Towards Hypothesis-driven Forensic Text Exploration System

Jenny Felser*, Dirk Labudde†, Michael Spranger‡
Mittweida University of Applied Sciences
09648 Mittweida, Germany
Email: *felser@hs-mittweida.de, †labudde@hs-mittweida.de, ‡spranger@hs-mittweida.de

***Abstract*—Short messages stored on mobile devices have become a crucial source of evidence in criminal investigations. However, the high volume of chat messages poses a challenge to the investigator. Topic modelling offers the potential to summarise the short messages compactly, thus effectively supporting the investigator in exploring the vast number of chat messages. This paper presents our preliminary work towards developing a forensic text exploration system based on topic modelling approaches. The two goals typically pursued by the investigator when exploring chat messages are to be supported. On the one hand, the investigator often already has a hypothesis about specific topics discussed in the chats and wants to find evidence. On the other hand, the investigator also wants to discover new topics and connections. Accordingly, in this work, we investigated unsupervised and semi-supervised approaches based on Latent Dirichlet Allocation (LDA) with the additional use of word embeddings. Overall, the evaluation of different methods using actual case data showed that the semi-supervised approach, combined with word embedding similarity, can find qualitatively better topics than unsupervised topic modelling approaches based on LDA.**

***Index Terms*—topic modelling; forensic text analysis; semi-supervised; hypothesis-driven analysis.**

I. Introduction

Nowadays, the analysis of short messages stored on mobile devices is an important part of forensic investigations. However, the high number of messages can also prove challenging for the investigator. Often, a single mobile phone stores more than 15,000 Short Message Service (SMS) and 150,000 messages from messenger services [1]. Furthermore, especially in the case of gang crime and organised crime, it is often necessary to examine the short messages of several mobile phones [1]. To assist the investigator in exploring the chat messages, the application of topic modelling is suggested. This allows to get an overview of the contents discussed in the messages and to summarise the messages as compactly as possible.

Topic modelling should best support both goals that investigators are pursuing when analysing forensic chat messages: On the one hand, investigators usually have some presumption about topics that have been discussed in the messages. Usually, at least they know the area of offence their case is about. In addition, they can obtain information about the circumstances of the offence from interrogations [2] or the case file [2]. Accordingly, one goal is to find evidence in the data for a certain hypothesis, respectively, that a topic was actually discussed in it. On the other hand, the investigators also want to discover new topics, for example about the motivation of the crime or previously unsuspected connections to certain people.

The basic aim of this paper is to investigate some methods of unsupervised topic modelling for the first scenario and semi-supervised topic modelling for the second scenario and to qualitatively evaluate and compare the results. More specifically, the unsupervised method used is weighted Latent Dirichlet Allocation (wLDA), as described by Wilson and Chew [3], while the keyword-Assisted Topic Model (keyATM) developed by Eshima et al. [4] was chosen as the semi-supervised method. In addition, an extension of keyATM is proposed based on a combination with the Cluster Words (CluWords) document representation presented by Viegas et al. [5], which additionally includes word embeddings.

The paper is organized as follows: At first, some related work is presented in Section II. Then, an overview of the data and methods is provided in Section IV. The experimental results are presented and discussed in Section V. Finally, Section VI concludes the paper.

II. Related Work

To the extent of our knowledge, topic modelling has only been used by a few works in the field of forensics with the aim of compactly summarising data sets in the context of forensic investigations [6]–[9]. Furthermore, they also did not focus specifically on communication data. Instead, de Waal et al. [8] extracted topics from all textual data that need to be investigated for a case, including emails and notes in text documents, while Noel and Peterson [9] used Word documents extracted from a hard disk store as the data basis. Both works [8], [9] applied the Latent Dirichlet Allocation (LDA), as described by Blei et al. [10], as the algorithm for topic modelling. Moreover, topic modelling was used by Li et al. [7] and Busso et al. [6] to support data exploration in specific offense areas or in the analysis of concrete cases. Li et al. [7] tried to uncover various topics in conversations about corruption on Twitter using the Biterm Topic Modeling (BTM) algorithm introduced by Yan et al. [11]. Moreover, Busso et al. [6] applied the Structural Topic Model (STM), as described by Roberts et al. [12], to identify topics in a series of racist and offensive letters. Thus, in summary, unsupervised probabilistic generative models such as the LDA and its extensions were mainly applied to forensic texts. These are suitable for the

