Probabilistic Neural Topic Models for Text Understanding

by

Gabriele Pergola

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Computer Science

December 2020

Declarations

I, Gabriele Pergola, declare that this thesis titled, 'Probabilistic Neural Topic Models for Text Understanding' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Acknowledgments

Completing this dissertation would not have been possible without the help of all the incredible people I was fortunate to meet during my studies.

First, I would like to thank my supervisor Yulan He. You have been the most understanding and supportive advisor I could have wished for, thank you for your sensible advice, patience and for fostering a productive research environment. Most of all, thank you for your example in being a leading researcher in your field while always being deeply human and understanding.

I was fortunate to attend many meetings with David Lowe and Maria Liakata and benefited from their support and advice. Thank you for your insights and time.

Thank you to all the present and past members of our research lab and group, you have all contributed immensely to my personal and professional time during these years: Lin, Vishwash, Thomas, Shan, Vassan, Noa, Lixing, Runcong, Fan, Elena, Fabio, Manlio, and all the others.

I owe a lot of my first steps into the academic life to my postgraduate mentor, Marco Ortolani, your support has been crucial for the beginning of this journey.

I am grateful to my community, Comunità Exodos, which has been a never ending source of energy for research and truth.

I am deeply thankful to my parents, Giuseppe and Rita, and my brother, Gianluca, for their understanding and unconditioned support: without them none of this would have been possible.

Finally, I will forever be thankful to my girlfriend, Marika, for her endless love, for starting and continuing with me this life journey.

Abstract

Making sense of text is still one of the most fascinating and open challenges thanks and despite the vast amount of information continuously produced by recent technologies. Along with the growing size of textual data, automatic approaches have to deal with the wide variety of linguistic features across different domains and contexts: for example, user reviews might be characterised by colloquial idioms, slang or contractions; while clinical notes often contain technical jargon, with typical medical abbreviations and polysemous words whose meaning strictly depend on the particular context in which they were used.

We propose to address these issues by combining topic modelling principles and models with distributional word representations. Topic models generate concise and expressive representations for high volumes of documents by clustering words into "topics", which can be interpreted as document decompositions. They are focused on analysing the global context of words and their co-occurrences within the whole corpus. Distributional language representations, instead, encode the word syntactic and semantic properties by leveraging the word local contexts and can be conveniently pre-trained to facilitate the model training and the simultaneous encoding of external knowledge. Our work represents one step in bridging the gap between the recent advances in topic modelling and the increasingly richer distributional word representations, with the aim of addressing the aforementioned issues related to different linguistic features within different domains.

In this thesis, we first propose a hierarchical neural model inspired by topic modelling, which leverages an attention mechanism along with a novel neural cell for fine-grained detection of sentiments and themes discussed in user reviews. Next, we present a neural topic model with adversarial training to distinguish topics based on their high-level semantics (e.g. opinions or factual descriptions). Then, we design a probabilistic topic model specialised for the extraction of biomedical phrases, whose inference process goes beyond the limitations of traditional topic models by seamlessly combining the word co-occurrences statistics with the information from word embeddings. Finally, inspired by the usage of entities in topic modelling [85], we design a novel masking strategy to fine-tune language models for biomedical question-answering. For each of the above models, we report experimental assessments supporting their efficacy across a wide variety of tasks and domains.

Contents

List of Figures

List of Tables

Acronyms

BOW Bag-Of-Words.

CBOW Continuous Bag-Of-Words.

CNN Convolutional Neural Network.

GAN Generative Adversarial Network.

GRU Gated Recurrent Units.

LDA Latent Dirichlet Allocation.

LM Language Model.

LSA Latent Semantic Analysis.

LSI Latent Semantic Indexing.

LSTM Long Short-Term Memory.

MLM Masked Language Model.

MTL Multi-Task Learning.

NER Named Entity Recognition.

NLP Natural Language Processing.

OOV Out-Of-Vocabulary.

PCA Principle Component Analysis.

QA Question Answering.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

SGD Stochastic Gradient Descent.

SGNS Skip-Gram with Negative Sampling.

SVD Singular Value Decomposition.

SVM Support Vector Machine.

tf-idf term frequency-inverse document frequency.

Chapter 1

Introduction

1.1 Motivation

Making sense of text is still one of the most fascinating and open challenges, thanks to and despite the vast amount of information continuously produced by recent technologies. The proliferation of sources of digital information — online newspapers, electronic health records, social networks, user reviews, and so on — has generated an unprecedented flow of text which inevitably calls for more efficient methods to analyse this digitised collective knowledge. Questions such as "what product features are the most appreciated by users?", "what topics are recently trending on social networks?", or "what are the treatments most frequently discussed for a specific disease?" need to be addressed by exploiting automatic methods due to the enormous quantity of involved documents.

However, along with the growing size of textual data, automatic approaches have to deal with the wide variety of linguistic features across different domains and contexts. For example, user reviews might be characterised by colloquial idioms, slang, or contractions; while clinical notes often contain technical jargon, multi-word phrases (i.e. single concepts unfolded across several words), with typical medical abbreviations and polysemous terms whose meaning strictly depends on the particular context in which they occur. Most of these words and phrases are critical in determining the precise meaning within documents, yet they are domain-specific, they might depend on the particular sentiment expressed (e.g., user reviews), or we might lack the necessary statistics to model them due to limited and costly data (e.g., medical documents). In addition, authors would frequently mix their opinions with factual descriptions, making it difficult to separate and process them appropriately.

For example, we report here two extracts from user reviews: "*Our children didn't manage to clean their plates! Plenty of food!*" and "*After one cycle the crockery is still dirty, it doesn't clean the plates even at full power.*", with the first one being about a restaurant and the second one about a dishwasher. Interestingly, the same expression "*not to clean the plates*" can be regarded as positive for *food*, while it bears a negative polarity for *kitchen equipment*, demonstrating the importance of jointly considering both topics and sentiments for better sentiment analysis.

Another prominent example comes from clinical notes, as medical concepts are often expressed in terms of multi-word phrases. For example, the phrases "*white blood cell*" or "*blood glucose*" would lose their meaning if decomposed as unigrams, and the word *cell* and *glucose*, if considered singularly, might be wrongly clustered together because of the shared *blood* term. Moreover, the electronic health record (EHR) narratives, for instance, are remarkably heterogeneous, ranging from discharge summaries to history of patients and consultations, resulting in high dimensional data that need to be processed to extract information.

Distributional Methods and Topic Model

Topic models have established themselves as effective tools to generate concise and expressive representations of high volumes of documents [16]. They cluster words into "topics", which can be interpreted as a document decomposition into a small set of themes to facilitate the high-level understanding of a corpus, otherwise too large to read. Analogously, machine reading comprehension (MRC) approaches allow users to identify information of interest indicating the span of text most likely associated with a user's query [6].

The recently developed distributional word and language representations based on neural models [36, 53, 79] combine well with the more traditional topical representations of documents, including topic models [13, 16, 34]. The former are distributional representations of text, thus they encode the text semantics through the local word context, typically defined by a small context window. They not only encode syntactic and semantic properties of words, but they are also highly efficient and parallelisable, scaling to large corpora. This entails the additional possibility to *pretrain* these representations so that we do not need to train models from scratch, and we can combine knowledge encoded from general-purpose corpora (e.g., Wikipedia) with domain-specific corpora (e.g., PubMed dataset [91]). By contrast, topic models provide information about the overall themes discussed in documents and can be leveraged to enhance the distributional models in detecting more precisely the topical context of the analysed text.

In this thesis, we propose and analyse different approaches to enhancing or combining distributional representations with topic modelling. We explore how hierarchical neural models for sentiment analysis can be modified to generate topical representations of text (§3). We propose a novel combination of neural topic models [84] and adversarial training [55] to distinguish between different types of topics based on their high-level semantics (§4). We propose a biomedical topic model using word embedding information to drive the inference of a probabilistic topic model (§5). Finally, inspired by the usage of entities in topic modelling [85], we design a novel masking strategy to fine-tune language models for biomedical question-answering (§6).

1.2 Research Objectives

The primary aim of this thesis is to investigate the combination of topic modelling principles and models with neural architectures and distributional word representations for text analysis, along with a systematic evaluation of its efficacy in generating topics that are accurate syntheses of the main themes in text and effective features for those downstream tasks where an enhanced degree of topic-awareness would be beneficial.

Our hypothesis is that topic models and neural architectures are a suitable combination for capturing high-level text semantics, such as the expressed sentiments or the domain-specific concepts. In particular, we posit that such a combination can be designed to consider simultaneously the global themes characterising a corpus and the local meanings of words and sentences in a document, and can be used to identify and separate topics based on their high-level semantics (e.g., topics about opinions or topics about facts or descriptions). We also posit that leveraging the large knowledge implicitly encoded in distributional representations of text would lead to more precise and expressive topics and features, and would be especially beneficial for domain-specific documents (e.g., clinical notes). Further, their combination would be a viable means to seamlessly integrate the vast structured knowledge available in technical domains by its first codification into distributional representations, which in turn, could be integrated into topic models. The above hypotheses inspired the methodologies presented throughout this thesis and can be summarised in the following research objectives (ROs):

RO 1 Combining global and local context of words. In topic modelling, topics in a corpus emerge from the word co-occurrences in documents (i.e. global context), thus leveraging whole documents as context to characterise the word meaning. Conversely, word embeddings encode the syntactic and semantic properties relying on the immediate surrounding terms occurring in a context window (i.e. local context). We propose to combine these two approaches to generate topics with greater consistency to the analysed text (§3, §5) .

RO 2 **Generating fine-grained topics.** The stylistic features characterising specific domains, such as user reviews or clinical notes, have a significant impact on the overall meaning of text. Their analysis requires models with high resolution, able to detect subtle differences that determine the shift in meaning across documents. Thus, we plan to perform a fine-grained detection of topics based on the expressed sentiments (§3), the different facets discussed, e.g., objective descriptions or opinions (§4), or the technical concepts characterising specific domains (§5).

RO 3 **Incorporating unstructured knowledge.** Word embeddings and language models implicitly encode a large volume of knowledge thanks to the unstructured text employed to train them. To leverage all this unstructured knowledge, we propose to exploit the distributed representations to model the word sentiment polarity (§3), domain-specific lexicon and concepts (§5).

RO 4 **Incorporating structured knowledge.** In the medical field, as in other technical domains, there has been a rich proliferation of resources providing ready-to-use structured knowledge. We aim at its seamlessly integration by identifying and encoding pivotal biomedical entities and concepts directly into topic models (§5) and language models (§6).

RO 5 **Evaluation on downstream tasks.** Along with the increased quality of topics in terms of expressiveness and coherence, meaningful semantic representations should have an observable impact on downstream tasks relying on those features. Therefore, we want to evaluate the influence of these topics on downstream tasks evaluating how enhancing and distinguishing polarity-bearing topics can lead to more accurate sentiment classification (§3, §4) , aspect extraction (§3), and how focusing pivotal entities in text can lead to better question-answering models (§6).

1.3 Contributions

The work in this thesis addresses the research objectives outlined making the following contributions:

C. 1 We introduce a new neural architecture, in particular, a *topic-dependent attention model* (TDAM) (§ 3), to combine the word global and local contexts by means of a new neural cell employing an auxiliary memory to keep track of the word occurrences across documents (**RO 1**), while simultaneously encoding the word embedding depending on the surrounding words (**RO 3**). The resulting embeddings are an accurate encoding of the themes in the corpus (**RO 2**), able to discriminate between the difference aspects discussed (**RO 5**).

C. 2 We design a probabilistic topic model, called *Context-aware Pólya urn model* (Context-GPU) (§ 5), to generate topics composed of topical phrases (**RO 2**) by combining the local and global context of words/phrases (**RO 1**, **RO 4**). The Context-GPU leverages the Pólya urn model to corroborate the word global and local contexts, determining the quality of the resulting topics. The window-based embedding not only improves the capability to detect semantic relatedness at the phrase level, but it also encodes word co-occurrences from an external source of knowledge (e.g., Wikipedia or the PubMed corpus) alleviating the lack of statistics for technical terms (**RO 3**).

C. 3 We propose a new model, namely a *disentangled adversarial neural topic model* (DIATOM) (§ 4), which is able to generate disentangled topics (**RO 2**) through the combination of a variational autoencoder and adversarial learning. We conduct an experimental assessment of the topic quality (**RO 5**), using more traditional topic quality scores (such as topic coherence, topic uniqueness, and perplexity), and devising a novel approach to measure the topic disentanglement based on the particular type (e.g., opinion or plot/neutral).

C. 4 We introduce the MOBO dataset (**RO 5**), a new collection of movie and book reviews paired with their plots, with annotated sentences which provide a research tool for the evaluation of topic types via topic labelling (§ 4).

C. 5 We introduce a *biomedical entity-aware masking* (BEM) strategy (§ 6) encouraging masked language models (MLMs) to learn entity-centric knowledge (§ 6.3). We first identify a set of entities characterising the particular domain (**RO 4**), using a biomedical entity recogniser (SciSpacy [126]), and then employing a subset of those entities to drive the masking strategy while fine-tuning. The resulting mechanism implicitly drives the encoding of new relations about biomedical entities, leveraging the named entity recogniser, and results in improved accuracy on several biomedical question-answering tasks (**RO 5**).

1.4 Publications

The work discussed in this dissertation relates primarily to the following articles (in order of publication):

- **Gabriele Pergola**, Lin Gui, Yulan He. *"A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews"*. In Proceedings of The North American Chapter of the Association for Computational Linguistics (NAACL) conference, 2021.

- **Gabriele Pergola**, Elena Kochkina, Lin Gui, Maria Liakata, Yulan He. *"Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies"*. In Proceedings of The the European Chapter of the Association for Computational Linguistics (EACL) conference, 2021.

- **Gabriele Pergola**, Lin Gui, Yulan He. *"TDAM: a Topic-Dependent Attention Model for Sentiment Analysis"*. Information Processing & Management, 2019.

- **Gabriele Pergola**, Yulan He, David Lowe. *"Topical Phrase Extraction from Clinical Reports by Incorporating both Local and Global Context"*. In Proceedings of The 2nd AAAI Workshop on Health Intelligence (AAAI18), 2018.

Below follows a list of co-authored publications, not included in this thesis:

- Runcong Zhao, Lin Gui, **Gabriele Pergola**, Yulan He. *"Adversarial Learning of Poisson Factorisation Model for Gauging Brand Sentiment in User Reviews"*. In Proceedings of The European Chapter of the Association for Computational Linguistics (EACL) conference, 2021.

- Junru Lu, **Gabriele Pergola**, Lin Gui, Binyang Li, Yulan He. *"CHIME: Cross-passage Hierarchical Memory Network for Generative Review Question Answering"*. In Proceedings of The 28th International Conference on Computational Linguistics (COLING), 2020.

- Lin Gui, Jia Leng, **Gabriele Pergola**, Yu Zhou, Ruifeng Xu, Yulan He. *"Neural Topic Model with Reinforcement Learning"*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), 2019.

1.5 Thesis Outline

We begin with CHAPTER 2 introducing the relevant background on probabilistic and neural topic modelling, along with an overview of masked language models. This chapter provides a brief review of the Latent Dirichlet Allocation (LDA) [16], pointing out the concepts and notation needed to outline the Gibbs Sampling [57] and the Pólya Urn Model [109]. In turn, these form the basis for the *Context-aware Pólya urn model* (Context-GPU) introduced in CHAPTER 5. Then, there follows an introduction to the neural variational inference [84], and its adaptation to a basic neural topic model [155]. The chapter ends with an analysis of neural language models [36] and the different masking strategies adopted to train them.

CHAPTER 3 introduces a topic-dependent attention model (TDAM). This is a hierarchical neural architecture inspired by the Hierarchical Attention Model (HAN) [194]. It is trained in a multi-task learning setting to combine the word global and local contexts with a modified GRU cell [29], and leverages an auxiliary memory to keep track of the word occurrences across documents used for the topic extraction. This chapter is based on the published work of Pergola et al. 2019.

CHAPTER 4 describes a novel topic model combining the adversarial training [55] with the neural topic model architecture [155] to distinguish between polarised opinion topics and topics about factual descriptions (e.g., movie and book plots). In this chapter, we also present the new MOBO dataset, a new collection of movie and book reviews paired with their plots and with manually annotated sentences, used to determine the topic type via topic labelling. This chapter is based on the published work of Pergola et al. 2021.

CHAPTER 5 presents the *Context-aware Pólya urn model* (Context-GPU), a probabilistic topic model based on the Pólya urn framework [123]. It modifies the topic inference by employing a weighting schema, where the word weights are determined simultaneously by the co-occurrence statistics and the word embedding similarity. To highlight the model effectiveness in the biomedical domain, we analyse clinical notes and generate topics composed of relevant medical phrases, overcoming the unigram limitation affecting traditional bag-of-words models [16]. The chapter is based on the published work of Pergola et al. 2018.

CHAPTER 6 introduces a *biomedical entity-aware masking* (BEM) strategy to fine-tune masked language models [36, 91] by leveraging the pivotal entities characterising the target domains. This masking strategy consists of a first step to detect biomedical entities in a corpus of interest by using a specialised named entity recogniser [126]. Then, these pivotal entities are chosen as tokens to be masked during the fine-tuning process, driving the model to a realignment of the word representations based on the medical entities. This results in improved performance on several biomedical QA tasks [164], traditionally characterised by a scarce availability of training resources [161]. The chapter is based on the published work of Pergola et al. 2021.

CHAPTER 7 finally summarises the contributions of each chapter, the limitations of the current work, and highlights some promising future research directions.

Chapter 2

Background

Chapter Abstract

In this chapter, we first introduce the background concepts and notation for probabilistic and neural topic modelling, with regard to the models presented in Chapter 3, 4 and 5. Then, there follows an overview of the distributional language representations, such as word embeddings and language models, whose concepts are particularly relevant for Chapter 3 and 6. Finally, we conclude this background chapter with an overview of the linguistic features characterising domain-specific text, of which medical documents are a prominent example analysed in Chapter 5 and 6.

2.1 Probabilistic Topic Modelling

Probabilistic topic models have established themselves as effective tools to generate expressive and concise representations of the main themes in text collections [16]. To this aim, they posit that data are generated according to some underlying probabilistic process, which in the case of text entails that documents are assumed to be the results of an underlying generative process depending on the model variables and parameters. Once a satisfying model, according to some metric, has been inferred, its variables encode the latent semantics within the analysed documents and their relations.

One of the most influential works in topic modelling has been the *latent Dirichlet allocation* (LDA) [16], where Blei et al. devised a model in which *topics* within a corpus are distributions over words, and in turn, documents are finite mixtures of these topics. LDA has given rise to a wide spectrum of extensions and applications, both in unsupervised [69] and supervised settings [176].

Most of them are unsupervised learning algorithms designed to automatically mine meaningful sets of words (i.e. topics) sharing a common semantics, yet several supervised variants have been proposed to improve the topic expressiveness or exploit the available meta-information for more label-oriented topical features [116].

All the aforementioned models work under the so-called *bag-of-words* assumption [16], which implies that neither the order of words in a document nor the order of documents matter. Despite the counter-intuitiveness of this assumption, LDA has remarkably succeeded in inferring the semantic structure of texts in several application scenarios [14]. However, some relevant features of the sentence structures remain still overlooked and additional details embedded through phrases might be ignored. This leads to a narrowed quality of topics for documents characterised by limited text and contextual information (e.g. tweets, user posts, etc.) [190]. Moreover, the bag-of-words assumption frequently entails the adoption of unigrams rather than *n*-grams which in turn leads to a sparseness problem due to the frequency distribution commonly characterising text in "natural language", leaving out structural information regarding the compositional semantics of text [96].

In the literature, it is possible to highlight at least two main strategies that attempt to tackle such limitations. Firstly, some extensions of LDA were designed to explicitly take into account the word order [173], or the document order when relevant (e.g. analyses of historical document sequences) [15]. Secondly, approaches were developed focusing on the preprocessing of documents. For instance, in order to overcome the text sparsity arising from a large number of short messages, documents are aggregated into long pseudo-documents that can be analysed as a whole [62].

2.1.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) [16] is a probabilistic generative model of corpora: documents are represented as random mixtures over topics, where each latent topic is a distribution over words. LDA is defined as a Bayesian model, and thus allows to determine the model's parameters via Bayesian inference. As a result, the document-topic distributions $\boldsymbol{\theta}$ and the topic-word distributions ϕ, which are the main parameters of the model, are treated as random variables. The Bayesian framework provides a rich and well-defined set of probabilistic techniques to reason about distributions over parameters conditioned on data (*posterior)*, and to infer the *likelihood* of different parametrisations of the model.

In particular, LDA defines some prior distributions over these parameters, $P(\theta|\boldsymbol{\alpha})$ and $P(\phi|\boldsymbol{\eta})$, where the parameters of the prior distributions are called *hyperparameters*. The advantage of being able to tune the priors of the model's distributions is to bias the parameters in favour of a simpler and more general model, preventing overfitting and driving a better generalisation to unseen data. Another advantage is the possibility to guide the parameters towards values influenced by the human expertise in a domain [71]. A commonly adopted prior in text modelling, including topic modelling, is the *Dirichlet distribution* [16, 199], due to its mathematical and geometrical properties along with its influence on the quality of the topic inferred [172].
The LDA model can be described through its *generative story*[1] as follows:

1. For k $\in$ {1,...,K}:
   (a) Draw K topic distributions,
   $\phi^{(k)} \sim$ Dirichlet(η)
2. For each document $d \in$ {1,...,D}:
   (a) Draw the document-topic distribution for d,
   $\theta_d \sim$ Dirichlet(α)
   (b) For each word token $i \in d$:
      i. Draw the new topic z_i,
      $z_{i,d} \sim$ Discrete(θ_d)
      ii. Draw the word value w_i from the topic-word distribution,
      $w_{i,d} \sim$ Discrete$(\phi^{(z_i)})$

The described process corresponds to the following joint likelihood over the model random variables:

$$P(\boldsymbol{d}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{\alpha}\, \boldsymbol{\eta}) = \prod_{i=1}^{I_d} P\left(z_{di} \mid \theta_d\right) P\left(w_{di} \mid z_{di}, \phi_k\right) \prod_{k=1}^{K} P\left(\phi_k \mid \boldsymbol{\eta}\right) \prod_{d=1}^{D} P\left(\theta_d \mid \boldsymbol{\alpha}\right) \tag{2.1}$$

with z_{di} being the topic assigned to the word token w_i in d. The unobserved variables to be learned are z, θ and ϕ, while the only observed data is $\boldsymbol{d}$.

[1] A *generative story* is a concise description of how the generative model assumes that its variables are generated.

The Bayesian inference process can be formalised by specifying the posterior distribution:

$$P(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{P(\boldsymbol{d}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}{P(\boldsymbol{d} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})} \tag{2.2}$$

Some approaches have been proposed to learn the specific values (i.e. points of estimates) which maximise the model posterior (*maximum a posteriori* - MAP) or likelihood (*maximum likelihood estimation* - MLE) [35]. However, we can generally obtain a better estimate by computing the actual full posterior distribution. To do so, we notice that given a set of variable assignments, the numerator in Equation 2.2 can be directly evaluated, the denominator is instead intractable due to the integration over all the possible assignments of $\boldsymbol{\theta}, \boldsymbol{\phi}$ and $\boldsymbol{z}$. We can then overcome the intractability with approximate inference techniques that estimate the full posterior distribution. Two commonly used techniques are the variational methods [16, 77], and the Markov Chain Monte Carlo (MCMC) methods [57]. Variational methods involve an optimisation process to find an approximation of the true posterior within a family of distributions analytically tractable, while MCMC methods determine a sampling process to directly estimate the true posterior. As a result, the variational methods are generally faster, while the MCMC ones guarantee to converge asymptotically to the actual posterior values (although we do not know how long the process would require).

2.1.2 Inference

LDA is a Bayesian model and can rely on the Bayesian inference framework to infer the model's parameters. However, as shown, the inference process becomes non-trivial as the model complexity and structure increases [69]. Thus, we proceed with a high-level description of the Bayesian inference for probabilistic models and then provide a brief overview of the methods that can be employed to overcome the intractability issues discussed in the previous paragraph.

