Towards a DDC-based Topic Network Model of Wikipedia

Tolga Uslu[1]*, Alexander Mehler[1], Andreas Niekler[2], and Daniel Baumartz[1]

[1]Goethe University, TextTechnology Lab,
Robert-Mayer-Straße 10, 60325 Frankfurt am Main, Germany
uslu@em.uni-frankfurt.de
mehler@em.uni-frankfurt.de
https://hucompute.org/
[2]Leipzig University, Natural Language Processing Group
Augustusplatz 10, 04109 Leipzig, Germany
aniekler@informatik.uni-leipzig.de
http://asv.informatik.uni-leipzig.de/

Abstract. This paper presents a network-theoretical approach to modeling the semantics of large text networks. By example of the German Wikipedia we demonstrate how to estimate the structuring of topics focused by large corpora of natural language texts. Algorithms of this sort are needed to implement distributional semantics of textual manifestations in large online social networks. Our algorithm is based on a comparative study of short text classification starting from two state-of-the-art approaches: Latent Dirichlet Allocation (LDA) and Neural Network Language Models (NNLM). We evaluate these models by example of (i) OAI metadata, (ii) a TREC dataset and (iii) the Google Snippets dataset to demonstrate their performance. We additionally show that a combination of both classifiers is better than any of its constitutive models. Finally, we exemplify our text classifier by plotting the topic structuring of all articles of the German Wikipedia.

Keywords: Topic model, topic networks, short text classification, LDA, NNLM, SVM

1 Introduction

In this paper, we develop a simple algorithm for modeling the semantics of large text networks. This is done by example of the German Wikipedia. Our aim is to model the structure and networking of topics as manifested by large corpora of natural language texts. Algorithms serving this task are needed to implement a distributional semantics of textual manifestations in online social

* Financial support by the Bundesministerium für Bildung und Forschung (BMBF) via the CEDIFOR project (https://www.cedifor.de/en/) as being performed by the TTLab at Goethe University Frankfurt (https://hucompute.org/) is gratefully acknowledged.

networks. One may want to know, for example, what topics are focused in a certain period of time in Twitter. Alternatively, one may want to know which fields of knowledge are either preferred or underrepresented in media such as Wikipedia or Wiktionary [20]. In order to answer questions of this sort, it is necessary to determine the topic distribution of each individual text aggregate of the focused media and to decide how the resulting distributions are to be networked. This is the task of the present paper.

Our algorithm for modeling the thematic structure of large text corpora utilizes a well-established topic classification, that is, the Dewey Decimal Classification (DDC). More specifically, we build on a comparative study of approaches to short text classification. Short texts (e.g. tweets) refer to situations in which only snippets (e.g., metadata, abstracts, summaries or only single sentences such as titles) are available as input for classification instead of full texts. One example of this is digital libraries working on OAI (Open Archives Initiative) metadata [28]. It also concerns text mining in online social media by example of chat messages, news feeds, tweets [24], or turn-taking in online discussions [7]. In all these cases the central information to be extracted is what the snippets are about in order to classify them thematically [29], to disambiguate or to classify their constituents [8] or to enrich them by means of external knowledge resources. The requirement to handle big data streams is another reason to process snippets instead of full texts even if being accessible. In each of these cases, classifiers are influenced more by the sparseness of the lexical content of short text. Therefore, one needs both fast and accurate classifiers that are expressive enough to overcome the problem of lexical sparseness.

In this paper, we present a network model of topic structuring that is based on a comparative study of text snippet classification starting from two state-of-the-art approaches: Latent Dirichlet Allocation (LDA) and Neural Network Language Models (NNLM). In the latter case we experiment with fastText [13], which has been developed to overcome problems of time-consuming deep learners. We test each of these approaches separately and also test a variant in which fastText is additionally fed with topics generated by LDA. We have found that both classifiers classify with similar quality. Feeding fastText with LDA-based topics has not accomplished any improvements. However, the combination of both classifiers has enabled us to improve the overall quality of classification.