 ISBN: 978-1-68558-089-6

second mentioned scenario regarding the analysis of forensic short messages, namely for the data exploration.

However, to the extent of our knowledge, no previous work in the forensic field has focused on the scenario where the investigator is looking for evidence of certain assumed topics in the data. The problem with unsupervised approaches is that they are not able to identify topics of interest to the investigator if they are present in the dataset only to a small extent [13], as is often the case in forensic communication data due to the prevalence of irrelevant small talk.

This problem can be addressed by incorporating the investigator's prior knowledge into the topic modelling process, for example, by using supervised approaches, e.g., [14], [15]. However, these require annotated training datasets to learn known topics. One approach to create an annotated dataset would be to collect as much case data as possible from different offence areas and label the messages with the known offence as their topic. Yet, legal questions in the respective country would first have to be clarified as to whether the merging of data from different cases is permissible. Instead, semi-supervised approaches come into consideration, which differ in the type of user input they integrate, e.g., [16]–[18]. For example, user feedback on the relevance of topics [19]–[21], information about thematic relationships between word pairs, e.g., [16], [22], [23] or user knowledge about known topics in the form of a few characteristic terms was included in topic modelling, e.g., [4], [17], [24]. The last case is most suitable for finding evidence for suspected topics. In these approaches, a distinction can be made between Targeted Topic Modelling, e.g., [17], [25], [26] and Seed-Guided Topic Modelling, e.g., [4], [24], [27].

Algorithms of Targeted Topic Modelling aimed at extracting fine-grained topics related to a specific aspect described by a single characteristic word, e.g., [17], [25], [26]. In forensic context, these approaches could be used to find different subtopics dealing exclusively with the crime under investigation, such as drug crime. The basic idea of these algorithms was to reduce the dataset to documents [17], [23], [25], word pairs [28] or words [26] that were relevant to the aspect, whereby the relevance determination was carried out with reinforcement learning [23] or based on external corpora [26], for example. However, especially when determining relevance at the document level, these approaches are accompanied by the risk that important case-relevant information can be lost if incorrectly classified as irrelevant.

In contrast, seed-guided topic modelling approaches, especially probabilistic generative models, e.g., [4], [13], [29], may be promising. Unlike other semi-supervised methods, e.g., [30], these have the advantage that they can overcome prior knowledge, if the desired topic, described by some relevant seed words, does not appear in the dataset at all, e.g., [13], [29], which is why these approaches are particularly suitable for testing hypotheses in a forensic context.

So far, however, semi-supervised probabilistic approaches have been applied and evaluated mainly on long, linguistically correct texts such as draft legislations [4] and customer reviews [31], [32]. An important contribution of this work is therefore to investigate the suitability of topic modelling for identifying case-relevant topics in forensic communication data, despite their particular challenges, such as their short length and low linguistic quality [33]. Furthermore, it is one of the first studies to include both goals, exploration and finding evidence, in topic analysis of forensic texts.

III. Data

For all experiments, WhatsApp messages from a real case about the financial support of a terrorist group, which were stored on the mobile phone of a suspected person, served as the data basis. The dataset is not publicly available, but was provided to the authors by a cooperating prosecutor for research purposes and has already been used in previous work [34]. The messages were exchanged in 146 chats between mid-December 2014 and mid-May 2019. The total of approximately 118,000 text messages in the data set were primarily in German and to a lesser extent in Turkish and Arabic. Since the focus of this work was on monolingual topic analysis, approximately 106,000 German messages were extracted by automatic language detection using Google's cld2 [35] and cld3 [36] models. Details on the vocabulary size, the number of unique tokens (besides words also punctuation marks, symbols, numbers and web links), the average frequency of words and the length of the messages can be taken from the upper section of Table I.