For a model in the Bayesian framework, we can specify a set of unknown parameters or latent variables $\boldsymbol{z}$ of interest along with the prior distribution $p(z)$ modelling our knowledge about them before we analyse data. Subsequently, we can specify a likelihood function $p(x|z)$ to quantify how much of the data x are related to z, and finally compute the posterior distribution $p(z|x)$ by applying the Bayes' rule: $p(z|x) = p(z)p(x|z)/p(x)$. The criterion most commonly used to guide the probabilistic models through the learning process is the estimation of the *maximum log likelihood*. Under this criterion, we try to infer the model's parameters

θ that maximise the sum of the log probabilities that the model assigns to the data (Equation 2.3).

A directed graphical model with parameters θ would represent a joint distribution $p_\theta(x, z)$ over both the observed variable x and the latent ones z, analogously to what is described for LDA in Equation 2.1. The marginal distribution $p_\theta(x)$ results are defined as follows:

$$p_\theta(x) = \int p_\theta(x, z)dz = \int p_\theta(z)p_\theta(x|z)dz \tag{2.3}$$

However, the marginal probability $p_\theta(x)$ is typically intractable, making the maximum likelihood learning difficult: while the likelihood $p_\theta(x|z)$ expresses the probability of data given the specified z, the marginal likelihood $p_\theta(x)$ measures how probable is x over the entire latent space of $\boldsymbol{z}$, resulting in a prohibitive and high dimensional integration, with no analytic solution [154]. Consequently, the $p_\theta(x)$ already shown intractable makes the posterior density $p_\theta(z|x)$ intractable as well, considering the Bayesian relation: $p_\theta(z|x) = p_\theta(z)p_\theta(x|z)/p_\theta(x)$. To circumvent this intractability issue, we need to resort the mentioned methods for inference approximation, such as *variational inference* and *Markov Chain Monte Carlo* (MCMC).

The *variational inference* casts the Bayesian inference as an optimisation problem introducing a parametrised posterior approximation $q_\theta(z|x)$; this is fit to the posterior distribution through the θ parameters, chosen to maximise a lower bound $\mathcal{L}$ on the marginal likelihood:

$$\log p(x) \geq \log p(x) - D_{\mathrm{KL}}(q_\theta(z|x)||p(z|x)) \tag{2.4}$$

$$= \mathbb{E}_{q_\theta(z|x)}[\log p(x, z) - \log q_\theta(z|x)] = \mathcal{L}. \tag{2.5}$$

where z are the latent variables of interest, $p(z)$ their prior distribution and x the observed data. If we keep maximising the bound $\mathcal{L}$ with regard to θ, we will minimise the KL-Divergence term, given that $\log p(x)$ is independent of θ.

Alternatively, the *Markov Chain Monte Carlo* (MCMC) method starts by taking a random draw z_0 from some initial distribution $q(z_0|x)$, but rather than optimising this distribution it applies a *stochastic transition operator* to z_0:

$$z_t \sim q(z_t|z_{t-1}, x). \tag{2.6}$$

The iterative applications of an appropriately chosen transition operator will result in a random variable z_T converging in distribution to the exact posterior $p(z|x)$. Using the MCMC we can approximate the exact posterior arbitrarily well applying the stochastic transition operator a sufficient number of times. While this makes MCMC methods asymptotically exact and is one of the significant advantages of using them, in practice, we do not know in advance how many times to keep iterating them, with an overall process that could require a rather long time depending on the application [151].

We proceed introducing an MCMC algorithm, called *Gibbs sampling*, that forms the basis for the proposed *Context-Generalised* Pólya Urn model [109, 123].

2.1.3 Gibbs Sampling

One of the most prominent methods for the inference approximation of the posterior is the Collapsed Gibbs Sampling (CGS) [49, 57, 58]. The CGS is an MCMC method increasingly adopted and whose simplicity has led to a wide variety of implementations and parallel architectures [117, 153]. The core idea behind it is to marginalise (*"collapse"*) out the parameters of the topic-word distributions $\boldsymbol{\phi}$ and document-topic distributions $\boldsymbol{\theta}$, and approximate these distributions through a sampling procedure of the model's latent variables $\boldsymbol{z}$ [57].

The Gibbs Sampling process considers in turn each word token w_i in a text collection, and conditioned on the current topic assignments for the other tokens, estimates the probability of assigning w_i to each topic k. A topic is than drawn and assigned to the current word token i following the conditional distribution: $P(z_i = k|\boldsymbol{z}_{-i}, w_i, d_i, \cdot)$, where z_{-i} denotes the topic assignments of others word tokens, "$\cdot$" all the other known indices $\boldsymbol{w}_{-i}$, $\boldsymbol{d}_{-i}$ along with the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ of the symmetric Dirichlet priors for topics and documents. To compute the mentioned conditional distribution, CGS maintains two count matrices: n_{kv}, representing the number of times that a word type v is assigned to a topic k, and n_{dk}, counting the number of word tokens in a document d assigned to a topic k. As was shown in Griffiths et al. (2004), these assignments can be ultimately computed using the n_{kv} and n_{dk} matrices:

$$P(z_i = k|\boldsymbol{z}_{-i}, w_i, d_i, \cdot) \propto \frac{n_{kv} + \eta}{\sum\limits_{v'=1}^{V} n_{kv} + V\eta} \cdot \frac{n_{dk} + \alpha}{N_d + \sum\limits_{k'=1}^{K} K\,\alpha_{k'}} \tag{2.7}$$

The Gibbs sampler is first run for a few iterations during which the Markov chain is typically in a low probability state (*burn-in* period). After this short period, it stabilises and starts retrieving the reliable estimates for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ and topic assignments z_i. The duration of the CGS process is finally determined by a fixed number of iterations, set a priori, during which the assigned topic $k \in \{1, ..., K\}$ to a word token w_i is sequentially updated following the aforementioned conditional probability.

2.1.4 Pólya Urn Model

The LDA inference process is intrinsically biased to promote together words that frequently occur in a corpus, overlooking less prominent but still correlated words, a well-known problem in literature [123]. To alleviate this shortcoming and increase the association strength between rare but still correlated words, a different inference process was proposed by Mimno et al. (2011) following the so-called *Generalised Pólya Urn* (GPU) model, that consists in incorporating a corpus-specific word co-occurrence metric into the generative process refining the original probabilities of related words under the same topic.

It is based on the interpretation of LDA as a Pólya urn model [109], a statistical model describing objects of interest (e.g. words or topics) in terms of coloured balls and urns. In particular, LDA follows the so-called *Simple Pólya urn* (SPU) model. During the main steps of the SPU generative scheme, a coloured ball is randomly drawn from an urn and is put back along with an additional new ball of the same colour to induce a self-reinforcement process known as "rich get richer": as a result, the probability of seeing a specific coloured ball from an urn increases every time this ball has been drawn.

Under the Pólya urn perspective, in the LDA generative process we have two types of urns: the document-topic and the topic-word urns. The document-topic urns hold balls whose colour corresponds to different topics in a document, while the coloured balls in the topic-word urns represent different words in a topic. As a result of our priors, initially, the document-topics urns contain α balls of K different colours (with K being the number of topics), and similarly, the topic-word urns contain β balls of V different colours (with V being the vocabulary size). The generative process proceeds as follows: a ball is extracted from the document-topic urn d_m, and its colour determines the new topic assignment z_i for the word w_i; then the ball is put back along with another ball of the same colour. Next, a ball is extracted from the topic-word urn z_i determining a new word $\hat{w}$ and, as before, the ball with an additional one of the same colour is put back into the urn. As a result, both the topic z_i and the word $\hat{w}$ increase their proportion in the document-topic and topic-word distribution, respectively.

The LDA generative process in terms of urns and balls can be described as follows:

1. For each document $d \in \{1,...,D\}$:
   (a) For each topic $k \in \{1,...,K\}$:
      i. Add α_{dk} balls of color k to the urn$_d$,
      $urn_d[k] = \alpha_{dk}$
2. For each topic $k \in \{1,...,K\}$:
   (a) For each word token $i \in \{1,...,V\}$:
      i. Add β_{ki} balls of color w_i to the urn$_k$,
      $urn_k[i] = \beta_{ki}$
3. For each document $d \in \{1,...,D\}$:
   (a) For each word token $i \in d$:
      i. Draw a ball and assign the topic k for the sampled color,
      $k \sim urn_d, \;\; z_i^{(d)} = k$
      $urn_d[k] = urn_d[k] - 1$
      ii. Draw the word value w_i from the topic-word distribution,
      $w_i \sim urn_k$
      $urn_k[i] = urn_k[i] - 1$
      iii. Place the ball back in the urn, along with a new ball of the same color,
      $urn_d[k] = urn_d[k] + 2$
      $urn_k[i] \;= urn_k[i] \;+ 2.$

The Pólya urn interpretation allows to easily modify the generative process and thus adapting the corresponding Gibbs sampling process accordingly. In Chapter 5, we describe in details the *Generalised Pólya urn model* [123] and the proposed *Context-aware Pólya urn model* (Context-GPU), a model leveraging word embeddings to determine the relevance of words in a topic (i.e. coloured balls), despite their low occurrences in corpus. As a result, the revised Pólya urn model provides a smooth mechanism to combine the local and global contexts characterising words in a corpus, joining the information from probabilistic topic models and word embeddings.

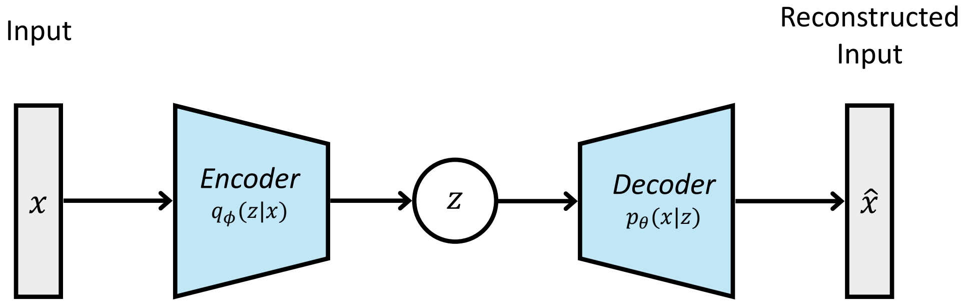


Figure 2.1: The Variational Autoencoder architecture.

2.2 Neural Topic Modelling

When designing a new topic model, even a small modification to the original components would require to re-derive again the variational inference method. This task could be even more demanding if the model needs to be scalable and parallelisable to cope with large datasets [17, 69]. Such a limitation has stimulated the development of neural variational inference approaches [84, 146] which can be easily adapted to new models given just the specification of the generative process.

One natural and promising application of these neural inference methods has been topic modelling. Mapping a document to a posterior distribution of latent topics is a task particularly suitable for neural models since these are universal function approximators, inherently oriented to generate "well-behaved" mapping, such that small changes in a document will produce only small changes in topics.

In the following, we first give a brief introduction to the *Variational Autoencoder* (VAE), a neural architecture based on the Autoencoding Variational Bayes (AEVB) [84] inference method. Next, we proceed by analysing a simple neural topic model based on the variational autoencoder architecture [155]. This would set the needed background to introduce the *Disentangled Adversarial Topic Model* (DIATOM) in Chapter 4.

2.2.1 Variational Autoencoder

Variational Autoencoders (VAEs) marry probabilistic graphical models with neural models [84].

As a neural model, VAEs directly inherit the architectural design from the Autoencoders (AEs) [66], to perform dimensionality reduction for unsupervised representation learning. It consists of two couple, but independently parametrised models: the *encoder* and the *decoder*. The resulting reconstruction mechanism forces the model to infer compact representations of data encoding the essential information to reconstruct them, and capture the meaningful factors of variations [10]. However, classic autoencoders do not necessarily rely on continuous representations of data and tend to have limited interpolation capability. Instead, VAEs are designed to project the input data into a continuous latent space, allowing easy interpolation and random sampling; consequently, for a given input x, the generative network is exposed to a range of variations associated with the same input, forcing the decoder to not just reconstruct the input data but to perform an interpolation in the continuous space. Although several autoencoder variations have been proposed to improve their generalisation [147, 170], VAEs have shown consistent generalisation performance in several applications [56]. In terms of graphical models, the encoder component $p_\theta(x|z)$ is a conditional Bayesian network of the form $p(\boldsymbol{z}|\boldsymbol{x})$, and the decoder component is a also a Bayesian network of the form $p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. Each of these conditionals are determined with a complex neural models. E.g. if f is a neural network, then $\boldsymbol{z}|\boldsymbol{x} \sim f(\boldsymbol{x}, \epsilon)$, with ϵ a noise random variable. The resulting learning procedure is a mix of traditional expectation maximisation, which thanks to the reparameterisation trick is performed through backpropagation on the neural layers [84, 146, 150]. A schematic depiction of a standard VAE architecture is depicted in Figure 2.1.

A latent variable model $p_\theta(x, z)$ with distributions parametrised by neural networks is also called a *deep latent variable model* (DLVM). A major advantage of DLVM models is that even though the prior or conditional distributions in the directed model are relatively simple (e.g. conditional Gaussian), the marginal distribution $p_\theta(x)$ can approximate complicated underlying distribution with almost arbitrary dependencies. The *Variational Autoencoder* [84] can be viewed as a DLMV model coupling an encoder (or *inference network*) $q_\phi(z|x)$ with a decoder (or *generative network*) $p_\theta(x|z)$.

Inference

To address the intractable posterior inference (Equation 2.3), in VAEs we leverage the parametric inference model $q_\phi(z|x)$ and optimise the variational parameters ϕ so that:

$$q_\phi(z|x) \approx p_\theta(z|x) \tag{2.8}$$

This approximation allows us to optimise and derive a lower bound on the marginal likelihood $p(x)$. In most cases, the encoder is just a single neural model used to perform posterior inference over all (or large subsets of) the samples in the dataset and optimises its neural weights and biases included as variational parameters ϕ of the distribution $q_\phi(z|x)$. This is called *amortized* variational inference [50], and it is one of the advantage compared to more traditional variational inference methods, avoiding a per-sample optimisation loop and leveraging the *stochastic gradient descent* (SGD) [19, 183] efficiency.

We can rewrite the Equation 2.3 for the marginal likelihood in terms of VAE components as follows:

$$\log p_\theta(x) =$$

$$= \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x) \right] \tag{2.9}$$

$$= \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{p_\theta(x,z)}{p_\theta(z|x)} \right] \right] \tag{2.10}$$

$$= \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{p_\theta(x,z)}{q_\phi(z|x)} \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \right] \tag{2.11}$$

$$= \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{p_\theta(x|z) p_\theta(x)}{q_\phi(z|x)} \right] \right] + \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \right] \tag{2.12}$$

$$= \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z) \right] - \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{q_\phi(z|x)}{p_\theta(z)} \right] \right] + \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \right] \tag{2.13}$$

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z) \right] - D_{KL}(q_\phi(z|x) \,||\, p_\theta(z))}_{\substack{= \mathcal{L}_{\theta,\phi}(x) \\ \text{(ELBO)}}} + \underbrace{D_{KL}(q_\phi(z|x) \,||\, p_\theta(z|x))}_{\geq 0} \tag{2.14}$$

In Equation 2.14, the first term is the objective to be optimised from the VAE, it is common to several variational methods and generally called *variational lower bound* or *evidence lower bound* (ELBO). The second term is the non-negative Kullback-Leibler (KL) divergence between $q_\phi(z|x)$ and the intractable $p_\theta(z|x)$. Due to the aforementioned non-negativity, the ELBO is a lower bound on the log-likelihood of the data, which is the reason why the ELBO is commonly used as an optimisation objective of VAEs. The two terms composing the ELBO in Equation 2.14 are the conditional distribution $p_\theta(x|z)$ of the generative network and the KL divergence between the approximation $q_\phi(z|x)$ and the prior $p(z)$. The former can be conveniently compute using sampling techniques, while the latter has a closed-form solution when adopting Gaussian distributions for the encoder and the prior, resulting in an overall tractable and differentiable lower bound.

We can rewrite Equation 2.14 by omitting the last term as follows:

$$\log p_\theta(x) \geq \mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}(q_\phi(z|x) \,||\, p_\theta(z)) \tag{2.15}$$

From Equation 2.15, we notice that maximising the ELBO with regard to the parameters θ and ϕ will concurrently maximise the marginal likelihood and minimise the KL divergence between the approximation $q_\phi(z|x)$ and $p_\theta(z|x)$, resulting in a model better fitting the data with an improved $q_\phi(z|x)$ approximation.

This is the core idea of the variational autoencoder: it maximises the ELBO as a proxy to maximise the likelihood of data, while simultaneously regularising the process with a constraint on the form of the approximate posterior through the KL divergence term.

An efficient approach for optimising the ELBO with regard to θ and ϕ is using stochastic gradient descent (SGD). The overall ELBO results from the sum of the ELBO on the single N samples:

$$\theta^*\phi^* = \underset{\theta,\phi}{\arg\max} \sum_{n=1}^{N} \mathcal{L}(x^n, \theta, \phi). \tag{2.16}$$

We can then simply derive unbiased gradients of the ELBO with regard to the generative model parameters θ as follows:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta},\phi}(\mathbf{x}) &= \nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right] && (2.17)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\nabla_{\boldsymbol{\theta}}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right)\right] && (2.18)\\
&\simeq \nabla_{\boldsymbol{\theta}}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right) && (2.19)\\
&= \nabla_{\boldsymbol{\theta}}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right) && (2.20)
\end{aligned}$$

with $\mathbf{z}$ being, in this particular case, a random sample from $q_\phi(z|x)$.
Unbiased gradients for the inference network instead are not obvious to derive, since in this case the distribution $q_\phi(z|x)$ cannot be easily derived being a function of ϕ:

$$\begin{aligned}
\nabla_{\phi}\mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \nabla_{\phi}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right] && (2.21)\\
&\neq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\nabla_{\phi}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right)\right] && (2.22)
\end{aligned}$$

We cannot compute the gradient, and thus perform the backpropagation algorithm, for a process including random sampling. An effective workaround consists of operating a change of variables, known in the literature as *reparameterisation trick* [84, 146], which moves the sampling out to an input layer. We examine as an

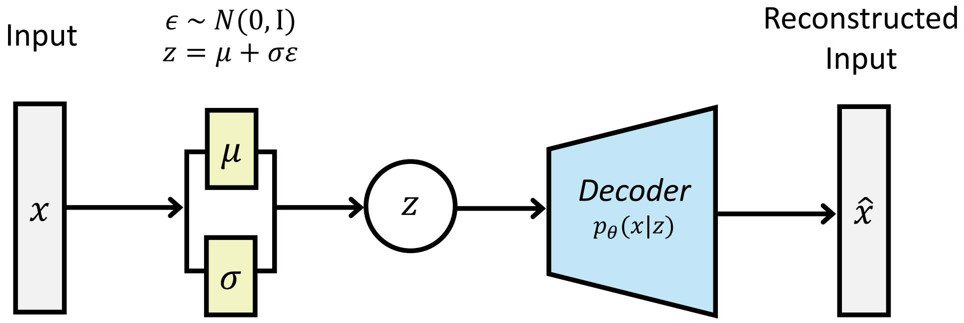


Figure 2.2: The Variational Autoencoder architecture highlighting the *reparameterization trick* [84].

example the commonly adopted Gaussian encoder $q_\phi(z|x) = \mathcal{N}(z; \mu, \mathrm{diag}(\sigma^2))$, that uses the neural encoder to determine the set of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Before applying the reparameterization, it is defined as:

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}) = \mathrm{NeuralEncoder}_\phi(x) \tag{2.23}$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, diag(\boldsymbol{\sigma})) = \prod_i \mathcal{N}(z_n; \mu_n, \sigma_n^2) \tag{2.24}$$

with $\mathcal{N}$ being the PDF of the univariate Gaussian distribution. Then, applying the reparameterization:

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{2.25}$$

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}) = \mathrm{NeuralEncoder}_\phi(\mathbf{x}) \tag{2.26}$$

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon \tag{2.27}$$

with $\odot$ denoting the element-wise product. This leads to a new posterior $q_\phi(\mathbf{z}|\mathbf{x})$ in terms of ϵ_n rather than z_n, and therefore differentiable [84]. Figure 2.2 depict the variational autoencoder architecture with the highlighted *reparameterization* mechanism.

2.2.2 A Simple Neural Topic Model

We now introduce a basic neural topic modelling (NTM) architecture, as proposed in Srivastava et al. (2017), describing how the VAE framework can be adopted to explicitly infer topics within documents. In the wake of this work, many recent variants have been proposed to incorporate more sophisticated priors [118], adopting different metrics [125], to allow supervision [22], or using graph-based methods [205].

In terms of priors, rather than using a Dirichlet prior as in LDA, in the NTM we employ a logistic normal prior on $\boldsymbol{\theta}$ to facilitate the inference [155]. Then, for a document d, we draw the latent variable $\boldsymbol{z_d}$ from a multivariate normal, transformed via a *softmax* function to lie on the simplex.
The generative story is described in the following:

1. For each document $d \in$ {1,...,D}:
   (a) Draw a document-topic distribution from a logistic normal prior,
   $(\boldsymbol{\mu_\phi}, \log \boldsymbol{\sigma_\phi}) = \text{NeuralEncoder}_\phi(x)$
   $\boldsymbol{z}_d \sim \mathcal{N}(\boldsymbol{\mu_\phi}, \boldsymbol{\sigma^2}), \quad \boldsymbol{\theta_d} = f_\phi(\boldsymbol{z}_d)$
   (b) For each word token $w_i \in d$:
      i. Draw the word value w_i from the topic-word distribution,
      $w_{i,d} \sim p(w \mid \boldsymbol{\theta}_i)$
   (c) Generate labels, if available,
   $\boldsymbol{y_d} \sim p(\boldsymbol{y} \mid f_y(\boldsymbol{z_d}))$

where f_ϕ and f_y are MLPs, and θ_d is a K-dimensional latent topic representation for d. The probability of a word $w_{i,j}$ can be parametrised by a *softmax* function or another MLP.

The inference process follows what we previously described for the VAE inference. In particular, each document d is assumed to have a latent representation z_d, which can be interpreted as its membership to each topic, and can be inferred through the VAE sampling process [84]. We assume a variational approximation to the posterior $q_\phi(\boldsymbol{z_d}, \boldsymbol{w_d}, \boldsymbol{y_d}))$, and aim to minimise the KL divergence with the true posterior $p(\boldsymbol{z_d}, \boldsymbol{w_d}, \boldsymbol{y_d}))$, with $\boldsymbol{\phi}$ variational parameters of a neural model to be optimised. Following analogous manipulations to the one described in §2.2.1, we can derive the ELBO [22, 84]:

$$\begin{aligned} \log p_z(d) = \; & \mathbb{E}_{q_\phi(\boldsymbol{z}_d | \boldsymbol{d}, \boldsymbol{y}_d)} \left[\sum_d^D log\, p(w_{i,d} \mid \boldsymbol{z}_d) \right] \\ & + \mathbb{E}_{q_\phi(\boldsymbol{z}_d | \boldsymbol{d}, \boldsymbol{y}_d)} \left[log\, p(\boldsymbol{y}_d \mid \boldsymbol{z}_d) \right] \\ & - D_{KL}(q_\phi(\boldsymbol{z}_d \mid \boldsymbol{d}, \boldsymbol{y}_d) \,||\, p(\boldsymbol{z}_d)) \end{aligned} \tag{2.28}$$

The normal prior on $\boldsymbol{z}$ takes the form of a network which outputs two vectors, $\mu_d = f_{\boldsymbol{\mu}}(d, \boldsymbol{y}_d)$ and $\boldsymbol{\sigma}^2 = f_{\boldsymbol{\sigma}}(d, \boldsymbol{y}_d)$, for a resulting approximate posterior

$q_\phi(\boldsymbol{z}_d \mid d,\, \boldsymbol{y}_d) = \mathcal{N}(\boldsymbol{\mu}_\phi,\, \boldsymbol{\sigma^2})$.
The resulting neural computations [118, 155] are:

$$\boldsymbol{\pi}_d = f_e([\boldsymbol{W}_x \boldsymbol{x}_d; \boldsymbol{W}_y \boldsymbol{y}_d;]) \tag{2.29}$$

$$\boldsymbol{\mu}_d = \boldsymbol{W}_\mu \boldsymbol{\pi}_d + \boldsymbol{\beta}_\mu \tag{2.30}$$

$$\log \boldsymbol{\sigma}_d^2 = \boldsymbol{W}_\sigma \boldsymbol{\pi}_d + \boldsymbol{\beta}_\sigma \tag{2.31}$$

where $\boldsymbol{x}_d$ is the document-term vector with the word w_i frequency within the document d, and f_e a multilayer perceptron. The weights and bias matrices and the f_e parameters are part of the variational parameters ϕ being optimised.
The intractable expectations in Equation 2.28 needs to be approximated, and to preserve the differentiability with respect to ϕ, we apply the reparameterization trick [84]:

$$\epsilon^{(s)} \sim \mathcal{N}(0, \mathbf{I}) \tag{2.32}$$

$$\mathbf{z}_d^{(s)} = \boldsymbol{\mu}_d + \boldsymbol{\sigma}_d \odot \epsilon^{(s)} \tag{2.33}$$

Thus, substituting the Monte Carlo approximation computed for a single sample of ϵ in Equation 2.28:

$$\log p_z(d) \approx \sum_d^D log\, p(w_{i,d} \mid \boldsymbol{z}_d^{(s)})$$

$$+\ log\, p(\boldsymbol{y}_d \mid \boldsymbol{z}_d^{(s)}) \tag{2.34}$$

$$-\ D_{KL}(q_\phi(\boldsymbol{z}_d \mid d, \boldsymbol{y}_d) \,||\, p(\boldsymbol{z}_d)) \tag{2.35}$$

This approximation can be optimised using stochastic gradient descent with respect to the model's parameters specified, with the KL divergence in Equation 2.35 that can be computed in a closed form.