As a gold standard of topic modeling we use the DDC, which is the most common thematic classification system in (digital) libraries. One advantage of this approach is that it provides access to extensive training and test data. In addition to that we consider two tasks of short text classification in order to enable comparisons with state-of-the-art approaches: the first uses the TREC (Text Retrieval Conference) dataset [26], the second the Google Snippets dataset [21]. As a result of these evaluations we receive a classifier that allows for determining the topic distribution of all articles of the German Wikipedia so that we can finally model the networking of these topics. In this way, we exemplify how to map text corpora on networks of topics described by them.

The paper is organized as follows: Section 2 discusses related work of text classification. Section 3 describes the series of topic classifiers with which we experiment in Section 3. In Section 5, the best performer of this evaluation is applied in order to visualize the thematic structure of Wikipedia. Finally, in Section we draw a conclusion and give an outlook on future work.

2 Related Work

Since our paper deals with the DDC-related classification of short texts, we consider two areas of related work: text snippet classification and topic modeling used for content analysis of online social networks.

By exploring OAI Metadata, [28] present an SVM-based classifier that considers all three levels of the DDC. A basic restriction of this approach relates to the fact that it only processes OAI records of a certain minimal length. In contrast to this, we do not consider such a lower bound so that we face a more realistic scenario in which the topic of a snippet is highly underrepresented by its vocabulary. Thus, unlike [28], we consider all 2nd-level DDC categories: in the case of English texts this 2-level approach even deals with a larger set of target classes than the 3-level approach of [28] (who are considering only 88 classes in total). Likewise, we aim at overcoming problems of computational complexity as exemplified by the approach of [27]. This research shows that DDC-related text categorization, especially by example of short texts, has been a desideratum so far.

The classification of text snippets, regardless of the classification scheme, has made significant progress with the utilization of neural networks for text classification. [29] show that the projection of similar text snippets onto a matrix can be a very helpful input to training a convolutional neural network that outperforms approaches based on other neural networks [14,16], LDA [21,5] or SVMs [23]. These approaches concentrate on single aspects like syntactic rules, topic modeling of text snippets or semantic similarity measurement. Our case study examines sources of information that have not previously been investigated together in the context of classifying text snippets. This includes

1. information about n-grams,
2. information provided by dataset-external semantic knowledge as given by topic models derived from general corpora, and
3. information provided by NLP tools about tokens, lemmas and parts of speech.

We integrate these information sources into our model and compare the performance of a neural network and an SVM-based approach as two competing instances of our model.

The usage of topic models and thematic classifications as an input to graph structures has been explored in different ways. Mostly, the connections in such graphs are built by topical similarities of the documents [4,17]. In this way, one can observe, for example, which sources or authors are highly connected in

the resulting graph. On the other hand, social networks can be analyzed with respect to topical preferences manifested by their textual content [19,3,9,25]. Our approach also adds the network perspective regarding topic distributions. However, we additionally explore the networking of topics as a function of the polysemy of the underlying textual aggregates.

3 Models of Topic Classification

In this section we describe the models that we used for topic-related text classification: based on LDA (Sec. 3.1), on neural networks (Sec. 3.2), on neural networks fed by LDA-based topics (Sec. 3.3), on neural networks fed by vectors representing word significance distributions (Sec. 3.4), and based on a combination of a SVM and a NNLM-based classifier (Sec. 3.5).

3.1 LDA-based classification (SVM-LDA)

Topic models, as the Latent Dirichlet Allocation (LDA) model, utilize large text corpora to infer a latent distribution of words over a given number of topics so that each document can be described as a mixture of those topics when exploring co-occurrences of their lexical constituents [2]. The parameters of the LDA model, ϕ (word-topic distribution) and θ (document-topic distribution) can be estimated using either a variational inference scheme or Gibbs samplers on a training set of documents [11]. One of the great benefits of topic models is the generalization of the model. The topic structure of documents which do not belong to the training set can be inferred using the fixed model parameters even if additional unknown vocabulary is included. In this way, each document of a corpus can be described in terms of its topic distribution regarding the parameters of a topic model that has been generated by means of a reference corpus.