TABLE I
Statistical description of the data set used

Property/ Statistic	Result
vocabulary size (# unique words)	36467
# unique tokens	39039
average frequency of words	22.62
∅ message length (in words)	7.75
# conversations	15625
∅ number of messages per conversation	6.81
∅ conversation length (in words)	52.78

As can be seen from the table, the messages contained on average less than eight words including stopwords. Since the short length of the messages poses a known problem for topic modelling [37], messages that occurred in a common temporal context were aggregated into related conversations, as explained by [33], which were subsequently considered as one document. The formation of conversations was carried out with the Mobile Network Analyzer (MoNA), a forensic tool for analysing mobile communication data [1]. Information about the number of conversations, the average number of messages that made up a conversation and the length of the conversations can be found in the bottom section of Table I.

IV. Methods

In order to find suitable approaches for both scenarios of forensic data analysis, exploration and hypothesis testing, initial experiments on unsupervised and semi-supervised topic

 ISBN: 978-1-68558-089-6

modelling were carried out. All algorithms were trained on the conversation documents described in Section III. The preprocessing as well as the training of the topic models was conducted with the statistical software R.

A. Preprocessing

Before performing topic modelling, extensive preprocessing was applied to these conversation documents, which, as shown by Churchill and Singh [38], is essential for good results in topic modelling, especially with noisy data such as forensic short messages. This included the performance of the following cleansing steps:

1) Removal of redundant whitespace
2) Removal of web links, email addresses, and mentions, as they did not contribute to the content
3) Removal of emojis, as they usually have little meaning without context in topic-word distributions
4) Removal of punctuation marks and then numbers
5) Removal of German, Turkish and English stopwords using the stopword lists provided by Diaz [39]. The removal of English and Turkish stop words was necessary despite the reduction of the data set to German messages, as it could not be excluded that Turkish idioms or anglicisms were used in messages classified as German.
6) Removal of the 100 most frequent words and the 100 words with the lowest Inverse Document Frequency (IDF)
7) Removal of all modal and auxiliary verbs as well as the most common German verbs manually selected from [40]
8) Conversion to lower case
9) Lemmatisation using the TreeTagger [41], [42], in particular to reduce the high sparsity of the communication dataset by decreasing the vocabulary size [43]
10) Tokenisation in unigrams

B. Unsupervised topic modelling

wLDA [3] was chosen as unsupervised approach. This algorithm differs from the standard LDA by integrating a term weighting scheme based on Pointwise Mutual Information (PMI) [44] into Collapsed Gibbs Sampling [45], which penalises terms that occur in many documents and are often not meaningful in topics. The term weighting scheme was used in addition to the stop word removal in order to prevent irrelevant high-frequency words, typical for colloquial texts [46], from dominating the topics. Using the cleaned conversation documents as input, the topic model was trained over 1,500 iterations, where the hyperparameters α as prior for the document-topic-distribution and β as prior for the topic-word distribution [4] were set to 0.08 and 0.01. The number of topics was set to 13 in accordance with the semi-supervised approach, which is explained in the following subsection.

C. Semi-supervised topic modelling

As a semi-supervised method, the keyATM model, as described by Eshima et al. [4], was chosen because it extends wLDA and, accordingly, unlike other seed-guided topic modelling algorithms, gives less weight to uninformative words when estimating topics. The basic idea of keyATM consists in the introduction of an additional topic-word distribution containing only seed words [4].