2.3 Distributed Language Representations

In this section, we briefly review the foundations and current works in modelling word and language distributed representations. We first analyse some of the most used word embedding methods, which have led to the development of a flourishing line of research on neural architectures for text analysis. Then, we introduce the deep-contextualised language models, a recently developed family of neural language models aiming at generating word representations depending entirely on the specific context surrounding them.

2.3.1 Word Embedding

In *word embeddings*, a continuous vector is used to encode the word meaning through the context in which the word is likely to occur; those word vectors yield a low-dimensional vector space whose dimensions can be interpreted as latent features describing syntactical or semantic properties. These word embedding methods are based on the so-called *'distributional hypothesis'* stating that word semantics is implicit in the word co-occurrences, so that words occurring in similar contexts would tend to share a similar meaning [45, 61].

Traditional word distributional representations had been devised relying on the construction of the *pointwise mutual information (PMI) matrix* [30] and its factorisation via singular value decomposition (SVD) [34]. PMI approaches build word matrices computing the pointwise mutual information between a word w_i and a set of context words (i.e. a set of words that frequently appeared within a context window). A subsequent variant was proposed, called *positive-PMI (PPMI)* [21], that simply replaces all the negative values in the PMI matrix with 0. As a result, PPMI was proven to outperform PMI on detecting word semantic similarities on several tasks [92]. A low-dimensional approximation of the PPMI matrix can be computed by applying a truncated singular value decomposition (SVD), yielding a more compact representation of word correlations. This can be interpreted as a generalisation of Latent Semantic Analysis (LSA) [34] where the PPMI matrix replaces the original term-document matrix [92].

Although those studies have proven the capability of encoding syntactic and semantic properties, the generation of word embeddings was too computationally expensive for large corpora, hindering their applications and their capability to encoded rich language statistics for a wide vocabulary. The neural-based methods were groundbreaking alternatives thanks to their efficiency in modelling text. Word embeddings, based on neural models, were originally introduced by Bengio et al. (2003) and Collobart et al. (2008), but were then widely adopted following a study by Mikolov et al. 2013. In particular, they introduced a novel design, known as the *skip-gram with negative-sampling training method (SGNS)*: an efficient neural embedding algorithm to compute word embeddings widely popularised by the *word2vec* software [121, 122]. This method overcomes the heavy computations required to scale to a higher number of documents by boosting the production of a vocabulary representation through neural representations [7, 175]; once generated, the vector representations can be not only inspected to discover word relations but also fed into new neural models as pretrained embeddings [10], allowing a smooth and cost-efficient methodology to incorporate knowledge into new models. Additional analysis showed that word2vec is implicitly factorising a PMI matrix [53] and that conventional approaches could achieve qualitative results comparable with word2vec when hyper-parameters are properly tuned [92], even though the required time complexity still remains prohibitively high. Following word2vec, several other word embeddings were proposed. One is *GloVe* [133], combining the advantage of both matrix factorisation and local context window approaches. Another one is *FastText*, proposed by Joulin et al. (2017). Although it achieves performance comparable with word2vec on semantic tasks, it is able to train word embeddings more efficiently and with greater generalisation capabilities. It is based on a hierarchical classifier that treats each word as made of character n-grams, yielding word vectors arising from the composition of n-gram representations. One of the advantages of this approach is a greater generalisation to out-of-vocabulary (OOV) words which, despite having never been seen before by the algorithm, can be represented leveraging some known prefixes or suffixes used as a clue of the word meaning.

2.3.2 Contextualized Language Models

The subsequent introduction of the "Transformer" architecture [169] has led to the development of state-of-the-art *contextualized* language models (LMs), such as *BERT* [36]. These new family of language models [141, 142, 193] is based on the Transformer architecture, a highly-parallelizable alternative to recurrent encoders (such as Long Short-Term Memory network [68]), with t introduction of novel objectives to bootstrap the training process (e.g. *masked language model*).

Compared to the Recurrent Neural Networks (RNN), the "Transformer" architecture relies solely on attention mechanisms, and uses an absolute-position embedding to mark words and keep track of their positions. It consists of multiple layers, where each layer contains multiple attention heads (e.g., BERT-*Base* has 12 layers with 12 attentions). For an input sequence of N tokens, an attention head takes as input a sequence of vectors $h = [h_1, ..., h_N]$; then, each vector h_n is transformed via separate linear transformations into query q_n, key k_n and value v_n

vectors. For all pairs of words, each head computes the attention weights α through a dot product (normalised with softmax) between the query and key vectors. Finally, the resulting output o_n of the attention head is defined as:

$$O = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.36}$$

with d_k being the dimensionality of the key vectors, used as normalising factor to scale the magnitude of the dot product and to ease the gradient flow. The resulting attention weights O will regulate to what extent the representation for the current token is going to be influenced by every other token.

BERT: Bidirectional Encoder Representations from Transformers

BERT is a neural language model, pretrained on large English corpora, such as the BookCorpus [206] and English Wikipedia with over 3.3 billion tokens, on two language tasks. One is the *masked language modelling* task, in which the model tries to predict the words that have been masked out in an input text. The other task consists of the *next sentence prediction*, where the model tries to predict whether a sentence follows a given sentence, or is instead just a random combination. Once the model has been pretrained in a semi-supervised manner, it can be further trained (i.e. *fine-tuned*) using labelled data; an approach that has led to the current state-of-the-art results across a wide variety of tasks [46, 91, 105, 159, 186, 196].

During the preprocessing of the data, BERT adds a special token `[CLS]` at the beginning of its input and another one, `[SEP]`, to the end. `[SEP]` can be further used to separate multiple texts within the same input: this is the case, for example, in the reading comprehension task, with the input consisting of a question and its related contest separated by `[SEP]`. These special tokens do not have just a formatting purpose but assume a rather relevant role during the training, for example, in the sentiment analysis tasks, where the `[CLS]` token is commonly used to perform the final classification.

Masking Strategies

In the wake of BERT, several new language models have been proposed, adopting different masking strategies.

As previously mentioned, BERT is a masked language model (MLM) since it drives the training process following a strategy that randomly replaces a predefined proportion of words with a special [MASK] token, which then need to be predicted based just the remaining context. BERT first chooses 15% of tokens uniformly at random, and swaps 10% of them into random tokens (resulting in an overall 1.5% of the initial tokens randomly swapped). This introduces a rather limited amount of noise with the aim of making the predictions more robust to trivial associations between the masked tokens and the context. While another 10% of the selected tokens are kept without modifications, the remaining 80% of them are replaced with the [MASK] token.

Several alternative strategies have been recently proposed with different impacts on the final performance. RoBERTa [104] introduced a slightly different training procedure: it removed the next sentence prediction task and showed that performance did not decrease, and at times even improved on some downstream tasks. The authors also introduced a *dynamic masking* where every time a sequence is fed to the model a different masking pattern is generated, compared to the *static* approach followed in the original BERT implementation, where each sample was masked once during preprocessing. These adjustments allowed a longer training, employing much larger corpora: compare to BERT, RoBERTa is trained on the BookCorpus [206], English Wikipedia, the CommonCrawl News dataset [104], the OpenWebText [52] and the Stories corpus [163].

In SpanBERT, Joshi et al. (2020) proposed to mask and predict spans rather than tokens. ERNIE [201] instead is focused on masking phrases and named entities to improve the structural knowledge encoded. Some of these techniques have been combined in the T5 model [142], an encoder-decoder transformer-based model sharing the same model, objective, and training process across multiple NLP tasks, all reframed as "text-to-text" problems: document summarisation, sentiment classification, question answering, machine translation and so on. Although like BERT, T5 uses a denoising approach for the masking strategy, it masks multiple tokens, somehow similarly to what was proposed in SpanBERT. Along with T5, Raffel et al. (2020) introduced a novel dataset, named *Colossal Clean Crawled Corpus* (C4), a cleaned version of the CommonCrawl dataset [104] of approximately $\sim 700\,\mathrm{GB}$ of text.

An alternative masking strategy is based on the *permutation language modelling* (PLM) task, proposed to train the XLNet model [193]. The aim of the permutation language model is to pretrain the LM without the need to rely on data corruption, i.e. to use a [MASK] token which though does not appear during the fine-tuning process. To avoid this discrepancy between the pretraining and fine-tuning phases, it instead minimises the expected log-likelihood of a sequence with regard to all possible permutations of the sequence order. In particular, for an input sentence of N tokens, there are $N!$ different orders that can be used to optimise the objective. This is implemented by preserving the original sequence order with the related positional embedding, while sampling a different factorisation order for each iteration, and then adjusting the Transformers attention mask to apply such permutations and perform the relative predictions.

Model Specialization

A wide spectrum of specialised language models has been recently developed [8, 23, 91, 196] due to the possibility to process a large variety of data to fine-tune the models towards different domains and tasks.

Among the tasks where contextualized language models has had a remarkable impact, we have question answering [143], machine reading comprehension [159], named entity recognition (NER) [46], sentiment analysis [196] and so on. There have also been a few attempts in combining BERT with neural topic models, leading to TopicBERT [25]. This aims at reducing the number of attention computations required by leveraging the topic signal provided from a neural topic model, preserving document classification performance on-par with BERT with a 40% speed-up in time requirements. In another work, proposing t-BERT [130], the authors considered the sentence BERT encoding jointly with its topic distribution, thus improving the topic-awareness of the model and, in turn, its document classification performance.

In terms of specialized domain, BERT has been adapted to deal with financial and legal documents [23, 192], with patents [90]], code [158] tweets [128], and several others [186]. Particular attention has been devoted to the medical domain, where different corpora and tasks still require different adaptation techniques. BioBERT [91] is a biomedical language model based on the BERT-*Base* variant [36], with additional pretrain on biomedical documents from PubMed and PMC collections, and uses the same training settings adopted in BERT. SciBERT [8] follows the BERT's masking strategy to pretrain the model from scratch using a scientific corpus composed of papers from Semantic Scholar [4]. Out of the 1.14M papers used, more than 80% belong to the biomedical domain. They both showed state-of-the-art result compared to the non-BERT SOTA on several tasks, as Named Entity Recognition, Question Answering, and Relation Extraction [8, 91]. ClinicalBERT [3], it is also based on the BERT-*Base* variant [36], but it is more focused on clinical documents. In particular, it is pretrained using the clinical notes from the MIMIC-III dataset [76]. BioMed-RoBERTa [60] is instead based on RoBERTa-*Base* [104] using a corpus of 2.27M articles from the Semantic Scholar dataset [4]. BlueBERT [132], follows an approach rather similar to the one described in BioBERT, built on top of the BERT-*Base* variant, but then pretrained on PubMed abstracts as well as clinical notes from the MIMIC-III dataset [76].

2.4 Text Processing for Medical Documents

In what follows, we briefly depict some of the main linguistic features characterising documents of technical domains. Although most of the features we highlight are tailored for documents in the medical domain, they reflect features shared among domain-specific text, such as patent, legal and technical documents, and so on. Most of the faced issues are common among several domains and the devised solutions can be easily adopted in different contexts. The medical domain is also one of the most challenging scenarios (i.e. abbreviations, medical jargon, relevant a priori knowledge, etc.) and thus an optimum benchmark for natural language models.

Although many studies have been conducted in analysing unstructured text data [16, 36], documents in technical domains are still difficult to be analysed (e.g. medical reports, patents, legal documents, etc.) [62, 95]. Corpora of these domains are frequently characterised by technical jargon, abbreviations and multi-word phrases, i.e. concepts unfolded across several words rather than a single word. Clinical notes are a prominent example of this family as medical concepts are often expressed in terms of multi-word phrases. For example, the phrases "*white blood cell*" or "*blood sugar*" would lose their meaning if decomposed as unigrams; the word *cell* and *sugar* might be wrongly put under the same topic because of the shared *blood* term.

2.4.1 Challenging Features of Technical Documents

In the following, we summarise some of the challenges characterising the task of processing unstructured text in technical domains:

Compositional semantics: multi-word phrases are often used in technical documents to refer to a specific concept. For example, in the medical domain, phrases like "*blood glucose*" or "*white blood cell*" would lose their meaning if split into word unigrams. However, because of the computational complexity involved, many models still rely entirely on the bag-of-words assumption by ignoring the word order.

Technical jargon: technical documents, such as medical or legal records, contain a large number of jargon and terms rarely used out of those particular contexts. This entails a lack of statistics affecting language models that, trained on a different and possibly more general domain, perform very poorly on these corpora. In addition, medical reports use a large dictionary of Greek and Latin words entailing Latin stems and inflexions (e.g. *basophil*ia, Synechococc*us* elongat*us*) that could be addressed taking into account the morphological structures of words and phrases, and needs to be processed differently when automatically performing stemming.

Abbreviations: medical documents contain a wide variety of abbreviations, acronyms and neologism. Those terms not only are domain-specific but often they can be expanded in more than one concept depending on the context at hand (e.g. SBP can be expanded both as "*spontaneous bacterial peritonitis*" or "*systolic blood pressure*"). Moreover, many abbreviations are expressed through punctuation (e.g. "p.o." which means "by mouth") that needs to be preserved when preprocessing data.

Polysemy: the same term can be used to refer to different concepts depending on the context. Some terms have a different meaning based on the global context in which they are used; for example, the word "*column*" can refer to a pillar or to a spine. Yet even within one specific domain, the same word can completely change its meaning; for example, the word "*inflammation*" can have at least five different meanings depending on the context in which it occurs [140].

Lack of structure: clinical notes, for instance, are free-text where physicians summarised their analyses. However, often they don't have a fixed structure, and although some notes are divided into sections, there is no standardisation.

Data availability: some of the current most effective models [23, 90, 91] need an intensive pretraining to work properly, using a large amount of data to then be possibly fine-tuned on the domain of interest. However, technical documents tend to be expensive to produce, requiring highly specialised staff, and they might contain sensitive data about people, events or institutions; hence, institutions owing large datasets refrain from releasing their data, and to date, in the medical domain, there are few publicly available medical text datasets that are suitable for pretraining models for specific tasks [76, 164, 177, 203].

Chapter 3

TDAM: a Topic-Dependent Attention Model

Chapter Abstract

In this chapter, we introduce a topic-dependent attention model (TDAM) for sentiment classification and topic extraction. TDAM assumes that a global topic embedding is shared across documents and employs an attention mechanism to derive local topic embedding for words and sentences. These are subsequently incorporated in a modified Gated Recurrent Unit (GRU) for sentiment classification and extraction of topics bearing different sentiment polarities. Those topics emerge from the words' local topic embeddings learned by the internal attention of the GRU cells in the context of a multi-task learning framework. We first introduce the related literature, then the hierarchical architecture, along with the new GRU unit. Finally, the experiments conducted on users' reviews demonstrate classification performance on a par with state-of-the-art methodologies in terms of sentiment classification and topic coherence for supervised topic extraction. In addition, our model is able to extract coherent aspect-sentiment clusters despite using no aspect-level annotations for training.

3.1 Introduction

In recent years, attention mechanisms in neural networks have been widely used in various tasks in Natural Language Processing (NLP), including machine translation [5, 106, 169], image captioning [187], text classification [26, 107, 194, 204] and reading

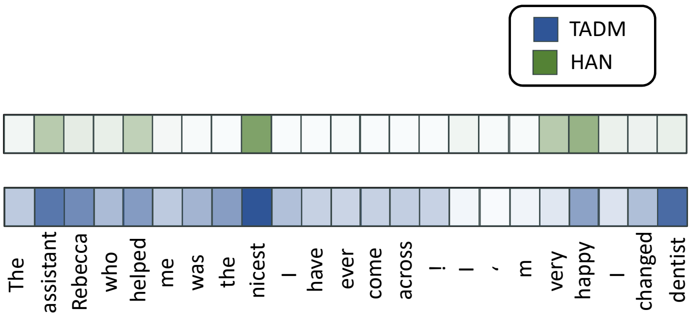


Figure 3.1: Attention weights from the *Topic-Dependent Attention Model* (TDAM) and *Hierarchical Attention Network* (HAN) [194]. TDAM highlights and gives more relevance to both sentiment and topical words.

comprehension [63, 180]. Attention mechanisms are commonly used in models for processing sequence data that instead of encoding the full input sequence into a fixed-length vector learn to "attend" to different parts of the input sequence, based on the task at hand. This is equivalent to giving the model access to its internal memory which consists of the hidden states of the sequence encoder. Typically soft attention is used which allows the model to retrieve a weighted combination of all memory locations.

One advantage of using attention mechanisms is that the learned attention weights can be visualised to enable an intuitive understanding of what contributes the most to the model's decision. For example, in sentiment classification, the visualisation of word-level attention weights can often give us a clue as to why a given sentence is classified as positive or negative. Words with higher attention weights can sometimes be indicative of the overall sentence-level polarity (for example, see Figure 3.1). This inspires us the development of a model for the extraction of polarity-bearing topics based on the attention weights learned by a model.

However, simply using the attention weights learned by the traditional attention networks such as the Hierarchical Attention Network (HAN) [194] would not give good results for the extraction of polarity-bearing topics, since in these models the attention weight of each word is calculated as the similarity between the word's hidden state representation with a context vector shared across all the documents. There is no mechanism to separate words into multiple clusters representing polarity-bearing topics.

Therefore, we propose a novel Topic-Dependent Attention Model (TDAM)[1] in which a global topic embedding (i.e., a matrix with K topic vectors) is shared

[1] `https://github.com/gabrer/topic_dependent_attention_model`

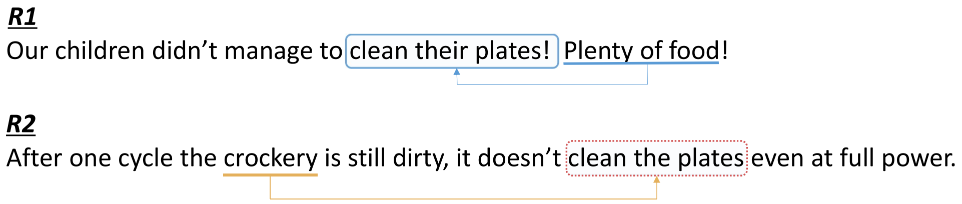

R1
Our children didn't manage to clean their plates! Plenty of food!
R2
After one cycle the crockery is still dirty, it doesn't clean the plates even at full power.

Figure 3.2: An example of topics bearing polarities.

across all the documents in a corpus and captures the global semantics in multiple topical dimensions. When processing each word in an input sequence, we can calculate the similarity of the hidden state of the word with each topic vector to get the attention weight along a certain topical dimension. By doing so, we can subsequently derive the local topical embedding for the word by the weighted combination of the global topic embeddings, indicating the varying strength of the association of the word with different topical dimensions. We use Bidirectional Gated Recurrent Unit (BiGRU) to model the input word sequence; we modify the GRU cells to derive a hidden state for the current word which simultaneously takes into account the current input word, the previous hidden state and local topic embedding.

Our proposed formulation of topical attention is somewhat related to the consciousness prior proposed in Bengio (2017) in which the conscious state value corresponds to the content of a thought and can be derived by a form of attention selecting a "small subset of all the information available" from the hidden states of the model. Analogously, we first assume the corpus is characterised by a global topic embedding. Then, we learn how to infer the local topic mixture for each analysed word/sentence combining hidden states and global topic embedding with attention.

We describe TDAM and present its application to sentiment classification in reviews by a hierarchical and multi-task learning architecture. The aim is to evaluate a review's polarity by predicting both the rating and the domain category of the review (e.g. *restaurant*, *service*, *health*, etc.). Often these reviews contain statements that can be fully specified only by the contextual topic. To illustrate, in Figure 3.2 we show two review extracts, one for a restaurant and another for a dishwasher. Interestingly, the same expression "*not to clean the plates*" can be regarded as positive for food while it bears a negative polarity for kitchen equipment. Thus, it is important to jointly consider both topic and sentiment shared over words for better sentiment analysis.

In particular, we make the following contributions:

- We design a neural architecture and a novel neural unit to analyse users' reviews while jointly taking into account topics and sentiments. The hierarchical architecture makes use of a global topic embedding which encodes the shared topics among words and sentences; while the neural unit employs a new internal attention mechanism that leverages the global topic embeddings to derive a local topic representation for words and sentences.
- We assess the benefit of multi-task learning to induce representations that are based on documents' polarities and domains. Our experiments show that combining the proposed architecture with the modified GRU unit is an effective approach to exploit the polarity and domain supervision for accurate sentiment classification and topic extraction.
- As a side task to evaluate the sentence representations encoded by TDAM, we extract *aspect-sentiment* clusters using no aspect-level annotations during the training; then, we evaluate the coherence of those clusters. Experiments demonstrate that TDAM achieves state-of-the-art performance in extracting clusters whose sentences share coherent polarities and belong to common domains.

To evaluate the performance of our model, we conduct experiments on both Yelp and Amazon review datasets (see §3.4.1). We compare the sentiment classification performance with state-of-the-art models (§3.5). Then, visualisation of topical attention weights highlights the advantages of the proposed framework (§3.5.2). We also evaluate how meaningful are the inferred representations in term of topic coherence (§3.5.3) and based on their capability to cluster sentences conveying a shared sentiment about a common aspect (§3.5.4).

3.2 Related Work

Our work is related to three lines of research.

3.2.1 Hierarchical structure for text classification

Many works have recently proposed to incorporate prior knowledge about the document structure directly into the model architecture to enhance the model's discriminative power in sentiment analysis. A hierarchical model incorporating user and product information was first proposed by Tang et al. (2015) for rating prediction of reviews. Similarly, Chen et al. (2016) combined user and product information in a hierarchical model using attention [5]; here, attention is employed to generate hidden representations for both products and users. Yang et al. (2016) used a simple and effective two-level hierarchical architecture to generate document representations for text classification; words are combined in sentences and in turn, sentences into documents by two levels of attention. Liu et al.(2018) further empowered the structural bias of neural architectures by embedding a differentiable parsing algorithm. This induces dependency tree structures used as additional discourse information; an attention mechanism incorporates these structural biases into the final document representation. Yang et al. (2019) introduced Coattention-LSTM for aspect-based sentiment analysis which designs a co-attention encoder alternating and combining the context and target attention vectors of reviews.

3.2.2 Combining topics with sequence modelling

There has been research incorporating topical information into the sequence modelling of text or use variational neural inference for supervised topic learning. Dieng et al. (2017) developed a language model combining the generative story of Latent Dirichlet Allocation (LDA) [16] with the word representations generated by a recurrent neural network (RNN). Stab et al. (2018) proposed incorporating topic information into some gates in Contextual-LSTM, improving generalisation accuracy on argument mining. Abdi et al. (2019) proposed to directly incorporate word and sentence level features about contextual polarity, type of sentence and sentiment shifts by encoding prior knowledge about part-of-speech (POS) tagging and sentiment lexicons. Kastrati et al. (2019) enhanced document representations with knowledge from an external ontology and encoded documents by topic modelling approaches. Jin et al. (2018) proposed to perform topic matrix factorisation by integrating both LSTM and LDA, where LSTM can improve the quality of the matrix factorisation by taking into account the local context of words. Card et al. (2018) proposed a general neural topic modelling framework that allows incorporating metadata information with a flexible variational inference algorithm. The metadata information can consist of labels driving the topic inference and used for the classification task, analogous to what proposed in a Bayesian framework by Blei et al. (2008) with supervised Latent Dirichlet Allocation (S-LDA).

