChange
#ZeroWaste, #GlobalRecyclingDay
#CarbonNeutral, #CarbonNeutrality, neutrality
#EmissionsFree, ship, emissions-free
#FoodWastage, #FoodLoss, #FoodWaste
e-voter, #VotingElectronic, e-voters
inclusive, workplace, culture
#WomenInScience, #GenerationEquality

Table 1.: Sample of relevant proposed topics, identified by our method. Rows correspond to topics, represented by their tokens. For ease of understanding, we present tokens in their original instead of their pre-processed form. The topics contain related tokens and are relevant in the industrial automation domain. They are mostly about electric vehicles, energy, industrial equipment and technology, robotics, the internet of things (IoT), climate change, and workplace culture.

IoT, good, installed, innovation, questions
January, experience, fair, panel, robotics
learn, service, watch, offerings, solution
stand, SPS, hall, sign, accelerate
earnings, global, industries, increased
general, operations, best, exchange, article
close, support, application, download, benefits
future, three, key, role, include

Table 2.: Sample of proposed topics, identified by LDA. Rows correspond to topics, represented by their tokens. The topics are uninformative. There are some useful keywords like "IoT" and "robotics", but they are not related to other tokens in their topics.

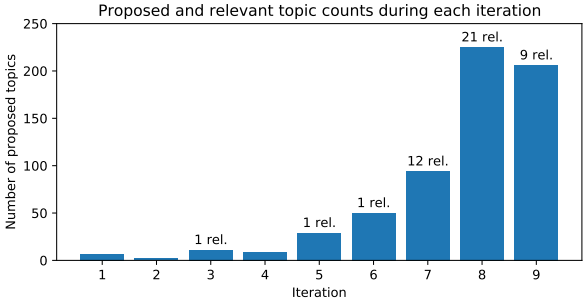

Figure 3.: The number of proposed and relevant topics per iteration when processing the evaluation dataset. The starting value of *MinPts* is 10 and decreases by 1 in each iteration. Only three relevant topics are identified in the first six iterations. Most of the relevant topics are found in iteration 8 when *MinPts* is set to 3.

reduction methods. We do so with Isomap [24], a non-linear dimensionality reduction technique that extends upon multi-dimensional scaling [25]. This is appropriate, because token embeddings likely contain essential nonlinear structures, which are invisible to linear dimensionality reduction techniques. The rough preservation of distances is important for understanding similarity between topics. We manually group the identified topics into nine broader categories to facilitate visualization and interpretation, as well as deal with topic overlap. We visualize the topics in Figure 4.

4 Discussion

We proposed a method of clustering fastText token embeddings with OPTICS on a dataset of short online industrial automation text documents. The method outperforms standard topic extraction techniques by proposing dominant and non-dominant domain-specific topics. The topics consist of semantically related tokens, so they can be easily understood by a person.

A key observation and requirement is that similar token embeddings form dense regions. This criterion can be met by embedding-based language models. The use of fastText in particular facilitates this even further, because it can place similar embeddings close to each other solely based on their character-level similarity. In the extreme case of single-token documents, this would result in topics of similarly spelled words. On the other hand, having a set of documents with appropriate word collocations would combine word and character-level features to produce more informative topics. We can reduce the impact character-level similarity by increasing the minimum length of n-grams which should be considered in the fastText model training.

A dataset may contain non-dominant relevant topics. There will not be many tokens which belong to such

Goal topic	Electric vehicles	Climate change	Robotics	Clean energy	Smart technology
LDA	EV, car, talk	increase, carbon, footprint	AI, factory, robotics	/	IoT, IIoT, cost
LSA	/	/	robotics, safety, robot	project, digitalization, #EnergyEfficiency	IoT, good, #EnergyEfficiency
HDP	/	/	/	energy, production, international	IoT, solutions, technology
Our method	EV, chargers, EV charging	neutral, #ClimateAction, #ClimateChange	robots, cobots, #MobileRobots	#GridEnergy, #NetZeroCarbon, #EnergyTransition	IFA15, #SmartHome, #VoiceActivated

Table 3.: Topic extraction results on the industrial automation dataset. We selected five goal topics which we observed in a small random sample of documents and compared how well each method can identify them. The LDA, LSA, and HDP methods proposed 100 topics each. Each cell contains the three most appropriate tokens of the best proposed topic with respect to a goal topic. The baseline methods were sometimes unable to propose an appropriate topic; such cells are marked with "/". All methods produced more than three tokens per topic. In baseline methods, all unlisted tokens were completely unrelated to the goal topic. Our method identified more relevant topics than individual baseline methods and also produced related tokens (listed and unlisted).