For each desired topic, set of seed words were created based on the so-called term tree explained by Spranger et al. [1], which describes a complex system of syntagmas, referring to case-relevant terms that occur together in a conversation. Each syntagm was considered as a set of seed words. The term tree was created semi-automatically by expanding case-relevant words provided by the prosecutor in charge of the case with further relevant words using a suggestion system of the software MoNA [1], [34]. Each syntagm respectively each set of seed words included case-relevant terms, their synonyms, spelling variants and words that are syntagmatically related to the case-relevant terms provided. As an example, a selection of terms from three out of eight seed word sets is presented in Table II. Notably, keyATM enables the specification of a seed word as a topic label before fitting the model [4]. Throughout this table and subsequent ones, English translations of terms are provided in parentheses.

TABLE II
SELECTED TOPIC LABEL AND EXAMPLES OF USED SEED WORDS FOR SEMI-SUPERVISED TOPIC DETECTION WITH KEYATM.

Topic Label	**Seed Terms**
Geld (money)	Euro, überweisen (transfer), Zahlung (payment)
Terror	Waffe (weapon), Anschlag (attack), Gewalt (violence)
Verein (association)	Vereinsregister (association register), rechtsfähig (judicable), Vereinsgründer (association founder)

With the created seed word sets, keyATM was trained on the cleaned conversation documents, where the hyperparameters α and β and the number of iterations were set to the same values as for the training of wLDA, as described in Section IV-B. Suplementary, specific hyperparameters for keyATM were set to the default values as suggested in the reference paper by Eshima et al. [4]. In addition to the eight seed topics, keyATM enables to find a predefined number of unseeded topics, in this case five, which mainly serve as residual topics to bundle unimportant words together [4], [24].

D. Semi-supervised topic modelling with CluWords

keyATM already aims to ensure that the seed words and their related words have a high probability in the desired topic [4]. However, this requires that the words co-occur with the seed words in documents [4], [29]. To ensure that words that are semantically very similar to the seed words are assigned high probabilities in the corresponding topic, regardless of their co-occurrence frequency, keyATM was extended with an adapted CluWords document representation, originally proposed by Viegas et al. [5]. A CluWord is defined as a set of words that have a high word embedding similarity to a term [5]. The basic idea of the approach is to insert

CluWords into the original conversation documents and then perform topic modelling on this pseudo-documents [5].

For this, word embeddings were learned first, whereby fastText [47] was chosen as the method because it can handle out-of-vocabulary words. Since the dataset of forensic short messages was considered too small to obtain meaningful word embeddings from it, instead, the unsupervised fastText skipgram-model with a window size of five and character N-Grams with a length between two and six was trained on a large external dataset to represent words as 300-dimensional word vectors. This training dataset also consisted of informal texts, namely primarily 20 million tweets provided by [48].

Subsequently, for each topic label of the seed word sets, its CluWord was created, which consisted of all words of the dataset for which the cosine similarity between their word embeddings and the word embedding of the topic label was above a threshold value of 0.45 [5]. The pseudo-documents were created by enhancing each topic label in a conversation document with its CluWord. This approach differed from the original CluWords method [5] only in the fact that the latter inserted the semantically similar words to all terms. The decision to include only the similar words to the topic labels, rather than to all the seed words, was based on the fact that the actual relevance of some seed words to the case was unclear.

The procedure for training keyATM on these pseudo-documents was analogous to Section IV-C.

V. Results

In this section, the results of the three approaches to topic modelling are presented qualitatively. The topics "Geld", "Terror" and "Verein" were selected as examples for the semi-supervised approaches. To ensure comparability, as suggested, for example by [49], among the topics of the unsupervised algorithm wLDA, those that most resembled the topics "Geld", "Terror" and "Verein" of keyATM were selected, determing the similarity with the Jensen-Shannon divergence (JSD) [50].

A. Unsupervised topic modelling

The eight words with the highest probability in the selected topics of the wLDA are shown in Table III, which also indicates the most similar seed topic in parentheses.