3.2.3 Multi-task learning

Several variants of multi-task learning with neural networks have been recently used for sentiment analysis. Wu et al. (2016) proposed a multi-task learning framework for microblog sentiment classification which combines common sentiment knowledge with user-specific preferences. Liu et al. (2016) employed an external memory to allow different tasks to share information. Liu et al. (2017) proposed an adversarial approach to induce orthogonal features for each task. Chen et al. (2018) applied a different training scheme to the adversarial approach to minimise the distance between feature distributions across different domains. Zhang et al. (2018) proposed to use an embedded representation of labels to ease the generation of cross-domain features. Zheng et al. (2018) proposed to share the same sentence representation for each task which in turn can select the task-specific information from the shared representation using an ad-hoc attention mechanism. Wang et al. (2018) applied multi-task learning for microblog sentiment classification by characterising users across multiple languages.

3.3 Topic-Dependent Attention Model

We illustrate the architecture of our proposed Topic-Dependent Attention Model (TDAM) in Figure 3.3, which is a hierarchical and multi-level attention framework trained with multi-task learning.

Concretely, at the word sequence level (the bottom part of Figure 3.3), we add a word-level topic attention layer that computes the local topic embedding of each word based on the global topic embedding and the current hidden state. Such word-level local topic embedding indicates how strongly each word is associated with every topic dimension, which is fed into the Bi-GRU cell in the next time step for the derivation of the hidden state representation of the next word. Bi-GRU is used to capture the topical contextual information in both the forward and backward directions. We then have a word attention layer that decides how to combine the hidden state representations of all the constituent words in order to generate the sentence representation. At the sentence level, a similar two-level attention mechanism is used to derive the document representation, which is fed into two separate softmax layers for predicting the sentiment class and the domain category. Each of the key components of TDAM is detailed below.

3.3.1 Topic-Dependent Word Encoder

Given a word sequence $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iT})$, where $x_{it} \in \mathbb{R}^d$ is a word embedding vector with d dimensions, we use Bi-GRU to encode the word sequence. The hidden state at each word position, h_{it}, is represented by the concatenation of both

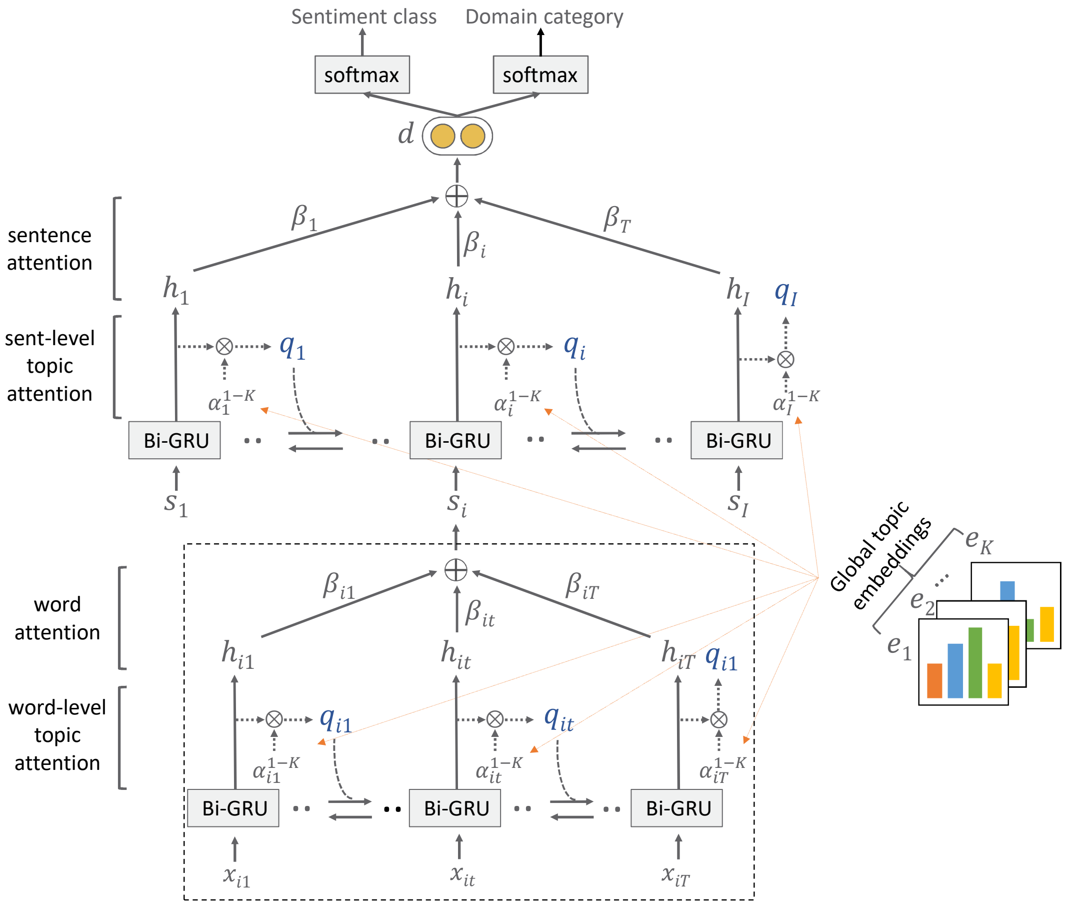


Figure 3.3: The Topic-Dependent Attention Model (TDAM) architecture.

forward and backward hidden states, $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$, which captures the contextual information of the whole sentence centred at x_{it}.

We assume there are K global topic embeddings shared across all documents, where each topic has a dense and distributed representation, $e_k \in \mathbb{R}^n$, with $k = \{1, ..., K\}$, which is initialised randomly and will be updated during model learning.

At each word position, we can calculate the word-level topic weight by measuring the distance between the word vector and each global topic vector. We first project h_{it} using a one-layer MLP and then compute the dot products between the projected h_{it} and global topic vectors $e_k, k = \{1, ..., K\}$ to generate the weight of local topic embedding for the corresponding word position[2]:

$$u_{it} = \tanh(W_w h_{it}) \tag{3.1}$$

$$\alpha_{it}^k = \text{softmax}(u_{it}^\intercal e_k) \tag{3.2}$$

[2] We drop the bias terms in all the equations for simplicity.

where $W_w \in \mathbb{R}^{n \times n}$ and $k \in \{1, ..., K\}$. The local topic embedding is then:

$$q_{it} = \sum_{k=1}^{K} \alpha_{it}^{k} \otimes e_k \tag{3.3}$$

with $q_{it} \in \mathbb{R}^n$, $\alpha_{it} \in \mathbb{R}^K$. Here, $\otimes$ denotes the multiplication of a vector by a scalar.

We add the local topic embedding into the GRU cell to rewrite the formulae as follows:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + V_r \boldsymbol{q_{t-1}}) \tag{3.4}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + V_z \boldsymbol{q_{t-1}}) \tag{3.5}$$

$$\hat{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1} + V_h \boldsymbol{q_{t-1}})) \tag{3.6}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \tag{3.7}$$

where $\sigma(\cdot)$ is the sigmoid function, all the W, U and Vs are weight matrices which are learned in the training process, $\odot$ denotes the element-wise product. The reset gate r_t controls how much past state information is to be ignored in the current state update. The update gate z_t controls how much information from the previous hidden state will be kept. The hidden state h_t is computed as the interpolation between the previous state h_{t-1} and the current candidate state $\hat{h}_t$.

In the above formulation, the hidden state in the current word position not only depends on the current input and the previous hidden state, but also takes into account the local topic embedding of the previous word. Since some of those words may be more informative than others in constituting the overall sentence meaning, we aggregate these representations with a final attention mechanism:

$$v_{it} = \tanh(W_v h_{it}) \tag{3.8}$$

$$\beta_{it} = \text{softmax}(v_{it}^{\intercal} v_w) \tag{3.9}$$

$$s_i = \sum_{t=1}^{t} \beta_{it} \otimes h_{it} \tag{3.10}$$

where β_{it} is the attention weight for the hidden state h_{it} and $s_i \in \mathbb{R}^n$ is the sentence representation for the ith sentence.

3.3.2 Sentence Encoder

Given each sentence representation s_i in document d where $i = \{1, ..., d_L\}$ and d_L denotes the document length, we can form the document representation using the proposed topical GRU in a similar way. For each sentence i, its context vector is $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$, which captures the contextual information of the whole document centred at s_i.

We follow an approach analogous to the topic-dependent word encoder and generate the local topic embedding for ith sentence:

$$u_i = \tanh(W_s h_i) \qquad W_s \in \mathbb{R}^{n \times n} \tag{3.11}$$

$$\alpha_i^k = \mathrm{softmax}(u_i^\intercal e_k) \qquad k \in \{1, ..., K\} \tag{3.12}$$

$$q_i = \sum_{k=1}^{K} \alpha_i^k \otimes e_k \qquad q_i \in \mathbb{R}^n \tag{3.13}$$

where q_i is local topic embedding for sentence i. We add the local topic embedding into the GRU cell as in Eq. 3.4-3.7.

Analogously to the word encoder, those sentences contribute differently to the overall document meaning; thus, we aggregate these representations with an attention mechanism similar to the final attention mechanism described in Section 3.3.1.

3.3.3 Multi-Task Learning

Finally, for each document d, we feed its representation m_d into the task-specific softmax layers, each one defined as follows:

$$p_d = \mathrm{softmax}(W_d m_d) \quad W_d \in \mathbb{R}^{C \times n} \tag{3.14}$$

where C denotes the total number of classes. The training loss is defined as the total cross-entropy of all documents computed for each task:

$$L_{task} = -\sum_{d=1}^{D} \sum_{c=1}^{C} y_{d,c} \log p_{d,c} \tag{3.15}$$

where $y_{d,c}$ is the binary indicator (0 or 1) if class label c is the correct classification for document d. We compute the overall loss as a weighted sum over the task-specific losses:

$$L_{total} = \sum_{j=1}^{J} \omega_j L(\hat{y}^{(j)}, y^{(j)}) \tag{3.16}$$

where J is the number of tasks, ω_j is the weight for each task, $y^{(j)}$ are the ground-truth labels in task j and $\hat{y}^{(j)}$ are the predicted labels in task j.

Dataset	**Yelp18**	**Amazon**
Sentiment classes	3	3
Domain categories	5	5
Documents	75,000	75,000
Average #s	9.7	6.7
Average #w	15.9	16.7
Vocabulary	$\sim 85 \times 10^3$	$\sim 100 \times 10^3$
Tokens	$\sim 11.7 \times 10^6$	$\sim 8.5 \times 10^3$

Table 3.1: Dataset statistics with #s number of sentences per document and and #w of words per sentence.

3.3.4 Topic Extraction

Once our model is trained, we can feed the test set and collect the local topic embedding q_{it} associated to each word (Eq. 3.3), collecting a set of n-dimensional vectors for each occurrence of words in text. This mechanism can be interpreted analogously to models generating deep contextualised word representations based on language model, where each word occurrence has a unique representation based on the context in which it appears [36, 138].

The local representation q_{it} in our model results from the interaction with the global topic embeddings, which encode the word co-occurrence patterns characterising the corpus. We posit that these vectors can give us an insight about the topic and polarity relations among words. Therefore, we first project these representations into a two-dimensional space by applying the t-SNE [167]; then, the resulting word vectors are clustered by applying the *K-means* algorithm. We create a fixed number of clusters k, whose value is tuned by maximising the topic coherence for $k \in [50, 100, 200]$. We use the distance of each word to the centroid of a topic cluster to rank words within a cluster. Similarly, we cluster sentences based on the representation resulting from the sentence-level topical attention layer. This encoding synthesises both the main topic and polarity characterising the sentence.

3.4 Experimental Setup

3.4.1 Datasets

We gathered two balanced datasets of reviews from the publicly available Yelp Dataset Challenge dataset in 2018 and the Amazon Review Dataset[3] [115], preserving the meta-information needed for a multi-task learning scenario. Each review is accompanied with one of the three ratings, *positive*, *negative* or *neutral* and comes from five of the most frequent domains[4]. Those ratings are the human-labelled review scores regarded as gold standard sentiment labels during the experimentation. For each pair of domains and ratings, we randomly sample 3,000 reviews, collecting a total of 75,000 reviews. To make it possible for others to replicate our results, we make both the dataset and our source code publicly available[5]. Table 3.1 summarises the statistics of the datasets.

3.4.2 Baselines

We train our proposed TDAM with multi-task learning to perform sentiment and domain classification simultaneously. We compare the performance of TDAM with the following baselines on both sentiment classification and topic extraction:

- **BiLSTM** [68] or **BiGRU** [29]: Both models consider a whole document as a single text sequence. The average of the hidden states is used as features for classification.

- **Hierarchical Attention Network (HAN)** [194]: The hierarchical structure of this attention model learns word and sentence representations through two additive attention levels.

- **Supervised-LDA (S-LDA)** [116]: It builds on top of the latent Dirichlet allocation (LDA) [16] adding a response variable associated with each document (e.g. review's rating or category).

- **SCHOLAR** [22]: A neural framework for topic models with metadata incorporation without the need of deriving model-specific inference. When metadata are labels, the model infers topics that are relevant to those labels.

[3] http://jmcauley.ucsd.edu/data/amazon/

[4] For Yelp: *restaurants, shopping, home services, health & medical* and *automotive*. For Amazon: *Pet supplies, electronics, health personal care, clothes shoes* and *home and kitchen*.

[5] `https://github.com/gabrer/topic_dependent_attention_model`

The baselines, such as BiLSTM, BiGRU and HAN, are additionally trained with multi-task learning, similar to the setup of our model.

3.4.3 Parameter Settings

For our experiments, we split the dataset into training, development and test set in the proportion of 80/10/10 and average all the results over 5-fold cross-validation. We perform tokenisation and sentence splitting with SpaCy[6]. We do not filter any words from the dataset during the training phase; although we use the default pre-processing for models like S-LDA and SCHOLAR. Word embeddings are initialised with 200-dimensional GloVe vectors [133]. We tune the models' hyperparameters on the development set via a grid search over combinations of learning rate $\lambda \in [0.01, 0.1]$, dropout $\delta \in [0, 0.6]$ and topic vector's size $\gamma_t \in [50, 200]$. Matrices are randomly initialised to be semi-orthogonal matrix [152]; all the remaining parameters are randomly sampled from a uniform distribution in $[-0.1, 0.1]$. We adopt Adam optimiser [83] and use a batch size of 64, sorting documents by length (i.e. number of sentences) to accelerate training convergence; we also apply batch normalisation as additional regulariser [32].

Once the model is trained, we extract the local topic embedding for each word occurrence in text as its contextualised word representation. These vectors are then projected to a lower-dimensional space by means of a multi-core implementation of a Tree-Based algorithm for accelerating t-SNE[7] [166]. Then, we cluster these words with K-means[8].

3.5 Evaluation and results

We report and discuss the experimental results obtained on three evaluation tasks, sentiment classification topic extraction and sentence cluster extraction.

3.5.1 Sentiment Classification

We train the models under two different settings: a single and a multi-task learning scenario, where we optimise over the only review polarity or over the combination of polarity and domain, respectively. For the latter, we denote the results with '-Mtl' in Table 3.2.

[6] https://spacy.io/

[7] https://github.com/DmitryUlyanov/Multicore-TSNE

[8] http://scikit-learn.org/stable/modules/clustering.html

Methods	**Yelp 18**	**Amazon**
BiLSTM	74.5 ± 0.2	72.1 ± 0.2
BiLSTM - Mtl	74.2 ± 0.2	71.8 ± 0.1
BiGRU	75.5 ± 0.1	72.5 ± 0.3
BiGRU - Mtl	75.4 ± 0.2	72.1 ± 0.3
HAN	83.7 ± 0.2	78.4 ± 0.2
HAN - Mtl	83.6 ± 0.3	78.2 ± 0.3
S-LDA	70.8 ± 0.2	64.6 ± 0.1
SCHOLAR	77.3 ± 0.2	71.4 ± 0.2
TDAM	84.2 ± 0.2	78.9 ± 0.2
TDAM - Mtl	**84.5** ± 0.3	**79.1** ± 0.2

Table 3.2: Sentiment classification accuracy and standard deviation over the 5-fold cross validation.

We can observe from the table that BiLSTM and BiGRU perform similarly. With hierarchical attention mechanisms at both the word level and the sentence level, HAN boosts the performance by nearly 10% on Yelp and 6% on Amazon compared to BiLSTM and BiGRU. For the neural topic modelling approaches, SCHOLAR outperforms traditional S-LDA by a large margin. However, SCHOLAR is still inferior to HAN. With our proposed topical attentions incorporated into the hierarchical network structure, TDAM further improves on HAN. When considering the sentiment classification task, the multi-task learning setting does not seem to bring any benefit to the baseline models, though it slightly improves the performance of TDAM. This confirms that providing more information is not necessarily beneficial for the models, and leveraging the topical information available for sentiment analysis requires tailored architectures, such as TDAM (via the modified GRU unit) or the contextualised language models [36]. These architectures are intrinsically designed to process the topical information, as confirmed by the improvement in sentiment classification and topic coherence (§3.5.3).

3.5.2 Effectiveness of Topical Attention

If we remove the topical attention and substitute our modified GRU with standard GRU, then the resulting architecture is similar to HAN [194] for a multi-task learning setting. In this section, we visualise the attention weights learned by HAN and TDAM to compare their results. Examples are shown in Figure 3.1. In TDAM, topical words such as *dentist* or the dentist's name, *Rebecca*, are regarded as relevant by the

	Yelp18			Amazon		
Topics =	50	100	200	50	100	200
HAN	-7.22	-7.05	-7.08	-13.21	-13.15	-13.14
HAN - Mtl	-7.04	-6.94	-6.93	-12.72	-12.20	-12.29
S-LDA	-6.26	-6.13	-6.15	-9.57	-9.41	-9.28
SCHOLAR	-6.24	-6.08	-6.11	-9.52	-9.46	-9.48
SCHOLAR-R	**-6.19**	-6.11	-6.08	-9.34	**-9.09**	-9.17
TDAM	-6.41	-6.12	-6.09	-9.62	-9.50	-9.46
TDAM - Mtl	-6.22	**-6.05**	**-5.93**	**-9.23**	-9.12	**-9.01**

Table 3.3: Topic coherence for different number of topics. The higher the better.

model. Along with them, it focuses on words bearing a strong sentiment, such as *nicest* or *happy*. These weights are compared with the attention weights learned by the HAN, showing that it primarily focuses sentiment words and overlooks other topical words, such as *dentist*.

3.5.3 Topic Coherence Evaluation

Among the baselines, S-LDA and SCHOLAR are topic modelling methods and therefore they can directly output topics from text. In addition, we can follow the topic extraction procedure described in Section 3.3.4 to extract topics from HAN to gain an insight into the learned representations. We thus compare the topic extraction results of TDAM with these three models. Also, as previously shown in [22], higher regularisation on SCHOLAR produced better topics. Therefore, we also report the results using SCHOLAR with higher regularisation, named as SCHOLAR-R.

To evaluate the quality of topics, we use the topic coherence measure[9] proposed in [148] which has been shown to outperform all the other existing topic coherence measures in matching the human judgement. We can observe from Table 3.3 that HAN gives the worse topic coherence results, showing that simply extracting topics using the attention weights is not feasible. With the incorporation of domain category information through multi-task learning, HAN-Mtl gives slightly better coherence results. Among topic modelling approaches, SCHOLAR-R with higher regularisation generates more coherence topics compared to SCHOLAR, which outperforms S-LDA. TDAM gives similar topic coherence results as SCHOLAR-R on some topic numbers. TDAM-Mtl improves over TDAM and generates the best coherence results on 2 out of 3 topic settings for both Yelp18 and Amazon, showing higher coherence scores overall.

[9] https://github.com/dice-group/Palmetto

3.5.4 Aspect-Polarity Coherence Evaluation

To assess the effectiveness of our proposed TDAM in extracting polarity-bearing topics, we use the annotated dataset provided in the SemEval 2016 Task 5 for aspect-based sentiment analysis[10]; this provides sentence-level annotations about different aspects (e.g. `FOOD#QUALITY`) and polarities (`pos`, `neut`, `neg`) in restaurant and laptop reviews.

We join the training set of restaurant and laptop reviews with the Yelp18 and Amazon dataset, respectively. With the same approach adopted for topic extraction, we use the test sets to generate sentence clusters and evaluate their *aspect-polarity coherence*, defined as the ratio of sentences sharing a common aspect and sentiment in a cluster. For the two topic modelling approaches, S-LDA and SCHOLAR, we generate sentence clusters based on the generative probabilities of sentences conditional on topics. Note that although the SemEval dataset provides the sentence-level annotations of aspects and polarities, these were NOT used for the training of the models here. We only use the gold standard annotations of aspects and polarities in the test set to evaluate the quality of the extracted polarity-bearing topics.

We generate multiple clusters, i.e. (50,100,150), representing polarity-bearing aspects and report the results in Table 3.4, which shows the ratio of sentence clusters with more than *threshold* sentences sharing a common aspect (values in brackets) or a common aspect-polarity. We can observe that the topic modelling approaches struggle in generating coherent aspect-polarity clusters with at least 50% of common aspect-polarities. The two hierarchical models, HAN and TDAM, have significantly more coherent aspect-polarity clusters compared to S-LDA and SCHOLAR, and both benefit from multi-task learning. For all the models, results on SemEval-Restaurant are better than those obtained on SemEval-Laptop. This might be partly attributed to the abundant restaurant reviews on Yelp18 compared to the laptop-related reviews on Amazon. Overall, TDAM-Mtl gives the best results.

We also show some example sentence clusters produced by HAN and TDAM under multi-task learning in Table 3.5. HAN discriminates rather effectively positive sentences (the majority in the cluster) from negative and neutral ones. However, despite several sentences sharing the same polarity, their topics/aspects are quite heterogeneous. TDAM phrases are rather coherent overall, both in terms of topics and expressed sentiment. Along with their aspects and polarity, it is worth noting that the average length of the top-10 sentences within the clusters is generally longer for HAN (9.1 and 14.4 words for the positive and negative clusters, respectively)

[10] http://alt.qcri.org/semeval2016/task5/

and slightly shorter for TDAM (7.8 and 13.5 words for the positive and negative clusters, respectively). TDAM tends to emphasise the encoding of more concise sentences and with a clearer dominant aspect (e.g. `FOOD#QUALITY`); while HAN, although clustering rather coherent sentences in terms of their shared sentiments, puts less emphasis on the analysed topic, resulting in a miscellany of aspects.

These results are encouraging. Our TDAM is able to detect coherent aspects and also polarity-bearing aspects despite using no aspect-level annotations at all. Considering it is very time consuming to provide aspect-level annotations, TDAM could be used to bootstrap the training of aspect-based sentiment detectors.

3.6 Summary

We have presented a new topic-dependent attention model for sentiment classification and topic extraction. The conjunction of the topical recurrent unit and the multi-task learning framework has been shown to be an effective combination to generate representations for more accurate sentiment classification, meaningful topics and for side tasks of polarity-bearing aspects detection.