In text classification, a vector space model is often used to derive elementary features for documents. The famous tf-idf scheme, entropy-based measures or the pointwise mutual information can be used as alternatives to weight the terms in the document vectors. In [27], lexical features are weighted using such term weights. The resulting feature vectors are used to train a Support Vector Machine (SVM) using a Negative Euclidean Distance Kernel (NDK) on a dataset of 4 000 German DDC classified documents. This approach achieves 0.723 in F-score with respect to the dataset.

Our approach uses additional information besides the tf-idf weights including the extraction of uni-, bi-, and trigrams and the additional use of topics as features within an SVM-classification scheme. That is, we informationally enrich each document in the training set. Unigram stop words are deleted from the set of features and words of the document collection were stemmed. Since a topic encodes an associated vocabulary context (e.g., the word-topic distribution), each document holds general information about other documents containing similar topics. This information can be useful in classification tasks if we augment the

lexical features with the topic structure for a category. Our hope is to enhance the results of [27] by the use of such topic model-related features. The here described approach uses the LDA model of [2] to infer topics on the dataset.

We considered a novel strategy to augment the lexical features with topic information. An LDA-model with 100 topics is inferred on "general" language data and the topic distributions of documents from both, training and test datasets, are determined with respect to this model.[1] This gives us an additional topic distribution on each document in the training and test sets. The language resource to build our model is based on corpora from the Wortschatz[2] project. We chose 3 million sentences from news data which were crawled in 2015 from German and English websites to build the respective models. We did not use Wikipedia-based data because of the possible domain similarity to our OAI-datasets in Section 4.

Additionally, we apply the tf-idf weighting scheme to the document term vectors (uni-, bi- and trigrams) in order to reduce the influence of general vocabulary. Then, we append the topic distribution for a document as a vector of probabilities to its vector of lexical features. To train the SVM, we used the R-version of liblinear with an L2-regularized logistic regression and estimated the C-parameter heuristically [6,12].

3.2 Neural network-based classification (NN)

For the neural network-based approaches, we started with the simple but very efficient classifier of [13] called fastText (see Figure 1 for a visual depiction of this model in our context). fastText uses a bag-of-words (bow) model and defines the occurrences of words in a document as input of the neural network. Since the order of words is ignored in the bow-model, fastText uses n-grams to capture some information about the local order. To avoid being forced to use default parameter settings, we have written a parameter analyzer, which searches the parameter space for better performing settings (according to a hill-climbing algorithm). Since the input corpora were not preprocessed, we applied various NLP tools to obtain additional information about tokenization, lemmas and parts of speech. We also used pretrained word embeddings to initialize the neural network.

3.3 Neural Network based classification combined with LDA (NN-LDA)

Since fastText only accepts text as input, we adapted its architecture so that we can process the textual content in conjunction with the topic distribution of a document. To this end, we extended the neural network underlying fastText to include not only input nodes for words, but also for each topic provided by the model of Section 3.1. Thus, when considering a distribution of 100 LDA-based

[1] Our experiments showed that 100 topics provided the best topic solution for the described experiments in terms of F1 performance of the final classifier. We tested 20, 50, 75, 100, 250 and 500 as values for the amount of topics to infer.