TABLE III
The eight most probable words from three topics of wLDA with high similarity to the selected topics of keyATM.

Topic 4 (Geld)	Topic 7 (Terror)	Topic 10 (Verein)
Euro	schlafen (sleep)	€
Geld (money)	schreiben (write)	Twitter
spielen (play)	nerven (annoy)	Stream
kaufen (buy)	Bett (bed)	first name user
holen (get)	erzählen (tell)	spenden (donate)
neu (new)	Arbeit (work)	Statement
schicken (send)	kennen (know)	first name user
PC	scheißen (shit)	zahlen (pay)

As can be seen, the fourth and seventh topics are difficult to interpret. For the fourth topic, this can be explained by the fact that topics about money and computer games seem to be mixed. Condering the seventh topic, the problem is that it generally does not contain meaningful terms, but mainly general ones. This was unexpected, as highly frequent words were removed or penalized by the adjusted Collapsed Gibbs Sampling method [4]. One possible explanation might be that the PMI weighting is unreliable for short texts, as noted by [51].

In contrast, the tenth topic could be considered relevant to the case, as it contained words such as "spenden" and "zahlen". That words like "Twitter" and the two individuals whose names appeared among the top words in the topic were related to fundraising activities and relevant to the case was evident from examining the context of these terms in the chat messages.

B. Semi-supervised topic modelling

Regarding the semi-supervised topic modelling, the eight most probable words of the three selected topics are displayed in Table IV, where the selected seed words of the respective topic are highlighted in bold and seed words of other topics are marked with an asterisk. As outlined in Table IV, the most probable words of the topic "Geld" include both seed words and intuitively associated terms like "kaufen" and "zahlen". However, these terms are quite generic, making it difficult to determine the topic's relevance to the case. Furthermore, keyATM could not identify the topic "Terror", but, instead, the topic consists of irrelevant and meaningless terms. These outcomes for both topics can be attributed to the fact that, according to Eshima et al. [4], the quality of topics is heavily dependent on the chosen seed word sets. Regarding the topic "Geld", the problem was that the seed words themselves, such as euro, were very general terms, while concerning the topic "Terror", one possible explanation for the poor results could be the low frequency of the seed words [4].

TABLE IV
The eight most probable words of the three topics "Geld", "terror" and "Verein" using the algorithm keyATM.

Geld (money)	Terror	Verein (association)
Geld (money)	Bild (image)	Stream
Euro	lachen (laugh)	boy's first name
schicken (send)	kennen (know)	€*
€	stehen (stand)	Twitter
holen (get)	süß (cute)	Event
Mail	kaufen (buy)	Twitch
kaufen (buy)	Hammer (hammer)	boy's first name
Handy (mobile phone)	Son	spenden (donate)

Regarding the seed topic "Verein", the most probable words included specific terms. However, the differences with the most similar unsupervised wLDA topic were minor, as this topic already contained relevant words. Nevertheless, keyATM enhanced interpretability through automatic label assignment.

 ISBN: 978-1-68558-089-6

C. Semi-supervised topic modelling with CluWords

Particularly concerning the topic "Terror", the inclusion of CluWords resulted in more relevant terms appearing among the most probable words. As shown in Table V, which lists the top eight words in the three topics, the topic "Terror" included terms like "Mord" and "Durchsuchungsbefehl".

TABLE V
THE EIGHT MOST PROBABLE WORDS OF THE THREE TOPICS "GELD", "TERROR" AND "VEREIN" USING KEYATM WITH CLUWORDS.