Methods	Topics	SemEval-Restaurant					SemEval-Laptop				
	threshold	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$
HAN	50	(0.52) 0.40	(0.28) 0.24	(0.14) 0.10	(0.04) 0.02	(0.00) 0.00	(0.18) 0.15	(0.15) 0.12	(0.08) 0.03	(0.03) 0.01	(0.00) 0.00
	100	(0.64) 0.47	(0.36) 0.27	(0.14) 0.11	(0.09) 0.07	(0.03) 0.03	(0.28) 0.26	(0.21) 0.20	(0.12) 0.08	(0.05) 0.04	(0.01) 0.01
	150	(0.70) 0.59	(0.39) 0.32	(0.21) 0.16	(0.14) 0.12	(0.12) 0.08	(0.37) 0.34	(0.28) 0.23	(0.14) 0.10	(0.8) 0.07	(0.04) 0.03
HAN-Mtl	50	(0.56) 0.40	(0.36) 0.28	(0.26) 0.18	(0.12) 0.10	(0.06) 0.02	(0.24) 0.19	(0.18) 0.15	(0.12) 0.04	(0.04) 0.02	(0.03) 0.01
	100	(0.64) 0.52	(0.51) 0.40	(0.26) 0.22	(0.17) 0.13	(0.10) 0.06	(0.31) 0.27	(0.26) 0.19	(0.16) 0.09	(0.08) 0.04	(0.04) 0.03
	150	(0.72) 0.63	(0.51) 0.43	(0.30) 0.22	(**0.21**) 0.12	(0.14) 0.08	(0.41) 0.38	(0.35) 0.24	(0.17) 0.12	(0.12) 0.08	(0.05) 0.03
S-LDA	50	(0.18) 0.12	(0.08) 0.03	(0.02) 0.01	(0.00) 0.00	(0.00) 0.00	(0.09) 0.07	(0.08) 0.05	(0.03) 0.01	(0.02) 0.00	(0.00) 0.00
	100	(0.24) 0.21	(0.11) 0.10	(0.03) 0.02	(0.00) 0.00	(0.00) 0.00	(0.15) 0.14	(0.10) 0.06	(0.04) 0.01	(0.01) 0.00	(0.00) 0.00
	150	(0.39) 0.35	(0.19) 0.16	(0.05) 0.04	(0.01) 0.01	(0.01) 0.01	(0.27) 0.24	(0.14) 0.11	(0.08) 0.04	(0.2) 0.02	(0.00) 0.00
SCHOLAR-R	50	(0.31) 0.18	(0.22) 0.10	(0.04) 0.03	(0.01) 0.01	(0.00) 0.00	(0.14) 0.10	(0.08) 0.04	(0.06) 0.02	(0.03) 0.01	(0.00) 0.00
	100	(0.39) 0.24	(0.24) 0.13	(0.08) 0.04	(0.03) 0.01	(0.01) 0.01	(0.21) 0.15	(0.16) 0.08	(0.09) 0.05	(0.03) 0.01	(0.01) 0.01
	150	(0.43) 0.36	(0.28) 0.19	(0.11) 0.10	(0.04) 0.03	(0.01) 0.01	(0.34) 0.26	(0.21) 0.13	(0.12) 0.06	(0.05) 0.02	(0.02) 0.01
TDAM	50	(0.54) 0.42	(0.30) 0.24	(0.12) 0.08	(0.06) 0.04	(0.02) 0.00	(0.19) 0.17	(0.15) 0.14	(0.09) 0.06	(0.03) 0.02	(0.00) 0.00
	100	(0.63) 0.55	(0.40) 0.31	(0.21) 0.16	(0.14) 0.10	(0.10) 0.06	(0.38) 0.29	(0.24) 0.18	(0.12) 0.10	(0.05) 0.04	(0.02) 0.02
	150	(0.73) 0.65	(0.43) 0.36	(0.28) **0.26**	(0.19) 0.15	(**0.16**) 0.13	(0.39) 0.37	(0.28) 0.25	(0.16) 0.13	(0.08) 0.08	(0.05) 0.03
TDAM-Mtl	50	(0.68) 0.51	(**0.52**) 0.38	(0.24) 0.20	(0.14) 0.12	(0.06) 0.04	(0.31) 0.25	(0.26) 0.17	(0.22) 0.13	(0.13) 0.07	(0.04) 0.02
	100	(0.72) 0.58	(0.47) 0.39	(0.31) 0.24	(0.19)0.16	(0.13) 0.12	(0.39) 0.32	(0.38) 0.24	(0.26) 0.15	(0.12) 0.09	(0.02) 0.0
	150	(**0.80**) **0.68**	(0.50) **0.40**	(**0.32**) 0.25	(0.20) **0.16**	(**0.16**) **0.14**	(**0.48**) **0.43**	(**0.42**) **0.31**	(**0.26**) **0.18**	(**0.17**) **0.11**	(**0.09**) **0.05**

Table 3.4: Ratio of clusters where at least x% sentences sharing the same aspect (values in brackets) and sharing the same aspect-polarity (i.e. both aspect and polarity are correct).

HAN		TDAM	
Positive polarity - Food#Quality			
1) wait the half hour with a cup of joe , and enjoy more than your average breakfast .	`FOOD#QUALITY pos`	1) the food was all good but it was way too	`FOOD#QUALITY neg`
2) space was limited , but the food made up for it .	`RESTAURANT#MISCELLANEOUS neg`	2) the pizza 's are light and scrumptious .	`FOOD#STYLE_OPTIONS pos`
3)the prices should have been lower .	`FOOD#STYLE_OPTIONS neg`	3) the food is great and they make a mean bloody mary .	`FOOD#QUALITY pos`
4) the crowd is mixed yuppies , young and old .	`RESTAURANT#MISCELLANEOUS neut`	4) great draft and bottle selection and the pizza rocks .	`FOOD#QUALITY pos`
5)making the cakes myself since i was about seven - but something about these little devils gets better every time .	`FOOD#QUALITY pos`	5) the food is simply unforgettable !	`FOOD#QUALITY pos`
6) mmm ... good !	`RESTAURANT#GENERAL pos`	6)the food is great, the bartenders go that extra mile.	`FOOD#QUALITY pos`
7) the service is so efficient you can be in and out of there quickly .	`SERVICE#GENERAL pos`	7)the food is sinful.	`FOOD#QUALITY pos`
8) service was decent .	`SERVICE#GENERAL neut`	8)the sushi here is delicious !	`FOOD#QUALITY pos`
9) their specialty rolls are impressive	`FOOD#QUALITY pos`	9) the food was great !	`FOOD#QUALITY pos`
10) it was nt the freshest seafood ever , but the taste and presentation was ok .	`FOOD#STYLE_OPTIONS neut`	10) good eats .	`FOOD#QUALITY pos`
Negative polarity - Food#Quality			
1) the pancakes were certainly inventive but $ 8.50 for 3 - 6 ″ pancakes (one of them was more like 5 ″)	`FOOD#STYLE_OPTIONS neg`	1) i may not be a sushi guru	`FOOD#QUALITY neg`
2)a beautiful assortment of enormous white gulf prawns , smoked albacore tuna, [..] and a tiny pile of dungeness	`FOOD#STYLE_OPTIONS pos`	2) rice is too dry , tuna was n't so fresh either .	`FOOD#QUALITY neg`
3) space was limited , but the food made up for it .	`RESTAURANT#MISCELLANEOUS neg`	3)the only way this place survives with such average food is because most customers are one-time customer tourists	`FOOD#QUALITY neg`
4) the portions are big though , so do not order too much .	`FOOD#STYLE_OPTIONS neut`	4) the portions are big though , so do not order too much .	`FOOD#STYLE_OPTIONS neut`
5) not the biggest portions but adequate .	`FOOD#QUALITY pos`	5) the only drawback is that this place is really expensive and the portions are on the small side .	`RESTAURANT#PRICES neg`
6)the waiter was a bit unfriendly and the feel of the restaurant was crowded .	`SERVICE#GENERAL neg`	6)but i can tell you that the food here is just okay and that there is not much else to it .	`FOOD#QUALITY neg`
7) food was fine , with a some little - tastier - than - normal salsa .	`FOOD#QUALITY pos`	7) and they give good quantity for the price .	`FOOD#STYLE_OPTIONS pos`
8) i got the shellfish and shrimp appetizer and it was alright .	`FOOD#QUALITY neut`	8) food was fine , with a some little - tastier - than - normal salsa .	`FOOD#QUALITY pos`
9) once seated it took about 30 minutes to finally get the meal .	`FOOD#GENERAL neg`	9)our drinks kept coming but our server came by a couple times .	`SERVICE#GENERAL pos`
10) the food is here is incredible , though the quality is inconsistent during lunch .	`FOOD#QUALITY pos`	10) nice food but no spice !	`FOOD#QUALITY neg`

Table 3.5: Clusters of positive and negative aspects about `FOOD#QUALITY` experiences from SemEval16. On the left, sentences were clustered with HAN; on the right, they were clustered with TDAM. The aspect and polarity labels for each sentence are the gold standard annotations.

Chapter 4

A Disentangled Adversarial Neural Topic Model

Chapter Abstract

In this chapter, we present a novel disentangled adversarial neural topic model (DIATOM), combining the NTM architecture with an adversarial training to disentangle opinion topics from plot and neutral ones. Existing topic models when applied to user reviews may extract topics associated with writers' subjective opinions mixed with those related to factual descriptions, such as plot summaries in movie and book reviews. It is thus desirable to automatically separate opinion topics from plot/neutral ones for more discriminative features and better interpretability. Although existing approaches have achieved significant results in topic extraction, surprisingly, very little work has been done on how to disentangle these latent topics. In the following, we first describe DIATOM, outlining the relevant literature. Then, we report an extensive experimental assessment based on a new collection of movie and book reviews paired with their plots, namely the MOBO *dataset, showing an improved coherence and variety of topics, a consistent disentanglement rate, and sentiment classification performance superior to other supervised topic models.*

4.1 Introduction

Variational Autoencoders (VAEs) [84] allow to design complex generative models of data since the inference process of VAE-based approaches has the advantage of being independent from the model architecture providing high flexibility in designing new neural components. In the wake of the renewed interest for VAEs, traditional probabilistic topic models [16] have been revised giving rise to several Neural Topic Model (NTM) variants, such as NVDM [118], ProdLDA [155], NTM-R [40], etc. However, existing topic models when applied to user reviews may extract topics associated with writers' subjective opinions mixed with those related to factual descriptions such as plot summaries of movies and books [97]. Although these approaches have achieved significant results via the neural inference process, surprisingly, very little work has been done on how to disentangle the inferred topic representations.

Disentangled representations can be defined as representations where individual latent units are sensitive to variations of a single generative factor, while being relatively invariant to changes of other factors [10, 64]. Inducing such representations has been shown to be significantly beneficial for their generalisation and interpretability [2, 131]. For example, an image can be view as the result of several generative factors mutually interacting, as the one or many sources of light, the material and reflective properties of various surfaces or the shape of the objects depicted [10]. In the context of topic modelling, documents result from a generative process over mixtures of latent topics, and therefore, we propose to consider these latent topics as generative factors to be disentangled to improve their interpretability and discriminative power. Disentangled topics are topics invariant to the factors of variation of text, which for instance, in the context of book and movie reviews could be the author's opinion (e.g. positive/negative), the salient parts of a plot or other auxiliary information reported. An illustration of this is shown in Figure 4.1 in which opinion topics are separated from plot topics.

However, models relying solely on sentiment information are easily misled and not suitable to disentangle opinion from plots, since even plot descriptions frequently make large use of sentiment expressions [129]. Consider, for example, the following sentence: "The ring holds a *dark* power, and it soon begins to exert its *evil influence* on Bilbo", an excerpt from a strong positive Amazon's review.

Therefore, we propose to distinguish opinion-bearing topics from plot/neutral ones combining a neural topic model architecture with an adversarial training. In this study, we present the DIsentangled Adversarial TOpic Model (DIATOM)[1], aiming at disentangling information related to the target labels (i.e. the review score), from other distinct aspects yet possibly still polarised (e.g. plot descriptions). We also introduce a new dataset, namely the MOBO dataset[1], made up of movie

[1] Source code and dataset omitted for the anonymous submission.

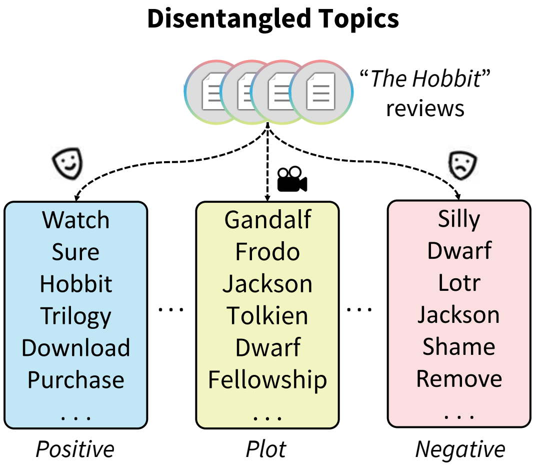


Figure 4.1: Disentangled topics extracted by DIATOM from the Amazon reviews for "The Hobbit".

and book reviews, paired with their related plots. The reviews come from different publicly available datasets: IMDB [108], GoodReads [174] and Amazon reviews [115], and encompass a wide spectrum of domains and styles. We conduct an extensive experimental assessment of our model. First, we assess the topic quality in terms of topic coherence and diversity and compare DIATOM with other supervised topic models on the sentiment classification task; then, we analyse the disentangling rate of topics to quantitatively assess the degree of separation between actual opinion and plot/neutral topics.
Our contributions are summarised below:

- We propose a new model, DIATOM, which is able to generate disentangled topics through the combination of VAE and adversarial learning.
- We introduce the MOBO dataset, a new collection of movie and book reviews paired with their plots.
- We conduct an experimental assessment of our model, highlighting more interpretable topics with better topic coherence and diversity scores compared to other state-of-the-art supervised topic models, and improved discriminative power on sentiment classification, and a consistent topic-disentanglement rate.

The rest of the chapter is organised as follows. We reviews the related literature on sentiment-topic models, neural topic models and the studies on disentangled representations (§4.2). Then, we present the details of our proposed DIATOM model (§4.3), followed by the experimental setup (§4.4) and results (§4.5). Finally, we conclude with a summary of the results (§4.6).

4.2 Related Work

Our work is closely related to three lines of research: sentiment-topic models, neural topic models and learning disentangled representations.

4.2.1 Sentiment-Topic Models

Probabilistic graphical models for topic extraction have been extensively studied. Beyond LDA [16]; a wide spectrum of models has extended LDA to more specific tasks using contextual information [15, 149, 181]. Supervised-LDA (sLDA) [116] is a general-purpose supervised extension that builds on top of LDA by adding a response variable associated with each document (e.g. a review's rating).

A category of extensions particularly relevant for this work is the sentiment-topic models. Examples include the Joint Sentiment-Topic (JST) model [97, 98] and Aspect and Sentiment Unification Model (ASUM) [74]. These models are able to extract informative topics grouped under different sentiment classes. Although they do not rely on document labels, they require word prior polarity information to be incorporated into the learning process in order to generate consistent results. Nevertheless, when provided with document-level class labels, JST can learn document-topic distributions influenced by the class information. The possibility to supervise the learning process with document labels and to avoid the necessity of prior information over words makes it suitable for a fair comparison with the model proposed in this work. Besides, these models require carefully tailored inference algorithms, and the standard Gibbs sampling algorithm used can have a high computational cost when fitting large-scale data, with time and memory scaling linearly with the number of documents, leading researchers to devise more sophisticated approaches to make it scalable [54].

Compared to DIATOM, the discussed sentiment topic models can only distinguish between *polarity*-bearing topics and neutral ones, remaining strictly aligned to the provided labels. Instead, along with neutral topics, DIATOM is able to generate opinion-bearing topics and plot topics that may still be polarised but not carrying any user's opinion.

4.2.2 Neural Topic Models

Neural models provide a more generic and extendable alternative to topic modelling, and therefore, have recently gained increasing interest. Some of them use belief networks [124], or enforce the Dirichlet prior on the document-topic distribution by means of Wasserstein Autoencoders [125]. Others adopt continuous representations to capture long-term dependencies or preserve word order via sequence-to-sequence VAE [20, 38, 188, 195] whose time complexity and difficulty of training, however, have limited their applications.

Neural Variational Document Model (NVDM) [118] is a direct extension of VAE used for topic detection in text. In NVDM, the prior of latent topics is assumed to be a Gaussian distribution. This is not ideal since it cannot mimic the simplex in the latent topic space. To address this problem, LDA-VAE [155] instead used the logistic normal distribution to approximate the Dirichlet distribution. ProdLDA [155] extended LDA-VAE by replacing the mixture model of LDA with a product of experts. SCHOLAR is a neural framework for topic models with metadata incorporation [22]. When metadata are document labels, the model infers topics that are relevant to those labels. Although some studies have applied the adversarial approach [55] to topic models setting a Dirichlet prior on the generative network [113, 178], it is still unexplored how to use this mechanism to disentangle opinion-bearing topics from plot or neutral topics.

Compared to these neural topic models, DIATOM is the first attempt using an adversarial mechanism to distinguish between topic types (i.e. opinion and plot topics), while not only generating topics aligned with the available target labels (i.e. opinion topics) but seamless incorporating the external signal of plot summaries to drive the generation of topics about salient parts of plots mentioned by users (i.e. plot topics) not related to the target classes (i.e. sentiment polarity).

4.2.3 Representation Disentanglement

Despite the lack of general consensus about a unique definition of disentangled representations [48, 65], it typically refers to representations that are only sensitive to one single generative factor of data and relatively invariant to other factors of variation [10]. One proposed definition builds upon the concept of statistical independence by minimising total correlation [2, 41], while an alternative approach explored the possibility to measure and track the changes in a single latent dimension as degree of disentanglement [64]. However, the disentanglement of representation achieved in DIATOM is instead analogous to the one presented in Thomas et al. (2017) and Bengio et al. (2017), where they impose additional constraints to the representations in the latent space that can be controlled exploiting a reinforcement learning mechanism determining the disentangled factors. In DIATOM, we alternatively make use of an adversarial approach over the available target labels.

Both Generative Adversarial Networks (GAN) [27, 99, 110, 114] and VAEs [28, 70, 131] have been successfully employed in disentangling features in computer vision tasks. Application in text processing has shown promising results [42, 59, 67, 75, 86], yet applications to topic modelling are still limited [182] and to the best of our knowledge, there is no work in separating opinion-bearing topics from plot/neutral topics.

4.3 DIATOM architecture

Our proposed DIATOM model is shown in Figure 4.2. Assuming a document $\boldsymbol{x}$ is associated with a sentiment label y_s, and each document can be represented by latent topics associated with sentiments ($\boldsymbol{z}_s$) and plots[2] ($\boldsymbol{z}_a$), we aim to learn a model maximising the joint data-label log-likelihood, $\log p(\boldsymbol{x}, y_s)$:

$$\begin{aligned}
&\log p(\boldsymbol{x}, y_s) = \log \int \int p(\boldsymbol{x}, y_s, \boldsymbol{z}_a, \boldsymbol{z}_s) d\boldsymbol{z}_a d\boldsymbol{z}_s \\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{z}_a|\boldsymbol{x}), q_\psi(\boldsymbol{z}_s|\boldsymbol{x}, y_s)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_a, \boldsymbol{z}_s)] \\
&\quad + \mathbb{E}_{q_\phi(\boldsymbol{z}_a|\boldsymbol{x}), q_\psi(\boldsymbol{z}_s|\boldsymbol{x}, y_s)}[\log p_\pi(y_s|\boldsymbol{x})] \\
&\quad - \mathrm{KL}\left(q_\phi(\boldsymbol{z}_a|\boldsymbol{x}) || p(\boldsymbol{z}_a)\right) \\
&\quad - \mathrm{KL}\left(q_\psi(\boldsymbol{z}_s|\boldsymbol{x}, y_s) || p(\boldsymbol{z}_s)\right) \qquad (4.1)
\end{aligned}$$

Inspired by Miao et al. (2016) and Card et al. (2018), we assume the document-level topic distribution for plots can be approximated by a multi-layer perceptron (MLP) taking as input a multivariate Gaussian distribution, and similarly for the topic distribution for sentiments. The multinomial distribution over words under a plot topic and an opinion topic can be parametrised by a weight matrix $\boldsymbol{W}$. The generative process is shown below.

- For each document $d \in \{1, .., D\}$,
  - Draw the latent plot-topics,
    $\hat{\phi} \sim \mathcal{N}(\mu_\phi, \Sigma_\phi), \quad \boldsymbol{z}_a = f_{\hat{\phi}}(\hat{\phi})$
  - Draw the latent opinion-topics,
    $\hat{\psi} \sim \mathcal{N}(\mu_\psi, \Sigma_\psi), \ \boldsymbol{z}_s = f_{\hat{\psi}}(\hat{\psi})$
  - For each word $n \in \{1, .., N_d\}$ in document d

[2] These are the topics not associated with the target sentiments, which can be either plot topics or neutral topics (not about plots). For notational convenience, we call both plot topics.

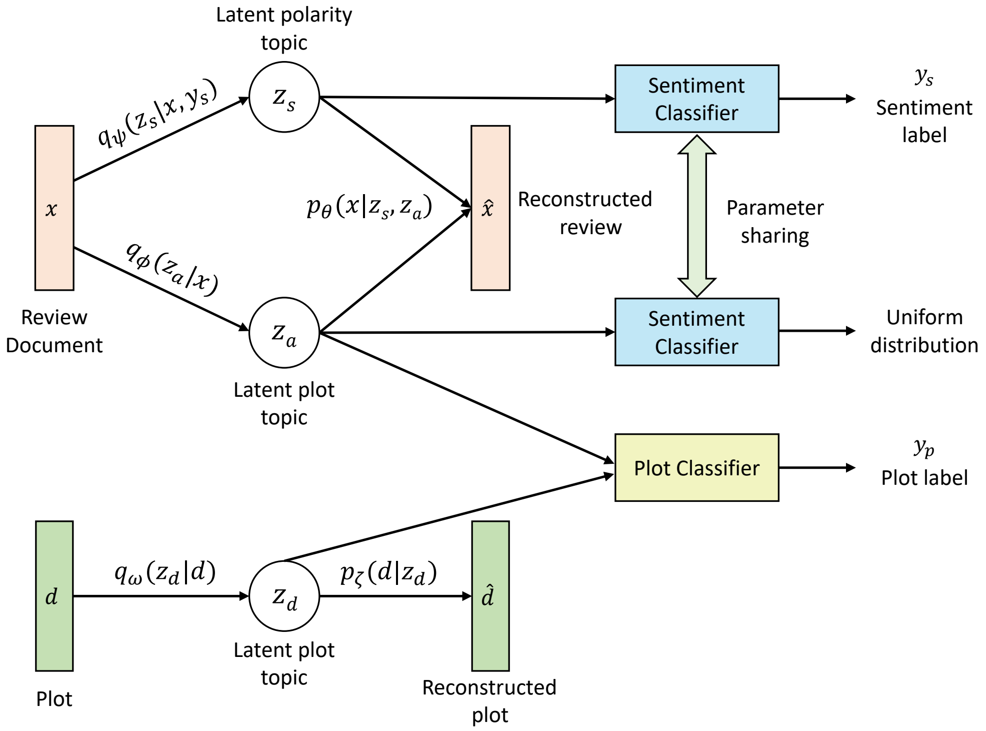


Figure 4.2: The DIATOM Architecture.

    * Draw $x_{d,n} \sim p(x_{d,n}|\boldsymbol{W}, \boldsymbol{z}_a, \boldsymbol{z}_s)$

- Generate the document-level sentiment label, $y_s \sim p(y_s|f_y(\boldsymbol{z}_s))$

where $f_{\hat{\phi}}$, $f_{\hat{\psi}}$ and f_y are MLPs, $\boldsymbol{z}_a$ is a K-dimensional latent topic representation of plots for document d, $\boldsymbol{z}_s$ is a S-dimensional latent topic representation of sentiments for document d. The probability of word $x_{d,n}$ can be parametrised by another network:

$$p(x_{d,n}|\boldsymbol{W}, \boldsymbol{z}_a, \boldsymbol{z}_s) \propto \exp\left(\boldsymbol{m}_d + \boldsymbol{W} \cdot (\boldsymbol{z}_a \parallel \boldsymbol{z}_s)\right) \tag{4.2}$$

where $\boldsymbol{m}_d$ is the V-dimensional background log-frequency word distribution, and $\boldsymbol{W} \in \mathbb{R}^{V \times (K+S)}$, while $\boldsymbol{z}_a \parallel \boldsymbol{z}_s$ is the concatenation of the two latent topic vectors.

Plot Inference Network

Following the idea of VAE which computes a variational approximation to an intractable posterior using MLPs, we define two inference networks f_{μ_ϕ} and f_{Σ_ϕ} which takes as input the word counts in documents:

$$\mu_\phi = f_{\mu_\phi}(\boldsymbol{x}) \quad \Sigma_\phi = \text{diag}(f_{\Sigma_\phi}(\boldsymbol{x})) \tag{4.3}$$

The outputs of both networks are vectors in $\mathbb{R}^K$. Here, 'diag' converts a column vector to a diagonal matrix. For a document $\boldsymbol{x}$, $q(\phi) \simeq \mathcal{LN}(\mu_\phi, \Sigma_\phi)$. With such a formulation, we can generate samples from $q(\phi)$ by first sampling $\epsilon \sim \mathcal{N}(0, I)$ and then computing $\hat{\phi} = \sigma(\mu_\phi + {\Sigma_\phi}^{1/2}\epsilon)$.