[2] http://wortschatz.uni-leipzig.de/en/download

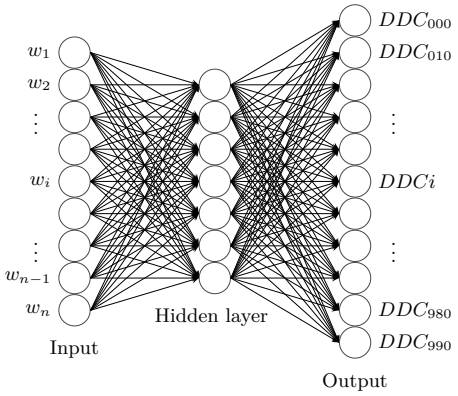


Fig. 1: Architecture of Model 3.2.

topics, we added 100 input nodes to the neural network, which are activated according to the topic values of the input document. In this way, our extension of fastText is additionally fed with numerical values signaling membership to topics derived from LDA.

3.4 Neural network-based classification combined with GSS (NN-GSS)

Taking profit of the fact that we adapted fastText to additionally accept numeric values as input, we calculated the GSS coefficient (Galavotti-Sebastiani-Simi) [10] for each pair of words in the input corpus and first-level categories of the DDC. In this way, each word of the input corpus is mapped onto a 10-dimensional feature vector whose dimensions denote the association of the word with respect to the given target category. Under this regime, the classifier uses feature vectors of GSS coefficients instead of the words themselves. This results in a neural network with $n \times m$ input nodes, where n is the size of the vocabulary of the input corpus and $m = 10$ the number of top-level DDC classes.

3.5 Combining both worlds (NN-SVM-LDA)

By means of an error analysis we found that the SVM and the NN-based classifiers make different errors, although achieving similar classification qualities. Therefore, we calculated a scoring for each document with respect to each target category based on the two best performing classifiers from the SVM- and NN-world, respectively, and experimented with two methods to combine their scorings:

1. voting for the target class as a function of the maximum score (not to be confused with majority voting) or
2. by means of the average score.

4 Classification Experiment

We test the models of Section 3 by example of four different data sets. Two of these samples represent OAI-based datasets (one in German and one in English) which were used regarding two classification tasks. The first task was to classify the first level of the DDC (10 classes in the English corpus (EN-10) and 10 classes in the German corpus (DE-10)). The second task was to classify the first two levels of the DDC (93 classes in the English corpus (EN-All) and 88 classes in the German corpus (DE-All)). The German corpus consists of 595 493 records with an average of 37.24 words per document. The English corpus consists of 1 222 948 records with an average of 50.69 words per document. Each corpus was randomly divided into training (70%) and test (30%) sets. In order to ensure comparability with state-of-the-art systems for classifying text snippets, we also evaluated our models using the TREC 2003 Question Answering dataset [18] and the Google Snippet dataset [22] as used in [29].

4.1 Classification

For the SVM-based classification using LDA-features we trained one SVM-model for each dataset and task. The results are shown in Table 1. The SVM-LDA model outperforms the models described in [27,28]. Furthermore, it performs as good as the NN-based model described in [29]. In examining the impact of all features, we find that the n-gram features have an impact similar to features provided by the topic model. The combination of both feature sets does not improve overall performance. In detail, the classification for the DE-10 dataset results in the following F1-scores for the different feature configurations: 1. unigrams – 0.786; 2. unigrams + topics – 0.805; 3. n-grams – 0.814; 4. n-grams + topics – 0.815. In general, it can be shown that using topic model features improves the quality of the classification, albeit to a limited extent. From a classification point of view, n-grams and LDA-based topics seems to encode related information within the feature space. This may give rise to future research.

In the case of classifying with neural networks, we carried out a parameter study to detect optimal parameter settings. To this end, we examined the following parameters:

- Learning rate (0.025 - 0.1)
- n-grams (1 - 5)
- Dimension (50 - 100)
- Epochs (500 - 10000)

The results are shown in Table 1. It shows that SVM-LDA performs better than its NN-based counterparts in the case of the English data sets, while the NN-based (lemma + POS) classifier outperforms its competitors in the case of the German data. However, the difference to SVM-LDA is very small. Additionally feeding the NN with LDA topics (NN-LDA) performs worse as does NN-GSS (DE-10). Further, lemma-level features perform very little better than token-level ones (DE-10 and DE-All).