Geld (money)	**Terror**	**Verein (association)**
Euro	Mord (murder)	€*
Geld (money)	Gesinnung (attitude)	first name user
kaufen (buy)	ermittlwn (investigate)	spenden (donate)
überweisen (transfer)	Hobbermittler (hobby investigator)	Statement
nah (close)	Verbrechen (crime)	first name user
ausgeben (spend)	Drohung (threat)	SWH
zahlen (pay)	Durchsuchungsbefehl (search warrant)	Twitter
kriegen (get)	Moschee (mosque)	Tipeee

However, further research is required to determine whether the topic "Terror" is actually related to aspects like search warrants or if its presence among the most probable words is solely due to similarity based on external word embeddings. In contrast to the topic "Terror", the most probable words of the other two topics, namely "Geld" and "Verein", strongly resembled the standard keyATM topics.

VI. CONCLUSION

Topic Modelling offers high potential for the analysis of forensic short messages, where it can be used both to find evidence for suspected topics and to explore the dataset. This paper presented our initial work on assisting the investigator with these two scenarios, for which unsupervised and semi-supervised topic modelling approaches were analysed. Overall, it was found that the unsupervised algorithm wLDA already succeeded in finding case-relevant topics. keyATM as a semi-supervised approach was able to detect a similar case-relevant topic as wLDA, but failed to find further rare topics in the messages despite the inclusion of prior knowledge. In contrast, the expansion of keyATM based on Word Embedding similarity proved to be more promising.

For this reason, there is potential in semi-supervised methods that simultaneously learn word embeddings and topics, such as the Keyword Assisted Embedded Topic Model (keyETM) proposed by Harandizadeh et al. [27]. Furthermore, a problem with semi-supervised topic modelling so far was that despite term weighting, many unimportant words appeared in the topics. To address this problem, future work intends to apply the semi-supervised Guided Topic-Noise Model (GTM) [13], which specifically addresses the high number of irrelevant words in colloquial texts. Basically, future experiments should be conducted on a comprehensive set of forensic datasets to definitely decide which approaches are particularly suited for forensic data analysis.

REFERENCES

[1] M. Spranger, J. Xi, L. Jaeckel, J. Felser, and D. Labudde, "MoNA: A forensic analysis platform for mobile communication," *Künstliche Intelligenz*, vol. 36, pp. 163–169, May 2022.

[2] H. Walder, T. Hansjakob, T. E. Gundlach, and P. Straub, *Criminalistic thinking (in German)*, 11th ed., ser. Fundamentals of criminalistics (in German). Heidelberg: Kriminalistik, 2020.

[3] A. T. Wilson and P. A. Chew, "Term weighting schemes for latent dirichlet allocation," in *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies (NAACL-HLT)*. Los Angeles, California: Association for Computational Linguistics (ACL), Jun. 2010, pp. 465–473.

[4] S. Eshima, K. Imai, and T. Sasaki, "Keyword-Assisted Topic Models," *American Journal of Political Science*, pp. 1–21, Feb. 2023.

[5] F. Viegas *et al.*, "CluWords: Exploiting semantic word clustering representation for enhanced topic modeling," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. New York, NY, USA: ACM, Jan. 2019, pp. 753–761.

[6] L. Busso, M. Petyko, S. Atkins, and T. Grant, "Operation Heron: Latent topic changes in an abusive letter series," *Corpora*, vol. 17, no. 2, pp. 225–258, Aug. 2022.

[7] J. Li, W.-H. Chen, Q. Xu, N. Shah, and T. Mackey, "Leveraging big data to identify corruption as an SDG Goal 16 humanitarian technology," in *Proceedings of the Global Humanitarian Technology Conference (GHTC)*. Seattle, WA, USA: IEEE, Oct. 2019, pp. 1–4.

[8] A. de Waal, J. Venter, and E. Barnard, "Applying topic modeling to forensic data," in *Proceedings of Advances in Digital Forensics IV*, ser. IFIP - The International Federation for Information Processing Book Series, vol. 285. Paris, France: Springer Science+Business, Jan. 2008, pp. 115–126.

[9] G. E. Noel and G. L. Peterson, "Applicability of Latent Dirichlet Allocation to multi-disk search," *Digital Investigation*, vol. 11, no. 1, pp. 43–56, Jul. 2014.