Sentiment Inference Network

Similarly, to compute a variational approximation to $q(\psi)$, we define two inference networks f_{μ_ψ} and f_{Σ_ψ} which takes as input the word counts in documents:

$$\mu_\psi = f_{\mu_\psi}(\boldsymbol{x}) \quad \Sigma_\psi = \mathrm{diag}(f_{\Sigma_\psi}(\boldsymbol{x})) \tag{4.4}$$

The outputs of both networks are vectors in $\mathbb{R}^S$. For a document $\boldsymbol{x}$, $q(\psi) \simeq \mathcal{LN}(\mu_\psi, \Sigma_\psi)$. We then generate samples from $q(\psi)$ by first sampling $\epsilon \sim \mathcal{N}(0, I)$ and then computing $\hat{\psi} = \sigma(\mu_\psi + {\Sigma_\psi}^{1/2}\epsilon)$.

Overall Objective

With the sampled $\hat{\phi}$ and $\hat{\psi}$, for each document $\boldsymbol{x}$, we can minimise the reconstruction loss with a Monte Carlo approximation using L independent samples:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x}} \approx \frac{1}{L}\sum_{l=1}^{L}\sum_{n=1}^{N_d} &\log p_\theta(x_{d,n}|\hat{\phi}^{(l)}, \hat{\psi}^{(l)}) \\ &- \mathrm{KL}\left(q(\boldsymbol{z}_a|\boldsymbol{x}) \;||\; p(\boldsymbol{z}_a)\right) \\ &- \mathrm{KL}\left(q(\boldsymbol{z}_s|\boldsymbol{x}, y_s) \;||\; p(\boldsymbol{z}_s)\right)\end{aligned} \tag{4.5}$$

where the first term in the RHS is given by Eq. (4.2). It has been previously shown in [84], if a standard multivariate normal prior is placed on the latent variables $\boldsymbol{z}_a$ and $\boldsymbol{z}_s$, then there is a closed form solution to the KL divergence terms above.

We assume that the latent topics associated with plots, $\boldsymbol{z}_a$, are independent of sentiment classes, and hence, when fed into a sentiment classifier, should generate a uniform sentiment class distribution (similar to adversarial learning). On the contrary, the latent topics associated with sentiments, $\boldsymbol{z}_s$, should bear essential information to discriminate between sentiment classes. Therefore, we define the following two objectives for sentiment classification; the former being the expected

KL divergence with the uniform distribution $\mathcal{U}$, and the latter a cross-entropy loss:

$$\mathcal{L}_{adv} = -\mathbb{E}_{q_\phi(\boldsymbol{z}_a)}\left[\mathrm{KL}\left(\mathcal{U}(0,M)\;||\;p(\hat{y}|\boldsymbol{z}_a)\right)\right] \tag{4.6}$$

$$\mathcal{L}_{sent} = -\mathbb{E}_{q_\psi(\boldsymbol{z}_s)}\sum_{c=1}^{M} y_c \log\left(p(\hat{y}_c|\boldsymbol{z}_s)\right) \tag{4.7}$$

where M is the total number of sentiment classes, and $\mathcal{U}(0,M)$ represents the uniform sentiment class distribution.

To further disentangle the latent topics associated with plots, $\boldsymbol{z}_a$, and latent topics associated with sentiment, $\boldsymbol{z}_s$, while concurrently minimise the redundancy in the final topic matrix, we apply an orthogonal regularizer over the decoder matrix $\boldsymbol{W}$. $\mathcal{L}_{orth}$ reaches its minimum value when the dot product between different topic distributions goes close to zero:

$$\mathcal{L}_{orth} = ||\,\boldsymbol{W}\cdot\boldsymbol{W}^T - \mathbb{I}\,|| \tag{4.8}$$

Our final objective function is:

$$\mathcal{L} = -\alpha\mathcal{L}_{\boldsymbol{x}} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{sent} + \eta\mathcal{L}_{orth} \tag{4.9}$$

where α, β, γ and η control the relative contribution of various loss functions.

Plot Network

An additional VAE is plugged to the model providing a supplementary signal for the latent plot topic extraction. This mechanism preserves the plot information that might contain some sentiment words and thus, be wrongly regard as a user's opinion. The inference network is defined analogously to Eq. 4.3, which instead of taking a review document, takes a plot summary as input. An additional cross-entropy objective is minimised to drive the latent plot topics ($\boldsymbol{z}_a$) which would have a similar discriminative power as the features ($\boldsymbol{z}_d$) directly derived from the plots when used for plot classification:

Statistics	**IMDB**	**GoodReads**	**Amazon**
No. of plots	1,131	150	100
No. of reviews	25,836	83,852	32,375
No. of reviews per plot (avg / max / min)	24 / 30 / 10	954 / 3,000 / 549	464 / 1525 / 272
No. of words per review (avg / max / min)	156 / 1419 / 5	63 / 3506 / 5	61 / 3226 / 5
No. of words per plot (avg / max / min)	624 / 5501 / 93	104 / 233 / 30	198 / 687 / 71
Pos / Neg / Neutral distribution	0.46 / 0.54 / 0	0.33 / 0.50 / 0.17	0.32 / 0.46 / 0.22
Training set	20,317	65,816	25,883
Development set	2,965	9,007	3,275
Test set	2,554	9,029	3,217
No. of annotated sentences	6,000	6,000	6,000

Table 4.1: The MOBO dataset statistics.

$$\mathcal{L}_{\boldsymbol{d}} = \mathbb{E}_{q_{\omega}(\boldsymbol{z}_d|d)}[p_{\zeta}(d|\boldsymbol{z}_d)] - \mathrm{KL}\left(q(\boldsymbol{z}_d|\boldsymbol{d}) \;||\; p(\boldsymbol{z}_d)\right) \tag{4.10}$$

$$\mathcal{L}_{plot_{z_a}} = -\mathbb{E}_{q_{\phi}(\boldsymbol{z}_a)} \sum_{p=1}^{P} y_p \log(p(\hat{y}_p|\boldsymbol{z}_a)) \tag{4.11}$$

$$\mathcal{L}_{plot_{z_d}} = -\mathbb{E}_{q_{\omega}(\boldsymbol{z}_d)} \sum_{p=1}^{P} y_p \log\left(p(\hat{y}_p|\boldsymbol{z}_d)\right) \tag{4.12}$$

where R denotes the total number of plots in each dataset. Finally, $-\mathcal{L}_{\boldsymbol{d}}$ and $\mathcal{L}_{plot}$ are added to the overall loss defined in Eq. 4.9.

4.4 Experimental Setup

We conduct thorough experimental evaluations to assess the quality and disentanglement rate of extracted topics. To assess the quality of topics, we compute their topic coherence [148] coupled with their topic uniqueness. Then, we additionally look at the discriminative power of the disentangled features on the sentiment classification task. To fully assess the disentanglement rate of different methods, we perform topic labelling to compute the sentiment polarity of each topic (if any) and then measure the overall disentanglement rate (Eq. 4.14). As a result, we obtain an estimate of the extent to which different models can accurately control the topic disentanglement rate. We introduce and use a new dataset, named the *MOBO* dataset, pairing movie/book plots with their users' reviews, and including human-annotated sentences.

MOBO Dataset

The MOBO dataset is a collection of reviews and plots about **MO**vie and **BO**ok, associated with human-annotated sentences: while the pairs of reviews and plots are used to enhance the generation of plot topics, the human-annotated sentences provide the necessary ground-truth to automatically evaluate the topics' polarity.

Movie and book reviews were collected and paired from 3 public datasets: the Stanford's IMDB movie reviews [108], the GoodReads [174] and the Amazon reviews dataset [115]. Among all the available reviews in the IMDB dataset, we keep the ones with a corresponding plot in the MPST dataset [80], a corpus of movie synopses. The GoodReads dataset comes already with books' reviews paired with the related plots; while from the Amazon dataset, among all the product reviews, we keep only the ones related to movies available on the store and whose descriptions consist of the movie plots[3].

With the help of 15 annotators we further labelled more than 18,000 reviews' sentences (~ 6000 per corpus), marking the sentence polarity (`Positive`, `Negative`), or whether a sentence describes its corresponding movie/book `Plot`, or none of the above (`None`)[4]. We ensured that each sentence was labelled by at least two annotators by assigning overlapping subsets of $\sim$ 2400 sentences. In case of disagreement, when no consensus was reached, a final choice was made through a majority vote involving a third annotator. The final inter-annotator agreement (Cohen's kappa) was computed between each pair of annotators sharing a common subset, with a minimum value of 0.572 and maximum of 0.831, for a resulting average of 0.758 (i.e., 0.786/0.739/0.748 for the IMDB, GoodReads and Amazon dataset, respectively.)[5]. It is worth noting that the difficulties in finding an agreement on the sentence annotations are, to some extent, reflected by the performance across the different datasets on the sentiment classification task, as shown in Section §4.5.2. We report the dataset statistics in Table 4.1.

Baselines

We compare the experimental results with the following baselines:

- sLDA [116]: a supervised extension of LDA adding a response variable associated with each document.

[3] The dataset provides a predefined split of the corpus which preserves on train, development and test sets the same distribution of reviews based on their corresponding plots.

[4] We use *Doccano* as framework for collaborative labelling: https://github.com/doccano/

[5] We publicly release the full set of sentences with and without annotations for future expansion.

- JST [98]: Joint Sentiment-Topic model built on LDA which is able to extract polarity-bearing topics.
- NVDM [118]: Neural Variational Document Model, a variational auto-encoder with an encoder network (i.e. an MLP) mapping the bag-of-words representations into continuous latent distributions, and a generative network (i.e. a softmax decoder) reconstructing the document representations.
- GSM [119]: based upon NVDM, the Gaussian Softmax topic model generating the topic distribution by applying a softmax function on the hidden representations of documents.
- NTM [40]: Neural Topic Model is a variation of NVDM by plugging the topic coherence metric directly into the model's objective.
- ProdLDA [155]: with an architecture similar to NVDM, ProdLDA introduces a Dirichlet prior in place of Gaussian prior over the latent topic variable.
- Scholar [22]: a neural framework based on variational inference for the generation of topics incorporating metadata information.

Parameter Setting

We perform tokenisation and sentence splitting with SpaCy[6]. When available, we keep the default preprocessing, as it is the case for sLDA and SCHOLAR. Along with stopwords, we also remove tokens shorter than three characters and those with just digits or punctuation. We set the vocabulary to the 2,000 most common words as the best trade-off for each dataset.

The 300-dimensional word vectors are initialised with a pre-trained BERT embedding [36]. Sentence embeddings are generated from the Sentence-BERT using a pre-trained BERT-large with mean-tokens pooling [144]. We use the predefined split of the MOBO dataset into training, development and test set in the proportion of 80/10/10 and average all the results over five executions.

Hyperparameters

We tune the models' hyperparameters on the development set via a random search over combinations of learning rate $\lambda \in [0.001, 0.5]$, dropout $\delta \in [0.0, 0.6]$ and topic vector size $\gamma_t \in [25, 50, 100, 200]$. Encoder and decoder are configured following

[6] https://spacy.io/

[155]. The hidden representation of documents is set to 100 and sentiment classifier's hidden size to 50. Matrices are randomly initialised with the Xavier and sparse methods [51, 112]. We employ the Adam optimiser [83], set the batch size to 64 and apply batch normalisation as additional regulariser [32].

Sequential Unfreezing

Instead of simultaneously training all the model components, we unfreeze them sequentially. We first freeze the sentiment classifier and update only the autoencoder. At the e_{th} epoch, we unfreeze the sentiment classifier uniquely on the polarised features to let the classifier training. Finally, at the $(e+n)_{th}$ epoch, we unfreeze the adversarial mechanism to drive the generation of neutral features fooling the classifier. We follow an analogous approach with regard to the plot classifier. The values of e and a are treated as hyperparameters and chosen through the random search. We found that the sequential unfreezing scheme leads to better topic disentanglement.

4.5 Experimental Results

We report the results in terms of topic coherence/uniqueness, sentiment classification and topic disentanglement rate. We also perform ablation studies to gain more insights into our model.

4.5.1 Topic Coherence and Uniqueness

Traditionally, topic models have been evaluated through the perplexity over held-out documents [172]. Lower perplexity implied better predictiveness as it aimed at measuring the model goodness-of-fit over a held-out set. However, it has been shown that better perplexity does not imply more comprehensible topics [22, 24] . That is why topic coherence was introduced [87], to evaluate topics regarding their understandability with a score closely matching human judgements. Normalized Pointwise Mutual Information (NPMI) is a topic coherence score shown effective in matching the human judgements [148], and measures the statistical independence of observing two words in close proximity based on the word co-occurrence statistics. The NPMI for a list of words $\boldsymbol{w}$ is defined in Eq. 4.13:

$$\text{NPMI}(\boldsymbol{w}) = \frac{1}{N(N\text{-}1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \tag{4.13}$$

Datasets	**Models**	**Topic Coherence / Topic Uniqueness**			
		25	50	100	200
IMDB	LDA	0.395 / 20.3	0.387 / 30.1	0.383 / 33.9	0.391 / 34.4
	sLDA	0.421 / 15.8	0.376 / 18.9	0.291 / 13.5	0.288 / 14.6
	JST	0.472 / 22.7	0.526 / 26.8	0.527 / 29.3	0.530 / 31.1
	NVDM	0.281 / 15.8	0.284 / 30.2	0.273 / **50.3**	0.266 / **54.8**
	GSM	0.384 / 22.4	0.402 / 21.0	0.410 / 39.7	0.394 / 42.4
	NTM	0.423 / 28.8	0.508 / 28.6	0.513 / 24.1	0.523 / 23.5
	ProdLDA	0.502 / 31.1	0.543 / 30.8	0.566 / 27.7	0.558 / 29.2
	Scholar	**0.550** / 28.4	0.616 / 27.0	0.618 / 29.7	0.593 / 31.5
	DIATOM	0.544 / **37.1**	**0.639** / **38.1**	**0.626** / 36.5	**0.615** / 30.7
	– w/o Plot Network	0.525 / 30.1	0.603 / 36.7	0.607 / 33.8	0.584 / 30.3
GoodReads	LDA	0.441 / 19.6	0.463 / 33.5	0.455 / 41.6	0.462 / 40.3
	sLDA	0.432 / 34.4	0.387 / 47.3	0.313 / 25.6	0.315 / 23.8
	JST	0.465 / 43.5	0.549 / 46.2	0.560 / 47.6	0.551 / 45.2
	NVDM	0.294 / 40.8	0.323 / 30.2	0.287 / 48.3	0.264 / 46.9
	GSM	0.411 / 24.8	0.481 / 40.1	0.482 / 38.1	0.473 / 41.4
	NTM	0.421 / 23.5	0.523 / 47.6	0.493 / 33.4	0.465 / 38.7
	ProdLDA	0.551 / 30.3	0.562 / 41.8	0.564 / 39.8	0.556 / 37.7
	Scholar	0.545 / 38.3	0.603 / 42.0	**0.681** / 41.2	**0.664** / 38.4
	DIATOM	**0.582** / **54.0**	**0.634** / **52.9**	0.628 / **54.9**	0.609 / **48.7**
	– w/o Plot Network	0.555 / 40.1	0.615 / 49.3	0.607 / 33.8	0.578 / 39.6
Amazon	LDA	0.430 / 28.9	0.447 / 47.5	0.438 / 64.8	0.445 / 59.3
	sLDA	0.421 / 67.7	0.393 / 62.1	0.323 / 87.5	0.331 / 74.8
	JST	0.450 / 73.0	0.558 / 71.2	0.544 / 78.8	0.518 / 70.9
	NVDM	0.278 / 42.4	0.310 / 32.5	0.281 / 38.4	0.261 / 49.1
	GSM	0.441 / 53.2	0.451 / 60.0	0.433 / 61.7	0.427 / 64.4
	NTM	0.493 / 52.8	0.501 / 53.1	0.547 / 55.3	0.508 / 59.3
	ProdLDA	0.492 / 63.4	0.543 / 51.4	0.528 / 58.7	0.551 / 62.1
	Scholar	0.548 / 60.5	0.587 / 65.1	**0.641** / 63.2	0.629 / 68.2
	DIATOM	**0.563** / **82.0**	**0.598** / **82.3**	0.626 / **80.8**	**0.636** / **78.5**
	– w/o Plot Network	0.539 / 30.1	0.584 / 78.3	0.611 / 73.4	0.618 / 74.7

Table 4.2: Topic Coherence and Topic Uniqueness results for 25/50/100/200 topics. The best result in each column and for each dataset is highlighted in **bold**.

where $P(w_i)$ and $P(w_i, w_j)$ are calculated based on the word co-occurrences in a reference dataset, and N is typically set to 10, thus considering the top-10 words of topics. The aforementioned definition normalises the PMI in $[-1, 1]$, so that for two words, -1 denotes no co-occurrences while +1 a complete co-occurrence. We evaluate topic coherence using the C_V metric, a slightly refined NPMI score using a boolean sliding window to determine the words' context [148].

We additionally monitor the topic uniqueness (TU) to measure word redundancy across topics. Following Nan et al. (2019), we use $cnt(l, k)$ to denote the total number of times the top word l in topic k appears among the top words across all topics, then $TU(k) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{cnt(l,k)}$. TU is inversely proportional to the number of times each word appears in topics; a higher TU score implies that the top words are rarely repeated and, therefore, more diverse and unique topics.

In Table 4.2, we report the topic coherence and the topic uniqueness values. The supervised document label information was incorporated into sLDA, JST, SCHOLAR and DIATOM. Other models are purely unsupervised. We can observe that among conventional LDA-based models, JST performs significantly better compared to both LDA and sLDA for different topic settings and across all datasets. Neural topic models give mixed results. In terms of topic coherence, the trend is SCHOLAR > PRODLDA > NTM > GSM > NDVM. However, when we examine the topic uniqueness values, we can see that higher topic coherence values do not necessarily lead to higher topic uniqueness values. This shows that the topic coherence value could sometimes be misleading since a high topic coherence could be due to the redundancy of words across topics. We also notice that models with supervised document label information (except sLDA) generally outperform the unsupervised ones. This shows that the document label information can indeed help to extract more meaningful topics. When compared our proposed DIATOM with the baselines, we can observe that it achieves better coherence and topic uniqueness values most of the time, showing the benefit of separating opinion-bearing topics from plot topics by adversarial learning. The importance of the plot network is evident from the results since removing the plot network ("−w/o Plot Network") leads to the degraded topic coherence and topic uniqueness measures.

4.5.2 Sentiment Classification

In this section, we compare DIATOM with other supervised topic models for sentiment classification. The purpose of this evaluation is to highlight the discriminative power of the generated representations for the labels of interest while having attractive and unique properties as topic models, rather than confronting them with current state-of-the-art for text classification. We additionally report some baseline results using a Support Vector Machine (SVM) which has been widely employed on these tasks [129] providing an understanding of the relative differences in performance of different approaches.

Table 4.3 shows the sentiment classification accuracy. In JST, the supervised document label information is only incorporated as prior to the model, while both sLDA and SCHOLAR treat the class label of each document as a response variable and jointly model both documents and their responses. We can observe that the latter is more effective in incorporating supervised information since both sLDA

Models	**IMDB**	**GoodReads**	**Amazon**
SVM			
+ TFIDF	0.672 ±0.02	0.711 ±0.01	0.661 ±0.02
+ TFIDF + Lexicon	0.683 ±0.02	**0.719** ±0.02	0.667 ±0.02
+ LDA	0.615 ±0.02	0.659 ±0.02	0.594 ±0.01
BERT	0.806 ±0.02	0.783 ±0.03	0.754 ±0.02
RoBERTa	0.827 ±0.02	0.811 ±0.03	0.759 ±0.02
XLNet	*0.843* ±0.01	*0.824* ±0.02	*0.782* ±0.02
sLDA	0.637 ±0.01	0.652 ±0.01	0.579 ±0.01
JST	0.639 ±0.01	0.518 ±0.01	0.538 ±0.01
SCHOLAR	0.645 ±0.02	0.673 ±0.03	0.613 ±0.02
DIATOM	0.726 ±0.03	0.704 ±0.02	**0.686** ±0.02
– w/o Plot Network	**0.734** ±0.03	0.695 ±0.03	0.603 ±0.02

Table 4.3: Sentiment classification accuracy with 50 topics over the test set. Best performance from pre-trained models are highlighted in *italic*, all the others in bold.

and SCHOLAR outperform JST in general. But DIATOM gives significantly better results all over the baselines with the improvement over the best baseline model, SCHOLAR, by 3-8%. In our models, features used for sentiment classification are opinion-bearing topics. This shows that separating opinion topics from plot/neutral topics is beneficial for sentiment classification. We also observe that the contribution of the plot network to sentiment classification is dataset-dependent. The usage of the plot network largely boosts the sentiment classification accuracy by over 8% on the Amazon dataset. But its effect is negligible on the other two datasets.

When compared with traditional sentiment classification models such as SVM, we found that DIATOM outperforms SVM trained with various features on both IMDB and Amazon. But it performs slightly worse than SVM trained with TFIDF features with or without the additional incorporation of sentiment lexicon features. Nevertheless, DIATOM gives superior performance compared to SVM trained on LDA topic features in the range of 5-11%, showing the effectiveness of using opinion topics for sentiment classification.

To provide a thorough assessment of the MOBO dataset on the sentiment classification task, we also report the results obtained using the recently developed transformer-based architectures [168]. In particular, we employed three state-of-the-art language models, i.e., BERT [36], RoBERTa [104], and XLNet [193], which outperform by a significant margin all the other baselines. However, compared to

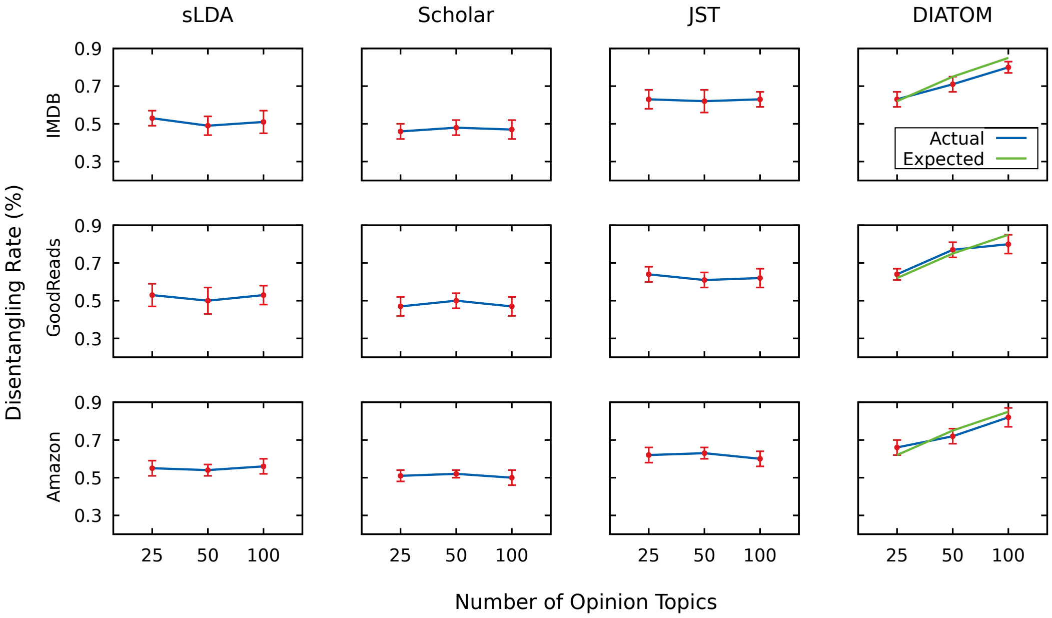


Figure 4.3: Disentangling rate of topic models across different number of topics.

the other architectures, these models are intensively pretrained on large external corpora and do not provide any topical representations of the documents.

4.5.3 Topic Disentanglement

None of the aforementioned measures can, however, capture how opinion and plot topics are distributed. To this aim, we use topic labeling to assign a proxy label (`Positive`, `Negative`, `Plot`, `None`) to each topic and then measure the topic-disentanglement rate ρ (Eq. 4.14) as the proportion of opinion-bearing topics with respect to the overall set of topics, complementary to the proportion of plot/neutral topics:

$$\rho = \frac{S}{S+K} \tag{4.14}$$

with S being the number of opinion topics and K the number of plot/neutral topics.