Corpus	Features	N-gram	F-scores
EN-10	NN: token-based	3	0.748
EN-10	SVM-LDA	1-3	0.771
EN-All	NN: token-based	3	0.698
EN-All	SVM-LDA	1-3	0.717
DE-10	NN: token-based	1	0.814
DE-10	NN: lemma + POS	2	0.816
DE-10	NN-GSS	–	0.792
DE-10	NN-LDA: lemma + POS + topics	2	0.795
DE-10	NN (lemma + POS) + SVM-LDA	1-3	0.820
DE-10	SVM	1-3	0.814
DE-10	SVM-LDA	1-3	0.815
DE-All	NN: lemma + POS	2	0.757
DE-All	NN: token-based	2	0.753
DE-All	SVM-LDA	1-3	0.750

Table 1: F-scores of text snippet classification based on four different corpora.

Method	Google Snippets	TREC
SVM-LDA (Section 3.1)	0.960	0.971
NN (Section 3.2)	0.962	0.974
[29]	0.851	0.972

Table 2: Comparison of our models to the best performing model in [29].

Next, we selected the best classifiers of both areas (SVM and NN) and further analyzed their classification quality. Although both classifiers perform similarly (81.4% and 81.6%), they make different mistakes. When always knowing the right class of a snippet and then selecting the classifier voting for it, we would achieve an F-score of 89.6% as a kind of an upper bound of an algorithmic combination of SVM-LDA and NN (lemmas + POS). However, we cannot presuppose this knowledge. Thus, we need to apply one of the combinations of Section 3.5. This produces an the F-score of 82% in the DE-10 experiment using the method of averaging scores.

Finally, we compared the best performers of Table 1 with those documented by [29]: Table 2 shows that we also outperform these competitors by example of the Google and the TREC data by more than 10%. Obviously, our approach is more than just competitive.

4.2 Discussion

Although we worked with the complete DDC corpus (as described at the beginning of this section) and therefore had to classify many small texts, we achieved

rather promising classification results. This holds for both the SVM-LDA and the NN-based classifier. Both classifiers outperform the approach of [28] (being based on a classical SVM) and the one of [27] (using a newly invented kernel function), even when using the full dataset rather than using only a subset of texts of a certain minimal length. In addition, both our classifiers outperform their competitors described in [29] (see Table 2).

In the case of SVM-LDA, we show that information provided by LDA has a positive impact on classification. The different errors generated by SVM-LDA and NN indicate that there is a high potential in the combination of both approaches. However, the neural network achieved worse results when directly using topic information provided by the LDA (NN-LDA – see Table 1). Therefore, information about topics as provided by LDA should be integrated into neural networks in other ways than by the one used here so that one can make better use of this information. The very same can be said about using GSS-weighted vectors (NN-GSS). Experiments of [15] and [16] show the potential of including word similarity information within a convolutional layer of a neural network. This type of semantic smoothing might also be interesting to explore similarities of documents that are used simultaneously for training the network. In this way, we may help to better integrate topic models and neural networks. This will also be an object of future research. In any event, we are now in a position to guess for any piece of text – down to the level of single words (supposed they have been seen during training) – what topic class of the DDC it likely belongs to. In this way, we have a very powerful topic model that can be used to study the topic distribution and topic networking of online social networks and related media.

5 A bird's eye view of topic networks

In this section, we experiment with the best performing (non-combined) topic classifier of Section 4, that is, NN (lemma + POS, DE-All), to model inter-topic structures. This is done by example of a complete release of the German Wikipedia (download: January 20th, 2017). That is, each of the 1 760 875 Wikipedia articles in this release is mapped onto a subset of DDC categories and each of the 53 122 347 links between these articles is mapped onto arcs between nodes denoting these categories. We address two tasks:

- Topic distribution and thematic dominance: Firstly, we try to determine for each article of this release what topics it deals with. This means that we assume a multi-label classification scenario in which the same article possibly manifests several topics to varying degrees (measured by the strength μ of classification).
- Topic linkage: Secondly, we use this information to generate a network that shows how these topics are interlinked. Through this network we provide two types of information: about the salience of topics and about topics being jointly manifested by articles.