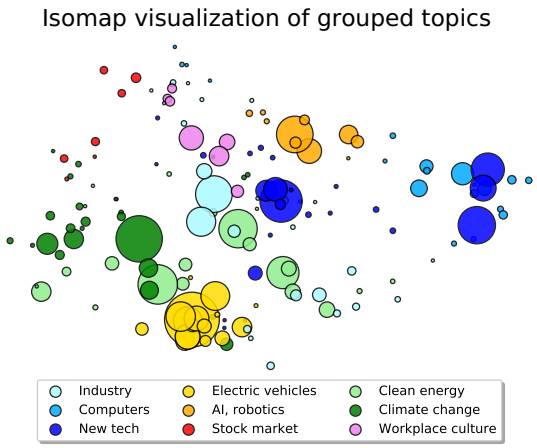

Figure 4.: Isomap visualization of the identified topic groups. Topic tokens are represented as circles with size corresponding to their frequency in the corpus. The plot shows topic importance and inter-topic relationships between the years 2010 and 2020 according to industrial automation companies. For example, the *clean energy* and *climate change* topics are close to *electric vehicles*, representing the sentiment that electric vehicles mean a step towards a cleaner world. The *AI, robotics*, and *workplace culture* topics are close together, possibly due to recent discussions of AI replacing human workers or simply incorporating AI in the standard industrial workflow.

topics. As a result, these will be found at small values of *MinPts* and the minimum cluster size clustering hyper-parameters. However, the user will also have to inspect many other, potentially irrelevant topics with a small number of tokens. In our experimentation, the highest number of produced topics was 225 at $MinPts = 2$, so checking the proposed topics was manageable. However, the number of proposed topics may be higher and the issue needs to be further considered to avoid sifting through too many irrelevant topics. A small improvement could be made by removing stopwords before clustering, but this may affect token neighborhoods and cause some small clusters not to be detected.

The OPTICS algorithm also outputs a hierarchy of clusters. In our dataset, this hierarchy was not noticeable from the reachability distance plot. We can inspect the generated hierarchy for a particular value of *MinPts*, but our method would need an additional procedure to combine hierarchies at different values of *MinPts* into a single one. We also tried running OPTICS once (not iteratively) with $\xi = 0.01$ and observed that there were not as many relevant topics as when using the original method. Such an approach did not produce a clear hierarchy, but it is possible that it would do so on a different dataset where word-level similarity is more significant than in short online posts.

An interesting direction for future work is using existing pre-trained language models instead of training one from scratch. This means having better embedding spaces which allow for topic interpretations with a hierarchy. It would be useful to reduce the number of irrelevant proposed topics as much as possible, likely

by using domain knowledge. It would also be useful to evaluate the method quantitatively. A suggestion for this is to consider a dataset and ask different users to select subjectively relevant topics from the list of proposed topics. This can be done at different hyperparameter settings and the user-specified topics may serve as targets in numeric evaluation. The method performs well on industrial automation posts, but testing it on data from different domains would help further assess its performance.

ACKNOWLEDGEMENTS

Our research was partially supported by the Slovenian Research Agency (ARRS) research core funding P5-0410.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[3] Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

[4] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88, 2010.

[5] Yuan Wang, Jie Liu, Yalou Huang, and Xia Feng. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1919–1933, 2016.

[6] Chi-Yu Liu, Zheng Liu, Tao Li, and Bin Xia. Topic modeling for noisy short texts with multiple relations. In *SEKE*, pages 610–609, 2018.

[7] Min Shi, Jianxun Liu, Dong Zhou, Mingdong Tang, and Buqing Cao. We-lda: a word embeddings augmented lda model for web services clustering. In *2017 ieee international conference on web services (icws)*, pages 9–16. IEEE, 2017.

[8] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[9] B. Shen and Yingsi Zhao. Optimization and application of optics algorithm on text clustering. *Journal of Convergence Information Technology*, 8:375–383, 2013.

[10] Ahmed Rafea and Nada A. Mostafa. Topic extraction in social media. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 94–98. IEEE, 2013.

[11] Kazuma Hashimoto, Georgios Kontonatsios, Makoto Miwa, and Sophia Ananiadou. Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of Biomedical Informatics*, 62:59–65, 2016.

[12] Suraj Subramanian and Deepali Vora. Unsupervised text classification and search using word embeddings on a self-organizing map. *International Journal of Computer Applications*, 156:35–37, December 2016.

[13] Xiangfeng Dai, Marwan Bikdash, and Bradley Meyer. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon 2017*, pages 1–7. IEEE, 2017.

[14] Guilherme Raiol De Miranda, Rodrigo Pasti, and Leandro Nunes de Castro. Detecting topics in documents by clustering word vectors. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 235–243. Springer, 2019.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[16] Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas. Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics*, pages 135–146, 2017.