For each topic, we first calculate its embedding by taking the normalised weighted average of the vectors of its top N words: $\overrightarrow{t_z} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \times \overrightarrow{w_i}$, where α_i is the normalised distribution of word w_i in topic z. We then retrieve the top 10 most similar sentences from the human-annotated sentence set measured by the cosine similarity between the topic embedding and each sentence embedding. The sentence embedding is computed using the Sentence-BERT encoder [144]. The most frequent label among the retrieved sentences is adopted as the topic's label.

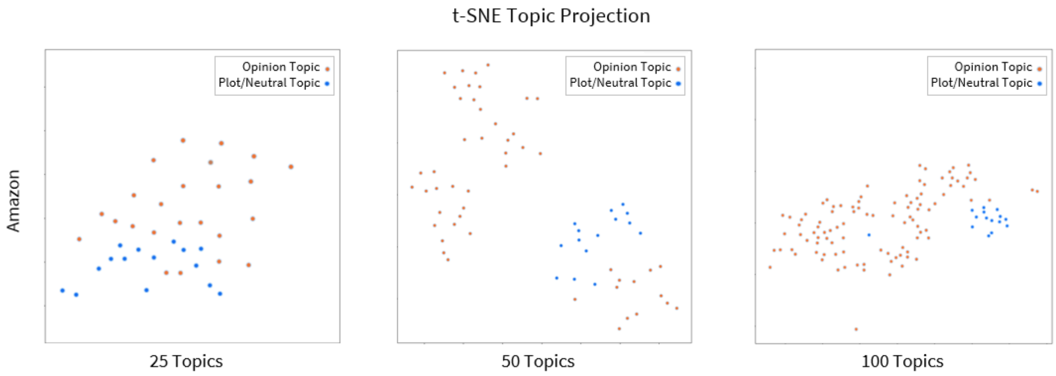


Figure 4.4: Example of t-SNE projection for the Amazon dataset of the topic distribution for different number of topics. Color are assigned according to plot/neutral and opinion topics.

To highlight the disentanglement capability of DIATOM, in Figure 4.3, we analyse how the proportion of opinion-bearing topics varies across standard and sentiment topic models. We notice that despite the signal from the document labels, sLDA and SCHOLAR tend to produce topics rather balanced in terms of neutral and opinion-bearing topics. JST has a more skewed distribution towards opinion topics. DIATOM instead generates an actual proportion of opinion topics approaching the expected proportion set up by the model, demonstrating the capability to control the generation of plot and opinion-bearing topics.

In Table 4.4 and Table 4.5, we show a set of topics grouped according to the disentanglement induced by DIATOM. For each topic, we report an excerpt of the most similar sentences retrieved. Aside from being overall coherent, we can guess rather paradigmatic themes as the IMBD-Topic 1 about peace and war between countries, or more peculiar plots related for instance to "*The Hobbit*" or "*Batman*" as in the Amazon topics. It is worth having a closer look at the IMDB-Topic 2, which despite the "negative" theme of depression and suicide, the model is able to correctly gather those words under the same plot topic. The opinion-bearing topics report a collection of commonly appreciated or critical aspects; some of them are mainly collections of related adjectives with the same polarity (e.g. IMDB-Topic 1), while others are made up of mixed terms describing the issues and the associated experience (e.g. Amazon-Topic 2).

4.5.4 Visualisation

Another way to look at the disentangled topics is through the visualisation of topic vectors.

Plot/Neutral Topics
IMDB - Topic 1 *Government, Country, Peace, Information, Free, Plane, Theory, Anti, Soldier, Hitler*
1. Groundbreaking in the realm of socially relevant drama, it dealt with issues such as abortion, domestic violence, student protest, child neglect, illiteracy, slumlords, the anti-war movement, [...]. 2. This effort by Charlie ultimately evolves into a major portion of the U.S. foreign policy known as the Reagan Doctrine, under which the U.S. expanded assistance beyond just the [...].
IMDB - Topic 2 *Window, Hospital, Apartment, Suicide, Commit, Pitt, Serial, Strange, Killer, Mental*
1. Even re-think why two boys/young men would do what they did - commit mutual suicide via slaughtering their classmates. 2. It’s the patented scene where the killer creeps up behind the victim.
GoodReads - Topic 1 *Cure, Plague, Trial, Betray, Thomas, Secret, Dashner, Ball, Betrayal, Wicked*
1. Blaming Cinder for her daughter’s illness, Cinder’s stepmother volunteers her body for plague research, an ”honor” that no one has survived. 2. By age thirteen, she has undergone countless surgeries, transfusions, and shots so that her older sister, Kate, can somehow fight the leukemia that has plagued her since childhood.
GoodReads - Topic 2 *Teenager, Fault, Illness, Mental, Depression, Maddy, Grief, Bully, Topic, Greg*
1. She’s got a lot of mental strength, having been ostracized for most of her life. 2. She went through a divorce, a crushing depression, another failed love, and the eradication of everything she ever thought she was supposed to be.
Amazon - Topic 1 *Dent, Gotham, City, Gordon, Bruce, Wayne, Harvey, Joker, Criminal, Nolan*
1. Being imprisoned Batman has enough time to paint a gigantic flaming bat on a bridge while people are literally being executed on the hour. 2. Batman gets with Catwoman... after how hard she sold him out?
Amazon - Topic 2 *Gandalf, Frodo, Jackson, Tolkien, Dwarf, Fellowship, Peter, Orc, Ring, Hobbit*
1. [...] the myriad inhabitants of Middle-earth, the legendary Rings of Power, and the fellowship of hobbits, elves, dwarfs, and humans–led by the wizard Gandalf (Ian McKellen) and the brave hobbit Frodo. 2. This is the beginning of a trilogy; soon to be finalized.

Table 4.4: Example of *plot/neutral* topics extracted by DIATOM and their associated most similar sentences.

Opinion-Bearing Topics
IMDB - Topic 1 *Badly, Stock, Remove, Poorly, Hype, Ridiculous, Insult, Disaster, Excuse, Lame*
1. I can't imagine how anyone could have read this badly written script and given it the greenlight. 2. Although there has obviously been a lot of money spent on them the numbers are badly staged and poorly photographed.
IMDB - Topic 2 *Exceptional, Recommend, Excellent, Craft, Believable , Overlook, Vhs, Solid, Festival, Amaze*
1. Overall, this is a good film and an excellent adaption. 2. It's great acting, superb cinematography and excellent writing.
GoodReads - Topic 1 *Negative, Judge, Note, Pretend, Embarrass, Quality, Extreme, Guilty, Fake, Borrow*
1. Can you give something negative stars? 2. And while it must be hard reading negative reviews you need to be able to deal with this in a graceful way (no one likes a sore loser).
GoodReads - Topic 2 *Teen, Nice, Normally, Little, Genre, Amuse, Theme, Enjoyment, Blow, Reread*
1. What would have made the book a lot more fun to read was more meatier characters in the other girls. 2. But I feel like that was part of the fun of it.
Amazon - Topic 1 *Expectation, Quality, Definitely, Great, Good, Worth, Graphic, Predictable, Compare, Decent*
1. Action is good. 2. Rachel Weisz was "mostly" good.
Amazon - Topic 2 *Price, Shame, Service , Normally, Purchase, Connection, Greed, Stream, Watch, Frustrate*
1. This experience leaves me skeptical of the Amazon Prime video service. 2. Look closely before purchasing.

Table 4.5: Example of *opinion* topics extracted by DIATOM and their associated most similar sentences.

Datasets	Models	Accuracy	TC / TU
IMDB	DIATOM	0.726 ± 0.03	0.639 / 38.1
	– w/o orth reg.	0.723 ± 0.01	0.582 / 27.5
	– w/o sent. class.	0.491 ± 0.03	0.601 / 35.4
	– w/o both	0.478 ± 0.03	0.544 / 25.4
	– w/o Plot Net	0.734 ± 0.03	0.603 / 36.7
GoodReads	DIATOM	0.704 ± 0.02	0.634 / 52.9
	– w/o orth. reg.	0.681 ± 0.02	0.612 / 41.1
	– w/o sent. class.	0.446 ± 0.02	0.638 / 47.6
	– w/o both	0.410 ± 0.02	0.552 / 39.6
	– w/o Plot Net	0.695 ± 0.03	0.615 / 49.3
Amazon	DIATOM	0.686 ± 0.02	0.598 / 82.3
	– w/o orth reg.	0.682 ± 0.01	0.605 / 55.3
	– w/o sent. class.	0.601 ± 0.03	0.573 / 76.9
	– w/o both	0.548 ± 0.03	0.567 / 52.1
	– w/o Plot Net	0.603 ± 0.02	0.584 / 78.3

Table 4.6: Ablation study over DIATOM by removing the orthogonal regularisation, the sentiment classifier or just the auxiliary Plot Network.

As an example, we plot in Figure 4.4 the 2-dimensional representation of the topic distributions projected by t-SNE for the Amazon dataset. Different colours represent different types of topics generated by DIATOM, namely plot/neutral in blue and opinion in red. We notice how consistently across a different number of topics, plot/neutral topics tend to cluster together, with the boundary close to some polarised topics likely to share common features, as shown in Figure 4.1 in which the plot topic and the negative topic share a common word *'Dwarf'*.

4.5.5 Ablation Study

We report in Table 4.6 the results of the ablation study on DIATOM. We observe that removing the orthogonal regularisation has a limited effect on sentiment classification, but causes a fluctuation on topic coherence and a clear drop in topic uniqueness. A significant classification performance drop is observed by removing the sentiment classifier, which essentially reduces DIATOM to an unsupervised model. Removing both the orthogonal regularisation and the sentiment classifier shows a major negative impact on both accuracy and the topic's quality.

Finally, we assess the influence of the plot network (§4.3), and while we do not notice any consistent impact across the datasets in terms of sentiment classification, the quality of topics has a notable drop in terms of coherence and diversity.

4.6 Summary

We have described DIATOM, a new neural topic model to generate disentangled topics through the combination of VAE and adversarial learning.

We reported the results of our experimental study based on the novel MOBO dataset highlighting the benefit of such an approach leading to topics with higher interpretability in terms of both topic coherence and topic uniqueness and more discriminative power reflected in better sentiment classification results compared to other supervised topic models.

Finally, we further discussed the model capability to consistently disentangle opinion-bearing topics from plot/neutral ones measuring the introduced disentangling rate.

Chapter 5

Topical Phrase Extraction from Clinical Reports

Chapter Abstract

Making sense of words often requires to simultaneously examine the surrounding context of a term as well as the global themes characterising the overall corpus. Several topic models have already exploited word embeddings to leverage the word local context, however, this has been weakly combined with the global context during the topic inference. In this chapter, we introduce Context-GPU, a topic model for topical phrase extraction, which by means of the Pólya urn model corroborates the word embedding information with the global context detected by the Latent Semantic Analysis. To highlight the effectiveness of this combined inference, the model was assessed in analysing clinical reports, a challenging scenario characterised by technical jargon and limited word statistics. Experimental results have shown it outperforms the state-of-the-art methods in terms of both topic coherence and computational cost.

5.1 Introduction

Topic models have been extensively used to generate synthetic representations of the main themes characterising a large document collection. Documents are traditionally represented under the bag-of-words assumption, a simple but effective representation that ignores the word orders, but in spite of this has shown remarkable results [16]. However, this assumption has commonly led to the extraction of unigram topics, relying on the word co-occurrence patterns across documents. This has notably narrowed the topic expressiveness as many domain-specific documents might include concepts that are unfolded in multiple words rather than in a single term, and the shared semantic of these words is solely based on their global context. Clinical reports are a prominent example of this family as medical concepts are often expressed in terms of multi-word phrases. For example, the phrases "*white blood cell*" or "*blood sugar*" would lose their meaning if decomposed as unigrams; in addition, the word *cell* and *sugar* might be wrongly put under the same topic because of the shared *blood* term.

Recently, word embeddings have gained an increasing interest thanks to their capability to leverage the word's local context, with an improved efficiency in representing words as continuous vectors of a low-dimensional space [36, 79, 120]. The resulting embeddings have been proved to encode numerous syntactic and semantic relations (e.g., similarities or analogies) based on the local context of words [92], and therefore, several works tried to combine topic models with word embedding [25, 94, 127, 130]. This generally resulted in an increased expressiveness of the discovered topics due to the word properties geometrically encoded in the word embeddings. However, the resulting models commonly entail a high computational cost, and the coherence of the generated topics is still negatively affected by the inherent limitations affecting these word embeddings, such as the *topic shifting* issue [145]. Indeed, words that share similar context windows might potentially be treated as directly co-related into the embedding space, with a misleading word similarity in case of antonyms (e.g. *tall* and *short*) or co-hyponyms (e.g. *schizophrenia* and *alzheimer*). In turn, this would lead the topic models to clustering words that are not strictly related despite sharing a common context or domain.

The computational cost required to combine word embeddings and topic models can be reduced by adopting the *Generalised Pólya urn model* [109]. Although the Latent Dirichlet Allocation (LDA) [16] already used the *Simple Pólya urn model*, its generalised version proposed in Mimno et al. (2011) allows incorporating word relatedness directly into the inference process, using the corpus statistics. We posit that a simple but effective extension of the *Generalised Pólya urn model* would consist of evaluating, instead, the word relatedness based on the word embedding scores. Concurrently, the coherence of the generated topics can be improved by mitigating the impact of the topic shifting issue by jointly considering the global and local context of a word, so that if two terms appear in similar context windows but do not share similar global contexts (i.e. corpus themes), they probably convey different topics.

We propose a Context-aware Pólya urn model (Context-GPU) to generate topics by extracting topical phrases combining the local and global context of words/phrases[1]. We first detect the medical phrases in clinical reports by means of an off-the-shelf medical concept extraction tool; hence, the phrases extracted are thus reliable and clinically relevant. Then, we use a modified Generalised Pólya urn model, which promotes words/phrases under the same topic if they are close neighbours in the window-based embedding (local context) as well as in the corpus-based embedding (global context) space. The window-based embedding improves the capability to detect semantic relatedness at the phrase level; also, it encodes word co-occurrences from an external source of knowledge (e.g. Wikipedia) alleviating the lack of statistics for technical terms. Simultaneously, the corpus-based embedding provides information about the global context, inducing coherent topics closer to the particular themes discussed. To the best of our knowledge, this is the first time local and global contexts are combined for topical phrases extraction. Our experimental results have shown the effectiveness of this approach outperforming the previous methods in terms of quality of topics, topic coherence and efficiency.

We proceed to describe the related work (§5.2). We then give a background of the Pólya urn model (§5.3) before presenting the proposed approach (§5.4). Finally, we discuss our experimental results with a thorough comparison with with the state-of-the-art approaches for topical phrase extraction (§5.5).

5.2 Related work

Our work is related to three lines of research, phrase embedding learning, topic modelling incorporating word embeddings and using latent topics for language model learning.

5.2.1 Phrase Embedding Learning

Distributional semantic models (i.e. word embeddings) have recently been applied successfully in many NLP tasks [92]. Neural network based approaches have become more efficient, allowing their use in multiple scenarios, thanks to the *skip-gram with negative-sampling training method* (SGNS), [120, 121]. It was widely popularised via *word2vec*, a software to create word embeddings. Recently, a new word embedding method has been proposed, called *FastText* [79], which treats each word as made of character n-grams. Vector representations are then computed from

[1] https://github.com/gabrer/context_gpu/

the sum of their n-gram representations. More traditional vector representations are based on a dimensionality reduction obtained by applying the Singular Value Decomposition (SVD) to the weighted document-term matrix of the corpus; Latent Semantic Analysis (LSA) [34] is a prominent method following this approach.

Phase embeddings can be simply taken as the average of their constituent word embeddings. If treating each phrase as a single term, its representation can also be learned from data directly using word representation learning methods such as LSA, SGNS or FastText. There have also been compositional semantic models that aim to build distributional representations of a phrase from its constituent word representations using Convolutional Neural Networks (CNNs) [89], based on features that capture phrase structure and context [198] or using convolutional tensor decomposition [72].

5.2.2 Topic Modelling Incorporating Word Embeddings

To exploit the information encoded into word embeddings, several models have been proposed combining topic models and word embedding representations. Gaussian LDA [33], for instance, use pre-trained word embeddings learned from large corpora (e.g., Wikipedia) to model topics as Gaussian distributions over the vector representations, defining topics as random samples from a multivariate Gaussian distribution whose mean is the topic embedding.

Nguyen et al. (2015) proposed to use the word embeddings pre-trained from a large external corpus as latent word features to define categorical distributions over words, which is called a latent feature component. The original topic-to-word Dirichlet multinomial component in LDA which generates the words from topics is then replaced by a two-component mixture of the original Dirichlet multinomial component and a latent feature component. But model learning is difficult because of the coupling between the two components.

An alternative approach is TopicVec [94] which replaces the multinomial topic-word distribution with a probability function, it computes a focus word from a topic and word neighbours within the embedding; in TopicVec this link function is in addition combined with a context word embeddings along with the topic embedding and the focus word embedding.

Li et al. (2016) measured the word relatedness based on pre-trained word embeddings and used it to modify the Gibbs sampling inference in a generalised Pólya urn model; overall, this strategy significantly reduces the computational cost compared to the aforementioned approaches. However, it is not only entirely focused on the short-text analysis (i.e. one document, one topic), but it does not exploit any global context to mitigate the effects of the topic shifting issues induced by word embeddings. Moreover, it did not explore the benefit of using a *n*-gram word embedding, such as FastText, against the word-oriented embeddings.

5.2.3 Using Latent Topics for Language Model Learning

While the aforementioned approaches incorporate word embeddings into the topic model generation, there have also been attempts to make use of latent topics to improve language models. Dieng et al. (2017) proposed TopicRNN in which the global semantics come from latent topics as in typical topic modelling, but local semantics is defined by the language model constructed using Recurrent Neural Networks (RNNs). The separation of global vs local semantics is achieved using a binary decision model for stop words. Topic vectors here are also sampled from a Gaussian distribution with zero mean and unit variance and are refined during language model learning. In a similar vein, Lau et al. (2017) proposed a topic-driven neural language model that also incorporates document context in the form of latent topics into a language model implemented using Long Short-Term Memory (LSTM) networks. They essentially treated the language and topic models as subtasks in a multi-task learning setting, and trained them jointly using categorical cross-entropy loss.

5.3 Pólya Urn Models

In this section, we give a background of both simple and generalised Pólya urn Models. We describe how they can be used for topic extraction, before presenting in the next section our proposed approach that extends them to exploit word contexts.

As shown in Mimno et al. (2011), the simple LDA model might not be able to fully capture the already available statistics of word co-occurrences in a corpus. Detecting semantic similarity between words is challenging due to the power-law characterisation of natural language, i.e., words sharing a common semantic might rarely co-occur together and hence being overlooked. A more effective model called Generalised Pólya urn model was proposed in Mimno et al. (2011), by extending the Simple Pólya urn model used in LDA where the topic-word component is updated in order to strengthen the associations between related words under the same topic.

5.3.1 Simple Pólya Urn Model

The generative process of LDA can be interpreted by means of Pólya urn model [109], a statistical model describing objects of interest (e.g. words or topics) in terms of coloured balls and urns.

In the context of topic models, balls can be considered as words and urns as topics; in particular, LDA follows the so-called *Simple Pólya urn* (SPU) model. In the main step of this process, a coloured ball is randomly drawn from an urn and is put back along with an additional new ball of the same colour; this induces a self-reinforcement process known as "rich get richer", since the probability of seeing a specific coloured ball from an urn increases every time this ball has been drawn.

Likewise, LDA follows the SPU model by employing two kinds of urns: topic-document and word-topic urns. The topic-document urns hold balls whose colour corresponds to different topics in a document, while the balls in the word-topic urns represent different words in a topic. The generative process proceeds as follows: a ball is extracted from the topic-document urn d_m, and its colour determines the new topic assignment $\hat{z}$, then the ball is put back along with another ball of the same colour. Next, a ball is extracted from the word-topic urn $\hat{z}$ determining a new word $\hat{w}$ and, as before, the ball with an additional one of the same colour is put back into the urn. As a result, both the topic $\hat{z}$ and the word $\hat{w}$ increase their proportion in the topic-document and word-topic distribution, respectively.

5.3.2 Generalised Pólya Urn Model

The described process is intrinsically biased to promote together words that frequently occur in a corpus, overlooking less prominent but correlated words. To alleviate this shortcoming and increase the association strength between rare but still related words, a *Generalised Pólya Urn* (GPU) model was proposed by Mimno et al. (2011). It incorporates a corpus-specific word co-occurrence metric into the generative process affecting the probabilities of related words under the same topic.

Unlike the aforementioned simple version, in a generalised Pólya urn model when a ball of colour $\hat{w}$ has been drawn, A_{vw} additional balls of several colours $v = \{1, ..., W\}$ are placed into the urn. This process increases, not only the probability of the observed word $\hat{w}$, but also the probability of its related words, and is commonly referred as *promotion* of the coloured balls [43]. Specifically, the LDA inference process now relies on a modified Gibbs sampling algorithm which simultaneously increases the probability of a word and its correlated terms at each iteration. Word relatedness is computed by weighting word co-occurrences using the standard

Inverse-Document Frequency (IDF) weighting strategy $\lambda_v = log(D/D(v))$, where D is the number of documents and *D(v)* is the number of documents where the word v occurs at least once; this weight has the beneficial property of being higher for rare words increasing their prominence.

However, the effectiveness of this approach strongly depends on how accurately word correlations are identified. Although the GPU framework proposed by Mimno et al. (2011) has improved the average quality of mined topics, it still relies exclusively on the global context of words (i.e. word co-occurrences in the corpus) and might completely overlook the sentence-specific meaning of a word conveyed by the word's local context.

This drastically narrows the model's capability to deal with multiple senses of words. For example, looking at the sentences *"White blood cell count is low."* and *"This raises the blood sugar back to its normal level."*, current models might put under the same topic words like *"cell"* and *"sugar"*, which are rather unlikely to appear coupled in a sentence. Moreover, similar issues can be experienced analysing documents characterised by technical jargon, which occur few times in corpus (i.e. poor statistics) and might exhibit a peculiar meaning for every phrase (i.e. multiple meanings).

5.4 Context-Aware Pólya Urn (Context-GPU) Model

In this section, we propose a modified Gibbs sampling algorithm to conduct a context-driven inference to cope with the described limitations. It exploits a word representation based on general sources of knowledge providing rich word statistics and takes into account simultaneously the local and global context of words to disambiguate irrelevant terms.

Our hypothesis is that the Generalised Pólya Urn model can be modified and enhanced to provide a framework combing the local and global context of words. Local context is determined by a word embedding based on context window and trained on a large source of general knowledge (e.g., Wikipedia). Rather, the global context relies on the word representations obtained considering the term co-occurrences within a corpus. As a result, both local and global context can be incorporated into a context-aware Pólya urn model called *Context-GPU*, a generative model which is able to capture the semantics of a word at both the sentence and document level, mitigating the effects of the topic shifting issue on the generated topics.

Before presenting our proposed Context-GPU, we first describe how we extract medical phrases from clinical documents.

5.4.1 Medical Phrase Extraction

Medical terms in clinical documents are often expressed in multi-word phrases, for example, "*arterial blood gas*" and "*heart transplant*". These phrases are not semantically decomposable, as once split into unigrams, they would lose their original semantic meanings.

We use an open-source clinical annotation tool *MedTagger*[2] which extracts and annotates concepts from clinical reports by leveraging knowledge bases, machine learning and syntactic parsing. The output of MedTagger provides detailed information about the medical concept detected, such as attributes, uncertainty, semantic group (i.e. Diagnosis, Test and Treatment), and so on. Also, it has achieved the state-of-the-art performance in terms of F-Measure (0.84) at the *i2b2 NLP challenge* on the concept mention task [100].

Clinical reports are also characterised by many occurrences of medical abbreviations to favour brevity due to a large amount of information that needs to be synthesised in a short time and limited space. Detection of medical concepts through MedTagger is not only much more reliable than other general-purpose techniques for phrase extraction, but also allows to effectively detect and preserve the medical abbreviations. Once medical phrases are detected, they are represented by compound words where constituent words are joined together by an underscore. E.g. words that compose the phrase "*short of breath*" are substituted by the compound word "*short_of_breath*", "*saphenous vein graft*" by "*saphenous_vein_graft*", and so on.