[10] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 14, pp. 993–1022, Mar. 2003.

[11] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. Rio de Janeiro, Brazil: ACM, May 2013, pp. 1445–1456.

[12] M. E. Roberts *et al.*, "Structural topic models for open-ended survey responses," *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, Oct. 2014.

[13] R. Churchill, L. Singh, R. Ryan, and P. Davis-Kean, "A guided topic-noise model for short texts," in *Proceedings of the ACM Web Conference 2022 (WWW '22)*. New York, NY, USA: ACM, Apr. 2022, pp. 2870–2878.

[14] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, Singapore: ACM, Aug. 2009, pp. 248–256.

[15] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*, ser. NIPS'07. Red Hook, NY, USA: Curran Associates Inc., Dec. 2007, pp. 121–128.

[16] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via Dirichlet Forest priors," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. Montreal, Quebec, Canada: ACM, Jun. 2009, pp. 25–32.

[17] S. Wang, Z. Chen, G. Fei, B. Liu, and S. Emery, "Targeted Topic Modeling for focused analysis," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, Aug. 2016, pp. 1235–1244.

[18] Y. Meng *et al.*, "Discriminative Topic Mining via Category-Name Guided Text Embedding," in *Proceedings of The Web Conference 2020 (WWW'20)*. New York, NY, USA: ACM, Apr. 2020, pp. 2121–2132.

 ISBN: 978-1-68558-089-6

[19] H. Kim, D. Choi, B. Drake, A. Endert, and H. Park, "TopicSifter: Interactive Search Space Reduction through Targeted Topic Modeling," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*. Vancouver, BC, Canada: IEEE, Oct. 2019, pp. 35–45.

[20] J. Choo, C. Lee, C. Reddy, and H. Park, "UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013.

[21] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 382–391, Jan. 2018.

[22] H. Kobayashi, H. Wakaki, T. Yamasaki, and M. Suzuki, "Topic Models with Logical Constraints on Words," in *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*. Hissar, Bulgaria: Association for Computational Linguistics (ACL), Sep. 2011, pp. 33–40.

[23] J. Chen *et al.*, "TAM: Targeted Analysis Model with reinforcement learning on short texts," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2772–2781, Jun. 2021.

[24] K. Watanabe and A. Baturo, "Seeded Sequential LDA: A semi-supervised algorithm for topic-specific analysis of sentences," *Social Science Computer Review*, pp. 1–25, May 2023.

[25] J. He, L. Li, Y. Wang, and X. Wu, "Targeted aspects oriented topic modeling for short texts," *Applied Intelligence*, vol. 50, no. 8, pp. 2384–2399, Mar. 2020.

[26] V. Rakesh *et al.*, "A Sparse Topic Model for extracting aspect-specific summaries from online reviews," in *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, ser. Track: Web Search and Mining. Lyon, France: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 1573–1582.

[27] B. Harandizadeh, J. H. Priniski, and F. Morstatter, "Keyword Assisted Embedded Topic Model," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM'22)*. Virtual Event AZ USA: ACM, Feb. 2022, pp. 372–380.

[28] J. Wang, L. Chen, L. Li, and X. Wu, "BiTTM: A Core Biterms-based Topic Model for Targeted Analysis," *Applied Sciences*, vol. 11, no. 21, pp. 1–22, Oct. 2021.

[29] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics (ACL), Apr. 2012, pp. 204–213.

[30] Y. Feng, J. Feng, and Y. Rao, "Reward-Modulated Adversarial Topic Modeling," in *Proceedings of the 25th International Conference on Database Systems for Advanced Applications (DASFAA)*, ser. Lecture Notes in Computer Science (LNCS), Y. Nah *et al.*, Eds. Jeju, South Korea: Springer International Publishing, Sep. 2020, pp. 689–697.