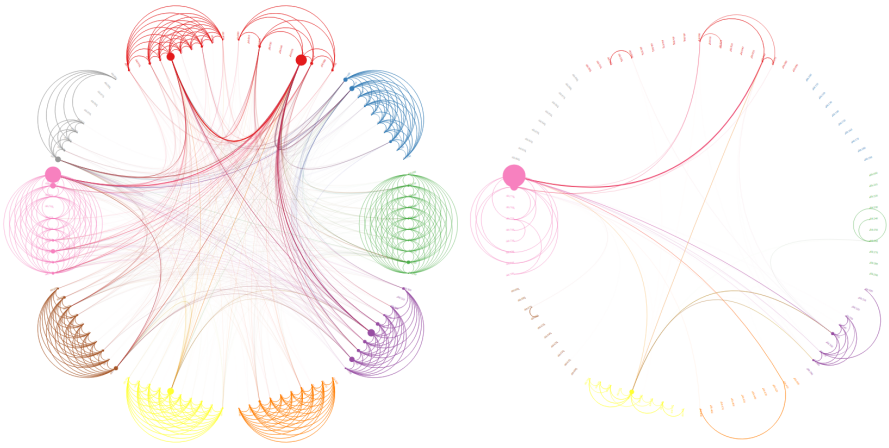
– Visualization: Our visual depiction of this topic network is based on the following statements:
  1. The more articles describing the same topic and the stronger they do, the more salient this topic becomes and the bigger its visual depiction.
  2. The more articles related to the topic A are linked with articles related to the topic B, the larger the visual representation of the arc from A to B.

The result of this visualization procedure is depicted in Figure 2 (a). It demonstrates that articles are usually so ambiguous (in terms of our classifier) that applying this algorithm of network induction to all Wikipedia articles ultimately brings us close to a completely connected topic network. Thus, in order to reveal more structure, we additionally experiment with varying thresholds of minimal classificatory membership by analogy to α-cuts in fuzzy set theory. This is demonstrated in Figure 2 (b): it shows that for a threshold of maximum class membership, we arrive at an extremely sparse network in which only a tiny fraction of topic-to-topic links survive. At this level, inter-topic structure almost diminishes: a single highly salient category emerges, that is, DDC class 790 (Recreational & performing arts). Note that in Figure 2 (b), salience of vertices is also a function of α: only those categorizations are counted per DDC class whose membership value μ is at least α; the same constraint also concerns the linkage of topic nodes.

Now the question is raised how the network of Figure 2 (b) passes over into the one of Figure 2 (a): how does it move from crisp to fuzzy categorization? In order to answer this question, we compute networks according to our algorithm of network induction by taking only those mappings of articles x to DDC categories A into account, whose class membership $\mu_A(x)$ satisfies the inequality $\mu_A(x) \geq \alpha$ while reducing α stepwise from 1 to 0.01 (in steps of 0.01). Then, for each of these α values we induce a separate network for which we compute a subset of graph invariants as depicted in Figure 3:

1. The unweighted C [30] and the weighted cluster value C_d^w of directed networks [1] estimating the probability with which nodes linked from the same node are themselves connected, taking into account the weights of these arcs.
2. The proportion of vertices belonging to the largest strongly lcc_s and weakly lcc_w connected component.
3. The cohesion value coh, that is, the proportion of existing arcs in relation to the number of possible arcs.