[17] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[18] Martin Ester and Hans-peter Kriegel and Jörg Sander and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *AAAI Press*, pages 226–231, 1996.

[19] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[21] George A. Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[22] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[25] Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

David Nabergoj received his M. Sc. degree in 2021 from the University of Ljubljana, Faculty of Computer and Information Science. He is currently researching normalizing flows and Bayesian statistics as part of a group in the Berkeley Center for Cosmological Physics.

Alessandro D'Alconzo is a senior data scientist at Siemens AG, Vienna. He received his PhD degree in Information Engineering from the Polytechnic of Bari (Italy), in 2007. His main research interests are in the field of network traffic monitoring and application of machine learning to industrial, and business analytics domains.

Danilo Valerio is a senior research scientist at Siemens Technology, Vienna. He received his PhD degree in Computer Science from the Faculty of Informatics in Vienna. His main research interests are in the field of machine learning in the domains of urban, industrial, and sales analytics.

Erik Štrumbelj is an associate professor at the Faculty of Computer and Information Science, University of Ljubljana. His main research interests lie in Bayesian statistics and machine learning.

APPENDIX

We provide a comprehensive list of identified topics for our method and baseline methods: latent semantic analysis (LSA), latent Dirichlet allocation (LDA), hierarchical Dirichlet process (HDP). We present the results for our method in Table 4 and the results for LDA, LSA, and HDP in Tables 5, 6, and 7 respectively. Tokens are presented in their original form for ease of understanding. The actual output tokens are in lowercase and without "#" characters.

EV, chargers, #EVCharging
cars, car, batteries
electric, charging, stations, charger
turbochargers, turbocharger, turbocharging
supercharger, supercharging, superchargers
hybrid, vehicles, buses, vehicle
#HybridElectric, E-Fan X, E-plane
gas, oil, natural, LNG
solar, PV, rooftop
wind, farm, turbine, turbines
SWT, #WindTurbines, #WindTurbine
water, treatment, wastewater
#WaterQuality, #SurfaceWater, #FishFarming
hydraulic, filtration, roller
drilling, drill, saws
thermostat, radiator, thermostats
Cobots, #MobileRobots, #CollaborativeRobots
collaborative, robot, LD
intelligence, artificial, nanotechnology
e-voter, #VotingElectronic, e-voters
Dow, Jones, indices, #RobecoSam
Nasdaq, quoted, OMX
IIoT, #EmpowerTheField, IoE
#PCbased, PCs, PC, PC-based
SJI, #PCBdesign, circuitboard, PCB
#EnergyTransition, #NetZeroCarbon, #GridEnergy
#ClimateAction, neutral, #ClimateChange
#ZeroWaste, #GlobalRecyclingDay
#CarbonNeutral, #CarbonNeutrality, neutrality
#EmissionsFree, ship, emissions-free
EP100, #NetZero, #CO2neutral
#EUEnergyDay, #SustainableDevelopment
#FoodWastage, #FoodLoss, #FoodWaste
friendly, efficient, environmentally, #EnergyEfficient
#PanelBuilder, #PanelBuilding
inclusive, workplace, culture
#WomenInScience, #IWD2020, #GenerationEquality
Boost40, TTTech, cybernetics
CX5230, #UltraCompact, #IndustrialPC, C6030
SPS, drives, IPC
#SmartHome, #VoiceActivated, #IFA15
Russwurm, #WebOfSystems, WOS
biomethane, #EndressHauser, biomethan
bioprocess, bioprocesses, single-use, bioprocessing

Table 4.: All topics that were identified by our method. Rows in both tables correspond to topics, represented by their tokens. The tokens of each topic are related and many of them are hashtags (tokens that start with the "#" character). Some topics are semantically similar to others.

share, located, deliver, focus, construction
read, ensure, addition, designed, July
acquisition, position, website, operating, latest
local, costs, leading, reduce, receive
forward, flow, cost, country, intelligent
day, help, decreased, ready, change
control, software, renewables, center, robots
annual, vice, president, Europe, easy
existing, safety, great, discover, higher
engineering, approx, total, potential, efficient
handling, services, large, plants, orders
binding, return, long, save, standards
order, plant, capacity, people, Australia
register, join, webinar, facility, continue
tonnes, year, received, scope, full
February, source, electric, range, cooperation
learn, service, watch, offerings, solution
stand, SPS, hall, sign, accelerate
earnings, global, industries, increased
group, capital, CEO, business, EBIT
india, increase, happy, changes, core
contribute, signed, awarded, advanced, enable
product, growing, quality, high, unit
January, experience, fair, panel, robotics
Copenhagen, effective, start, friday, efficiency
supply, China, work, electrical, link
international, IPC, technical, option, twincat
shares, blog, based, products, increasing
head, field, type, aging, library
supervisory, report, board, group, executive
IoT, good, installed, innovation, questions
financial, processing, place, drive, industry
contract, well, largest, strengthen, years
process, design, project, training, countries
booth, find, presented, June, infrastructure
general, operations, best, exchange, article
close, support, application, download, benefits
future, three, key, role, include
production, press, release, Yokogawa, projects
offer, free, working, exclusive, #LinesShare