[31] M. Tushev, F. Ebrahimi, and A. Mahmoud, "Domain-specific analysis of mobile app reviews using Keyword-Assisted Topic Models," in *Proceedings of the 44th International Conference on Software Engineering (ICSE '22)*. Pittsburgh Pennsylvania: ACM, May 2022, pp. 762–773.

[32] N. Amat-Lefort and S. J. Barnes, "Towards more convenient services: A text analytics approach to understanding service inconveniences in digital platforms," in *Proceedings of the 56th Hawaii International Conference on System Science (HICSS)*. Honolulu, Hawaii, USA: ScholarSpace, Jan. 2023, pp. 1346–1355.

[33] M. Spranger, F. Heinke, L. Appelt, M. Puder, and D. Labudde, "MoNA: Automated identification of evidence in forensic short messages," *International Journal on Advances in Security*, vol. 9, no. 1 & 2, pp. 14–24, Aug. 2016.

[34] J. Felser, J. Xi, C. Demus, D. Labudde, and M. Spranger, "Recommendation of query terms for colloquial texts in forensic text analysis," in *Proceedings of the International Workshop On Digital Forensics (IWDF)*. Hamburg, Germany: Gesellschaft für Informatik, Bonn, Sep. 2022, pp. 35–47.

[35] J. Riesa and I. Giuliani, "Compact Language Detector," Mountain View, California, United States, Aug. 2023, https://zenodo.org/records/7670098.

[36] J. Riesa, "Compact Language Detector," Mountain View, California, United States, Aug. 2023, https://github.com/CLD2Owners/cld2.

[37] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, pp. 1–30, Dec. 2020.

[38] R. Churchill and L. Singh, "textPrep: A text preprocessing toolkit for Topic Modeling on Social Media Data," in *Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021)*. Online Streaming: SCITEPRESS, Jan. 2021, pp. 60–70.

[39] G. Diaz, "Stopwords ISO," https://github.com/stopwords-iso/stopwords-iso, Aug. 2023.

[40] Lingolia Deutsch, "The 50 most important verbs in German (in German)," https://deutsch.lingolia.com/de/50-verben-deutsch, Oct. 2023.

[41] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK: Routledge, 1994, pp. 154–164.

[42] ——, "Improvements In part-of-speech tagging with an application to German," in *Natural Language Processing Using Very Large Corpora*, ser. Text, Speech and Language Technology (TLTB), S. Armstrong *et al.*, Eds. Dordrecht, Netherlands: Kluwer Academic Publishers, 1995, no. 11, pp. 13–25.

[43] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022.

[44] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," in *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, vol. 16. Vancouver, British Columbia, Canada: Association for Computational Linguistics (ACL), Jun. 1989, pp. 76–83.

[45] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, Apr. 2004.

[46] R. Churchill and L. Singh, "Percolation-based topic modeling for tweets," in *Proceedings of the 9th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'20)*. San Diego, USA: Association for Computing Machinery (ACM), Aug. 2020, pp. 1–8.

[47] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431.

[48] N. Kratzke, "Monthly Samples of German Tweets," Dec. 2021, https://zenodo.org/records/7670098.

[49] W. X. Zhao *et al.*, "Comparing Twitter and traditional media using Topic Models," in *Proceedings of the 33rd European Conference on IR Research (ECIR)*, ser. Lecture Notes in Computer Science (LNCS), P. Clough *et al.*, Eds. Berlin, Heidelberg: Springer Science+Business, Apr. 2011, pp. 338–349.

[50] S. M. Ali and S. D. Silvey, "A General Class of Coefficients of Divergence of One Distribution from Another," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 131–142, Jan. 1966.

[51] X. Li, A. Zhang, C. Li, J. Ouyang, and Y. Cai, "Exploring coherent topics by topic modeling with term weighting," *Information Processing & Management*, vol. 54, no. 6, pp. 1345–1358, Jun. 2018.

 ISBN: 978-1-68558-089-6