Finally, we plot aggregated values of graph invariants, that is, the product of C_d^w and C on the one hand and of coh and lcc_s on the other. We observe that compared to C_d^w (C), the values of $C_d^w \cdot coh$ $(C \cdot coh)$ are significantly smaller. This indicates that although clustering rapidly increases even for smallest decreases of maximum α, clustering rather concerns a small subset of vertices. At the same time, we observe that by weighting C_d^w with lcc_s, clustering does not decrease by far to the same extent (the same holds, though to a higher degree, for $C \cdot lcc_s$). This suggests that adding arcs as a result of reducing α contributes more to



(a) alpha=0 (b) alpha=1

Fig. 2: Comparison of the Wikipedia based DDC network with alpha = 0 and 1.

● DDC 0, ● DDC 1, ● DDC 2, ● DDC 3, ● DDC 4, ● DDC 5, ● DDC 6, ● DDC 7, ● DDC 8, ● DDC 9

the connectivity than to the clustering of the underlying networks. In other words: increasing the level of allowable ambiguity rather leads to connected topic networks than to networks exhibiting many local (triadic) clusters. If we compare the distribution of C_d^w as a function of α with C, we observe that for smaller values of $\alpha = 0.4$ C_d^w starts shrinking as C continues to grow: for this threshold value, smaller weights of edges begin to overlay higher edge weights. In other words, at this level, the categorization quickly becomes too much blurred. In any event, we also observe that under our model of topic classification, articles tend to be highly polysemous so that one rapidly approximates a highly connected graph ($lcc_w \sim 1$) that also exhibits high cluster ($C > 0.8$) and cohesion values ($coh > 0.2$).

Obviously, this analysis provides both (i) a bird's eye view on topic structuring as manifested by text networks as large as Wikipedia and (ii) an assessment of its ambiguity. The latter is done by analyzing the transition dynamics starting from clear classifications to highly ambiguous ones, taking into account clustering and connectivity.

6 Conclusion

In this paper, we developed a simple algorithm for analyzing and visualizing the topic structure of large text networks. To this end, we experimented with a series of classifiers in the context of three evaluation scenarios. This included an SVM-

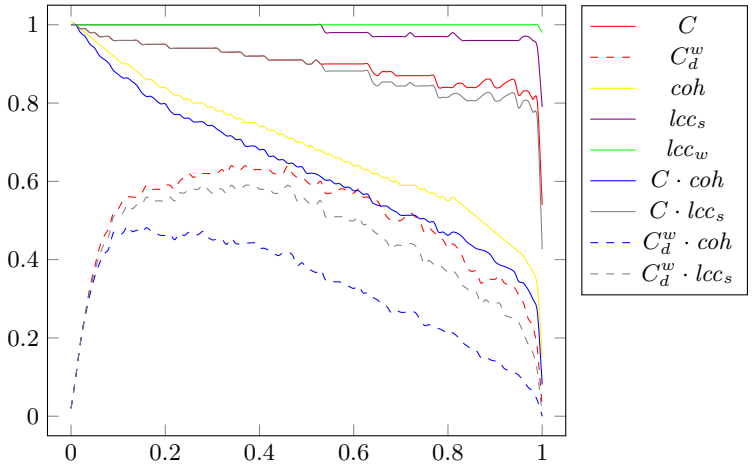


Fig. 3: Distribution of graph invariants of topic networks as a function of minimal class membership α.

based classifier exploring topics derived from LDA, a NNLM-based classifier (i.e, fastText) as well as combinations thereof. Using the best performer of these experiments, we have shown how to generate a bird's eye view of the salience and linkage of topics as manifested by hundreds of thousands of texts. In this context, we observed a very high degree of thematic ambiguity, which makes it necessary to search for more precise, less ambiguous classifiers. This will be the task of future work. Nevertheless, our paper shows a way to automatically visualize the thematic dynamics of textual aggregates as produced by large online social networks.