Table 5.: Sample of topics, identified by LDA. Rows in both tables correspond to topics, represented by their tokens. Most topics are completely uninformative. There are some useful keywords like "IoT" and "robotics", but they generally unrelated.

–, address, questions, co, a/s
visible, LinkedIn, –, join, follow
authority, supervisory, Danish, financial
learn, read, hall, booth, solutions
thx, retweets, favs, follow, learn
learn, read, hall, booth, find
read, hall, booth, forward, stand
hall, booth, read, energy, 's
's, find, DKK, energy, order
find, 's, DKK, energy, solutions
DKK, energy, 's, solutions, industry
's, DKK, energy, solutions, safety
forward, stand, hall, year, future
future, plant, power, industry, forward
video, watch, solutions, find, industry
booth, stand, forward, year, SPS
future, solutions, power, energy, industry
technology, solutions, power, safety, control
industry, solutions, future, power, year
technology, year, register, plant, free
industry, forward, register, year, booth
register, join, technology, solutions, forward
day, SPS, safety, report, help
safety, IIoT, digital, join, technology
control, help, report, annual, year
activate, larger, view, link, year
control, technology, year, day, activate
hannover, SPS, drives, IPC, power
safety, check, hannover, help, register
Hannover, help, join, power, stand
#LifeIsOn, check, digital, join, IoT
digital, help, safety, check, weg
ready, join, register, business, digital
join, technology, plant, register, control
digital, check, control, ready, lifeison
EcoStruxure, digital, check, day, business
help, order, check, business, join
business, work, digital, ready, people
weg, ready, production, register, robotics
ready, EcoStruxure, day, people, free

Table 6.: Sample of topics, identified by LSA. Rows in both tables correspond to topics, represented by their tokens. Most topics are completely uninformative as in LDA. Many tokens appear in several topics.

periscope, tonnages, percent, grandir
Leterrier, anticipare, targets, #Expo2020
industry-leading, 's, #VoiceActivated, ontvang
negative, #WOTC19, topped, skier
–, #DigitalEngineering, elevate, supervisory
proceso, GDSN, Dornbirner, ZRH
Pinterest, #ThinkTank, #ITxpo, sushi-sensor
Sydney, banana, Venkat, beruf
sequences, arena, anderen, eigentlich
busters, Caxitu, Turbocor, APIs
Santos, core, software-suite, 17
stick, hin, te, abreast
counterpart, walk, #BeeBetter, banken
Jurvetson, runden, entre, genau
invalides, group, dear, anticipare
plant, foothold, –, #SqueezeOut
intracellular, bronze, IoTSWC17, innovative
wenige, drawn, garantendoti, higher
Goodknecht, start, kijken
clothes, dijital, Canadian, roses
functionality, #PowerGrids, Lalbagh, #IndustrialDrives
#EPlanInfoCenterApp, space-savings, learn, Johnson
awaken, terry, sophisticated, Adelaide
Denice, #EatonEngaged, AFDB, signature
railways, #DeviceConfiguration, Pearson, frohe
cyanidation, self-diagnosis, oprex, #GenderEquality
withholding, wet, fossil-fueled, cons
Biraschi, uninsured, seated, provisions
host, svenska, Udo, AICCE
significative, bridge, tube, revisit
inaction, AP02, stabilise, dewatering
//eh.digital/2ldlg65, #EarlyDetection, #UseCases, oyu
proposal, booth, Stahl, built
#WorkTruckWeek, electrical-, arguably, visible
Highgrade, kennt, offering-, Brullon
WBCSD, dreiteiliger, //eh.digital/2ngktwn, ayudan
#EndressHauser, #DigitalTag, retweets, learn
#WeAreStillIn, indutry, #AutoID, shatter
future, foothold, #vision2020, Nooriabad
intuitive, weir, pitfalls, revitalize

Table 7.: Sample of topics, identified by HDP. Rows in both tables correspond to topics, represented by their tokens. While the topics are not informative, they do contain some relevant tokens, such as "#PowerGrids", "#GenderEquality", "fossil-fueled", and "dewatering".