References

1. Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The architecture of complex weighted networks. In Guido Caldarelli and Allesandro Vespignani, editors, Large Scale Structure and Dynamics of Complex Networks, pages 67–92. World Scientific, New Jersey, 2007.
2. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.
3. Youngchul Cha and Junghoo Cho. Social-network analysis using topic models. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 565–574. ACM, 2012.
4. Jonathan Chang and David Blei. Relational topic models for document networks. In Artificial Intelligence and Statistics, pages 81–88, 2009.
5. Mengen Chen, Xiaoming Jin, and Dou Shen. Short text classification improved by learning multi-granularity topics. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three, IJCAI'11, pages 1776–1781. AAAI Press, 2011.

6. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 9:1871–1874, 2008.
7. Eric N Forsythand and Craig H Martell. Lexical and discourse analysis of online chat dialog. In Semantic Computing, 2007. ICSC 2007. International Conference on, pages 19–26. IEEE, 2007.
8. Robert Gaizauskas, Emma Barker, Monica Lestari Paramita, and Ahmet Aker. Assigning terms to domains by document classification. In Proceedings of the 4th International Workshop on Computational Terminology, pages 11–21, 2014.
9. Brynjar Gretarsson, John O'donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. Topicnets: Visual analysis of large text corpora with topic modeling. ACM Transactions on Intelligent Systems and Technology (TIST), 3(2):23, 2012.
10. Luigi Galavotti and Fabrizio Sebastiani, and Maria Simi. Experiments on the use of feature selection and negative evidence in automated text categorization. Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, PT, pages 59–68, 2000.
11. Thomas L Griffiths and Mark Steyvers. Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1):5228–5235, 2004.
12. Thibault Helleputte. LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library, 2015. R package version 1.94-2.
13. Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. CoRR, abs/1607.01759, 2016.
14. Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.
15. Jonghoon Kim, François Rousseau, and Michalis Vazirgiannis. Convolutional sentence kernel from word embeddings for short text categorization. In EMNLP, pages 775–780, 2015.
16. Yoon Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.
17. John D. Lafferty and David M. Blei. Correlated topic models. In Advances in neural information processing systems, pages 147–154, 2006.
18. Xin Li and Dan Roth. Learning question classifiers. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pages 1–7. Association for Computational Linguistics, 2002.
19. Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. Computer Science Department Faculty Publication Series, page 3, 2005.
20. Alexander Mehler, Rüdiger Gleim, Wahed Hemati, and Tolga Uslu. Skalenfreie online soziale Lexika am Beispiel von Wiktionary. In Stefan Engelberg, Henning Lobin, Kathrin Steyer, and Sascha Wolfer, editors, Proceedings of 53rd Annual Conference of the Institut für Deutsche Sprache (IDS), March 14-16, Mannheim, Germany, Berlin, 2017. De Gruyter.
21. Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th International Conference on World Wide Web, WWW '08, pages 91–100, New York, NY, USA, 2008. ACM.
22. Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web, pages 91–100. ACM, 2008.

23. João Silva, Luísa Coheur, Ana Cristina Mendes, and Andreas Wichert. From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, 35(2):137–154, 2011.
24. Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10), pages 841–842, New York, NY, USA, 2010.
25. Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pages 807–816, New York, NY, USA, 2009. ACM.
26. Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '00, pages 200–207, New York, NY, USA, 2000. ACM.
27. Tim vor der Brück, Steffen Eger, and Alexander Mehler. Complex decomposition of the negative distance kernel. CoRR, abs/1601.00925, 2016.
28. Ulli Waltinger, Alexander Mehler, Mathias Lösch, and Wolfram Horstmann. Hierarchical classification of oai metadata using the ddc taxonomy. In Raffaella Bernardi, Sally Chambers, Björn Gottfried, Frédérique Segond, and Ilya Zaihrayeu, editors, NLP4DL/AT4DL, volume 6699 of Lecture Notes in Computer Science, pages 29–40. Springer, 2009.
29. Peng Wang, Bo Xu, Jiaming Xu, Guanhua Tian, Cheng-Lin Liu, and Hongwei Hao. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174, Part B:806 – 814, 2016.
30. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. nature, 393(6684):440–442, 1998.