However, treating phrases as compound words leads to more severe data sparsity since phrases sharing common lower-order *n*-grams, such as "*right coronary artery*" and "*left coronary artery*" would be considered as two totally different terms. Also, preserving both the multiple words and the compound phrase is not a solution, as the phrases are naturally less frequent than individual words and would be ranked with lower probabilities.

To this end, once multiple words are substituted by a compound word, in the Context-GPU we adopt the FastText embedding [79], a word embedding oriented by design to deal with sub-grams composing words. Thus, it naturally fits the need to detect the similarity between a phrase and its constituent words. For example, in our trained FastText embeddings the word *saphenous_vein_graft* has neighbours such as *saph*, *aphe*, but also *vein* and *graft*. Therefore, we combine FastText with the Pólya urn model to increase the probability to see under the same topic the words *vein* or *graft* once we come across the phrase *saphenous vein graft*, and vice versa.

[2] http://ohnlp.org/index.php/MedTagger

5.4.2 Local Context

Some commonly used embedding have been the SVD [92] and SGNS (i.e. word2vec) [120], and only recently FastText [79], and they all provide vectors encoding both syntactic and semantic information about a word at its local context window in a large corpus.

Two characteristic features of these embeddings are here exploited. The first is that words are represented by a vector trained with regard to the local contexts where the words are likely to appear. Therefore, it can be used to reinforce ties among words sharing common uses in phrases (e.g. *alzheimer* and *schizophrenia*). The second feature is that word embeddings are commonly trained on a large external source of data (e.g. Wikipedia), hence they can mitigate the low statistics of infrequent technical jargon or rare words in a corpus.

Words are considered related based on the geometric proximity of their vector representations. We propose two strategies to extract related words: a threshold and a Top-N approach. In the threshold approach, words are considered relevant when their respective cosine distances with the target word are lower than a predefined threshold. Alternatively, the Top-N approach extracts a fixed number N of the closest words regardless of their actual distances. In the former approach, the number of neighbours is not fixed for different words, while in the latter, the number is fixed but also unrelated words could be added to the neighbour set.

5.4.3 Global Context

Although topics extracted by combining word embedding and Pólya urn model are more consistent with the occurrence pattern of words in sentences, word embeddings have some well-known shortcomings related to antonyms (e.g. *tall* and *short*) or co-hyponyms (e.g. *schizophrenia* and *alzheimer*). Indeed, these are words that might share similar context windows and then be potentially treated as directly correlated into the embedding space. To avoid any semantic shift resulting from word ambiguities, we balance the local context information with the corpus-specific context computed by applying the Latent Semantic Analysis (LSA) [34].

In particular, we use LSA to learn latent topics from data by performing Singular Value Decomposition (SVD) on the $V \times D$ term-document count matrix where V is the vocabulary size and D is the number of documents. SVD factorises such a matrix into the product of three matrices, W, Σ, and $C^{\intercal}$. In $W \in \mathbb{R}^{V \times m}$, each row represents a word and each column represents a dimension in a latent space that is orthogonal to each other. Σ is a diagonal $m \times m$ matrix that contains singular values along the diagonal indicating how important each latent dimension is. In $C^{\intercal} \in \mathbb{R}^{m \times D}$, each row represents one of the latent dimensions and each column represents a document. If taking the top k latent dimensions in W, we will have a reduced matrix $W_k \in \mathbb{R}^{V \times k}$ where each word is essentially represented by a dense k-dimensional vector. Hence, using LSA, we will be able to generate another set of word embeddings based on global context. For each word, we can then retrieve its related words using the threshold or Top-N approaches mentioned above.

One may argue that topic models such as LDA already captures the global context information by compressing the original document into a lower-dimensional bag-of-topics representation. It is worth noting that LSA learns latent topics by performing SVD on the term-document count matrix, and as a result, the topics are assumed to be orthogonal. LDA uses generative probabilistic models to generate latent topics that are represented as word distributions, and it uses Dirichlet priors for both the document-topic and topic-word distributions. In LDA, topics are allowed to be non-orthogonal. So although both LSA and LDA try to capture the global context, the topic results would be somewhat different. It has been pointed out previously that in some scenarios LSA outperforms LDA providing better quality topics [13]. As will be shown in our experiments, additionally incorporating the global context derived by LSA into the context-aware Polya Urn model gives better performance.

Words are likely to express a common topic, not only when sharing a common local context window (i.e. FastText similarity), but also a global context (i.e. LSA similarity) depending on the analysed documents. Therefore, we first extract the word neighbours from the embeddings based on both the local and the global contexts. Then, we preserve only those terms in the intersection of the two sets, therefore improving the probability that words in a topic are related via both their local and global contexts.

5.4.4 Topic Inference

Given the set of documents $\mathcal{D}$ and the topic assignments $\mathcal{Z}$, the conditional posterior probability of a word w in a topic z follows the standard generalised Pólya urn model [123]:

$$P(w|z, \mathcal{W}, \mathcal{Z}, \beta, \mathbf{A}) = \frac{\sum_v N_{v|z} A_{vw} + \beta}{N_z + |\mathcal{V}|\beta} \tag{5.1}$$

where $\mathbf{A}$ is a *promotion matrix* that expresses whether two words are related to each other, i.e. if one should influence the expectation to draw the other one.

Algorithm 1 Training procedure of the Context-aware Pólya urn model.

```
Input: Corpus C, K topics, $\alpha$, $\beta$, thresholds $\tau$ and $\sigma$
Output: Posterior topic-word distribution
/* Medical phrase extraction */
$C_p \leftarrow MedTagger.PhraseDetection(C)$;

/* Local and global neighbors */
for $v \in \mathcal{W}$ do
    $\mathcal{P}_v \leftarrow WindowEmbedding.Neighbors(v)$;
    $\mathcal{Q}_v \leftarrow CorpusEmbedding.Neighbors(v)$;
end for

/* Promotion matrix */
$A_{v,w} \leftarrow ComputePromotionMatrix(\mathcal{P}_v, \mathcal{Q}_v)$

/* Generalised Pólya Urn sampling */
for $d \in \mathcal{D}$ do
    for $w_n \in \boldsymbol{w}^d$ do
        $N_{z_i|d_i} \leftarrow N_{z_i|d_i} - 1$
        for all $v$ do
            $N_{v|z_i} \leftarrow N_{v|z_i} - A_{vw_i}$
        end for
    end for
    sample $z_i \propto (N_{z|d_i} + \alpha_z)\frac{N_{w_i|z}+\beta}{\sum_{z'} N_{w_i|z'}+\beta}$
    for $w_n \in \boldsymbol{w}^d$ do
        $N_{z_i|d_i} \leftarrow N_{z_i|d_i} + 1$
        for all $v$ do
            $N_{v|z_i} \leftarrow N_{v|z_i} + A_{vw_i}$
        end for
    end for
end for
```

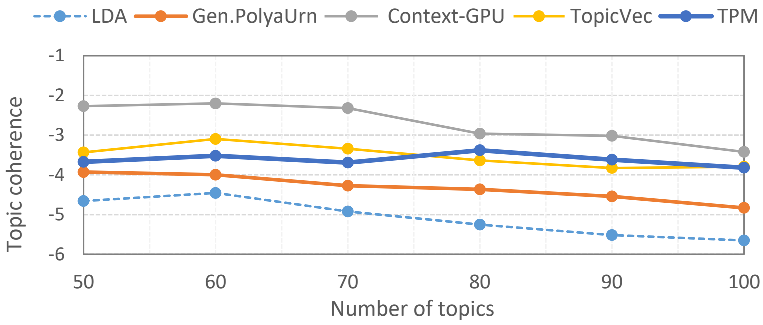


Figure 5.1: Context-GPU: topic coherence scores vs. number of topics.

The promotion matrix is critical for the overall algorithm performance, as it concisely expresses the available information about word relatedness. We propose to set the values of **A** by computing the word relatedness as a result of the $\mathcal{P}$ neighbours provided by the local context embeddings and the $\mathcal{Q}$ neighbours from the global context embeddings. For a word v, another word w is promoted if it is v's neighbour both at the local level (i.e., based on its local context embedding) and the global level (i.e., based on its global context embedding), as expressed in Eq. 5.2. Thus, only if both words are correlated in both the local and global context embedding space, their corresponding cell value in **A** is updated to increase their probabilities to be drawn under the same topic. In the particular case in which **A** correspond to the identity matrix, the model collapses into the Simple Pólya urn model, providing the posterior probability of a word w under a topic z for the standard LDA.

The training procedure of our proposed context-aware Pólya urn model is shown in Algorithm 1. The Gibbs sampling inference can be more complex and expensive due to the non-exchangeability property of words in the generalised Pólya urn model (i.e. under the same topic, the joint probability of words is not invariant to permutation). Therefore, we follow the same approach adopted in Mimno et al. (2011), considering each word as it was the last one during the inference process, ignoring the effect for subsequent words and their topic assignments.

$$A_{v,w} = \begin{cases} 1 & \text{if } w \in (\mathcal{P}_v \cap \mathcal{Q}_v) \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

5.5 Experiments

We assess the effectiveness of our proposed Context-GPU using the data released as part of the *i2b2 Natural Language Processing Challenges for Clinical Records* [165]. The corpus consists of 1,243 de-identified discharge summaries, characterised by medical jargon, which describes medications, dosages, modes (e.g. oral, intravenous, etc.), frequencies, reasons for the treatment, and so on. Hence, we adopted this dataset to assess the model efficacy to deal with multi-phrase concepts and domain-specific jargon.

Clinical reports are pre-processed by removing the common English stop words as well as the clinical-related stop words (e.g. "Dr.", "medical problem", "discharge", etc.). We filter out the most frequent ten words and the words occurring less than five times. We use the MedTagger software to detect medical phrases and represent them as bag-of-phrases within the documents. We do not perform stemming. As a result, in the "bag-of-words" setting the vocabulary size is 7,883, while in the "bag-of-phrases" setting it further increases to 9,932.

Word embeddings are trained on a snapshot of Wikipedia 2015, combined with the i2b2 dataset. We use the *hyperwords* library[3] [92] to train the 300-dimensional SVD and SGNS embeddings, configured with the default parameters. Likewise, we train the 300-dimensional FastText embedding using the library provided by Facebook Research[4], with n-gram sizes set between 2 and 6. The LSA representation adopted as local context is computed on the i2b2 dataset; we use the S-space library[5] to compute the final 300-dimensional representation of words and documents.

We train Context-GPU and set θ and σ to 0.7 and 0.8, respectively, based on a grid search of values in $[0.5, 0.6, 0.7, 0.8, 0.9]$, using 5-fold cross validation. We set the maximum number of Gibbs sampling iterations to 1500. We compare Context-GPU with the following baselines:

- LDA. We use the LDA implementation in MALLET[6] with the default settings and perform hyperparameter optimisation every 200 iterations.
- Generalised Pólya urn (GPU) model [123]. We implemented this algorithm by modifying the LDA implementation in the MALLET library.
- TopicVec [94]. We use the available implementation[7] with the default config-

[3] https://bitbucket.org/omerlevy/hyperwords

[4] https://github.com/facebookresearch/fastText

[5] https://github.com/fozziethebeat/S-Space

[6] http://mallet.cs.umass.edu

[7] https://github.com/askerlee/topicvec

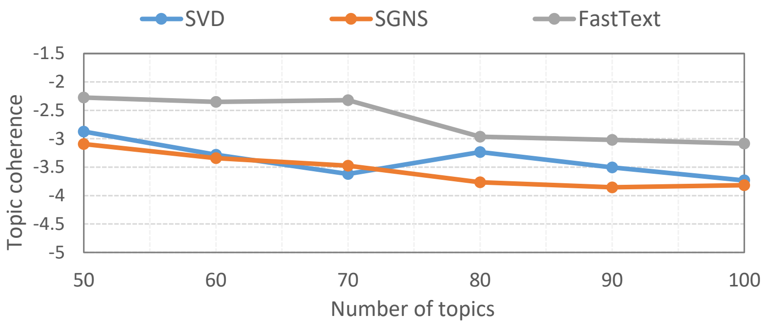


Figure 5.2: Context-GPU with different word/phrase embedding learning methods vs. number of topics.

uration, increasing the maximum iteration number.

- TPM [62]. We implemented the Topical Phrase Model which extracts medical topics using both MedTagger and a hierarchy of Pitman-Yor processes. It outperformed other topical phrase extraction models.

5.5.1 Topic coherence

We assess the generated topics by evaluating their topic coherence. We adopt the topic coherence measure proposed in Mimno et al. (2011), which relies on the co-occurrence statistics collected from the analysed corpus; this allows us to directly measure the coherence of topics with topical phrases (e.g. *short_of_breath*).

In our evaluations, we compute the topic coherence on the top 10 words/phrases using the implementation provided in the Palmetto library[8] [148]. In Figure 5.1, we report the topic coherence computed by averaging the coherence scores resulting for each topic. A peak of coherence is obtained around 60/70 topics for every model, suggesting a potentially suitable number of topics to discriminate the documents. GPU with only local context incorporated outperforms LDA, but its performance is worse compared to TopicVec or TPM. Context-GPU gives superior results over all the baseline models, in particular around 60 and 70 topics. This shows that incorporating the global context is essential to achieve a better topic coherence than only considering the local context. Also, our proposed Context-GPU only involves simple modifications to the GPU. Still, it appears to be more effective than more complex approaches for incorporating word embeddings in topics models (such as TopicVec) or assuming a generative process for documents following the HPYP process (such as TPM).

[8] https://github.com/dice-group/Palmetto

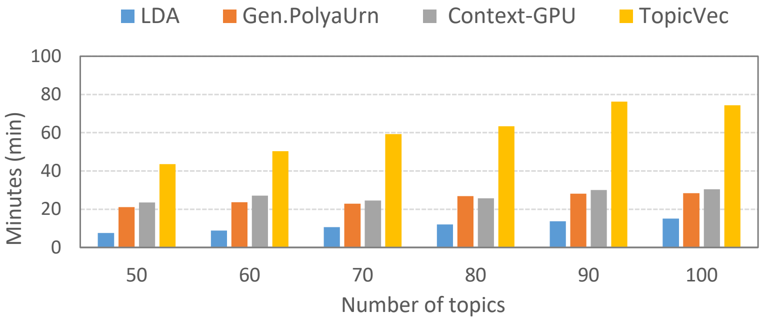


Figure 5.3: Context-GPU: execution time vs. number of topics.

To extract and represent topical phrases from text, we have explored two different methods that learn word/phrase representations: one that infers them directly from the data using the SVD; the second one, training an embedding over the combination of Wikipedia with clinical reports, using SGNS or FastText. In Figure 5.2, we compare these word/phrase embedding methods over our Context-GPU. We can observe that SVD and SGNS perform similarly in most cases and SVD even slightly outperforms SGNS when the topic number is set to 80 or 90. FastText outperforms the other two word/phrase embedding learning methods especially when the topic number is lower than 80. This shows that FastText built on character *n*-grams is more effective in capturing phase sub-structures.

Finally, we compare in Figure 5.3 the execution time required to train the models, excluding the constant time needed by each model to load the embeddings. We did not plot the training time for TPM as it required significantly more time (over 12 hours) compared to all the other models, showing that modelling the phrase generation using HPYP is very expensive. TopicVec is computationally more demanding than the others, while both GPU and Context-GPU have no noticeable differences, requiring three-fold the training time of LDA. Overall, Context-GPU appears to be more effective compared to TopicVec and TPM.

5.5.2 Topic Qualitative Assessment

We report in Table 5.1 some of the topics generated in a 70-topics run. We discuss only the topics of TopicVec and Context-GPU since TopicVec gives coherence scores on par with TPM but requiring significantly less training time. Topics mined with LDA are used as a baseline and are mainly composed of unigrams. This is probably due to the sparseness of the phrases, which tend to naturally occur less frequently than single words, and LDA does not have any compensation strategy to highlight rare yet still relevant words. TopicVec inference instead learns both word and topic embeddings simultaneously. It allows the model to take into account the local context of words, which in turn, alleviates the lack of global statistics for a term. Both the topics of TopicVec and Context-GPU make large use of topical phrases. However, in several topics of Context-GPU, we can distinguish a gradual definition of the analysed themes, which reflect better semantic coherence. For example, in Topic 4, it can be observed a gradual topic refinement under Context-GPU from the general purpose terms, such as *felt* or *insufficiency*, to more characterising words/phrases such as *shortness of breath*, *atrial fibrillation*. In addition, we can observe under the same topic both symptoms and medications, such as *dilated cardiomyopathy* and *Plavix 75 mg*. This qualitative analysis bring further insights on the expressiveness of topics extracted by the Context-GPU, compared to LDA and TopicVec, due to their internal coherence and the enhanced expressiveness of the adopted words/phrases.

It is worth noting, that even though the topic coherence commonly adopted to evaluate the quality of topics [123, 148] provides a reliable measure of the expressiveness and meaningfulness of the generated topics, it currently lacks any control of the medical consistence of the concepts reported within the generated topics, what could be referred as *medical topic coherence*. A future research direction could be focused on this human and automatic evaluation, and the possible approaches to increase this medical topic coherence rather than just the general-purpose one.

5.6 Summary

We have described a novel approach that effectively combines the local and global context of words and phrases. It first detects reliable medical phrases, and then generates topics using our proposed Context-aware Pólya urn model, a statistical model combining the word semantics encoded by the context-based and corpus-based embeddings. In particular, we have employed the LSA and FastText embeddings. The former encoded the topical information with regard to the corpus themes; the latter allowed a fine-grained detection of the word semantics depending on the local context in which they occurred. An experimental comparison with the state-of-the-art methods has shown an improved coherence of final topics and a significantly decreased computational cost.

LDA			
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
chemotherapy	fever	stroke	congestive heart failure
dilantin	urinalysis	carotid	diuresis
oncology	culture	weakness	ejection fraction
xrt	bacteria	speech	approximately
oncologist	white blood cell	stenosis	felt
cycle	polys	confusion	orthopnea
breast cancer	infection	head ct	digoxin
left breast	band	neurology	dyspnea
seizure	fluid	morning	weight
cancer	white	mass	insufficiency
TopicVec			
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
carotid	diuresis	dyspnea on exertion	congestive heart failure
coronary artery	torsemide	ejection fraction	fibrillation
magnesium	cardiomyopathy	pulmonary	ejection fraction
saphenous vein graft	shortness of breath	atrial fibrillation	insufficiency
potassium chloride	torsemide 100 mg	diuresed	calcium
coronary artery bypass grafting	spironolactone 25 mg	congestive heart failure	intubation
mitral insufficiency	diuretic	ischemia	thyroid
mitral regurgitation	aldactone	diabetes mellitus	vascular congestion
potassium	pleural effusion	propafenone	tricuspid regurgitation
substernal	pulmonary edema	volume overloaded	right knee
Contex-GPU			
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
pregnancy	mitral regurgitation	coronary artery disease	congestive heart failure
ultrasound	digoxin	cardiac transplant	pulmonary edema
postpartum hemorrhage	pleural effusion	cardiomyopathy	orthopnea
endometrial biopsy	orthopnea	right coronary artery	nonischemic
total abdominal hysterectomy	dilated cardiomyopathy	pravachol 20 mg	diastolic dysfunction
postpartum	plavix 75 mg	paroxysmal atrial fibrillation	cardiomyopathy
vomiting	shortness of breath	cyclosporine	heart failure
salpingo oophorectomy	dyspnea on exertion	herpes zoster	shortness of breath
physical examination	tachyarrhythmia	fenofibrate tricor	cardiac catheterization
fibroid	pulmonary edema	right coronary artery	atrial fibrillation

Table 5.1: Topics generated by LDA, TopicVec, and Context-GPU in a 70-topics run.

Chapter 6

Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies

Chapter Abstract

In this chapter, we present an approach to incorporate biomedical external knowledge into pre-trained language models by a fine-tuning process focused on pivotal entities, characterizing the domain at hand. This approach is inspired by the word of Krasnashchok et al. 2018, showing how promoting entities in Latent Dirichlet Allocation leads to better and more interpretable topics. Analogously, in this work, we propose a masking strategy to learn and realign the language model representations around the promoted biomedical entities. This is a first step, paving the way to future works for seamless integration of topic representations and language models [25, 130].

6.1 Introduction

Biomedical question-answering (QA) aims to provide users with succinct answers given their queries by analyzing large-scale scientific literature. It enables clinicians, public health officials and end-users to quickly access the rapid flow of specialized knowledge continuously produced. This has led the research community's effort towards developing specialized models and tools for biomedical QA and assessing their performance on benchmark datasets such as BioASQ [164], or the CovidQA collection [161], the first manually curated dataset about COVID-19 related issues.

Producing such data is time-consuming and requires involving domain experts, making it an expensive process. As a result, high-quality biomedical QA datasets are a scarce resource. One recently released CovidQA collection [161], the first manually curated dataset about COVID-19 related issues, provides only 127 question-answer pairs. Even one of the largest available biomedical QA datasets, BioASQ, only contains a few thousand questions.

There have been attempts to fine-tune pre-trained large-scale language models for general-purpose QA tasks [104, 142, 143] and then use them directly for biomedical QA. This is due to their domain adaption ability (i.e. *transfer learning*) which made it possible to leverage the broad knowledge already encoded in them [139]. Furthermore, there has also been increasing interest in developing domain-specific language models, such as BioBERT [91] or RoBERTa-Biomed [60], leveraging the vast medical literature available. While achieving state-of-the-art results on the QA task, these models come with a high computational cost: BioBERT needs ten days on eight GPUs to train [91], making it prohibitive for researchers with no access to massive computing resources, delaying the applications to novel emerging domains.

An alternative approach to incorporating external knowledge into pre-trained language models is to drive the LM to focus on pivotal entities characterizing the domain at hand during the fine-tuning stage. Similar ideas were explored in works by Zhang et al. [201], Sun et al. [157], which proposed the ERNIE model. However, their adaptation strategy was designed to generally improve the LM representations rather than adapting it to a particular domain, requiring additional objective functions and memory. In this work, we aim to enrich existing general-purpose LM models (e.g. BERT [36]) with the knowledge related to key medical concepts. In addition, we want domain-specific LMs (e.g. BioBERT) to re-encode the already acquired information around the medical entities of interest for a particular topic or theme (e.g. literature relating to COVID-19).

Therefore, to facilitate further domain adaptation, we propose a simple yet unexplored approach based on a novel masking strategy to fine-tune an LM. Our approach introduces a *biomedical entity-aware masking* (BEM) strategy encouraging masked language models (MLMs) to learn entity-centric knowledge (§6.3). We first identify a set of entities characterizing the domain at hand using a domain-specific entity recognizer (SciSpacy [126]), and then employ a subset of those entities to drive the masking strategy while fine-tuning. The resulting BEM strategy is applicable to a vast variety of MLMs and does not require additional memory or components in the neural architectures. Experimental results show performance on a par with the state-of-the-art models for biomedical QA tasks (§6.5) on several biomedical QA datasets, with an improved perplexity scores over the domain-specific documents (§6.3). A further qualitative assessment provides an insight into how QA pairs benefit from the proposed approach.

6.2 Related Work

Our work is mainly related to two lines of research on masking strategies for language models and model specialization to particular domains and tasks.

6.2.1 Masking Strategies

In the wake of BERT [36], several new language models have been proposed, adopting different masking strategies.

Several alternative strategies have been recently proposed with different impacts on the final performance. In RoBERTa [104], even though the words are chosen with the same criterion used in BERT, the authors introduced a *dynamic masking* where every time a sequence is fed into the model a difference masking pattern is generated, compared to the *static* approach followed in the original BERT implementation, where each sample was masked once during preprocessing. They also introduced a slightly different training procedure: they removed the next sentence prediction task and showed that performance does not decrease, and at times even improved on some downstream tasks.

In SpanBERT, Joshi et al. (2020) proposed to mask and predict spans rather than tokens. ERNIE [201] instead is focused on masking phrases and named entities to improve the structural knowledge encoded. Although their adaptation strategy was designed to generally improve the LM representations rather than adapting it to a particular domain, it requires additional objective functions and memory.

A rather different strategy is based on the *permutation language modeling* (PLM) task, proposed to train the XLNet model [193]. The aim of the permutation language model is to pre-train the LM without the need to rely on data corruption, i.e. to use a `[MASK]` token which though does not appear during the fine-tuning process. To avoid this discrepancy between the pre-training and fine-tuning phases, it instead minimizes the expected log-likelihood of a sequence with regard to all possible permutations of the sequence order.

6.2.2 Model Specialization

A wide spectrum of specialized language models has been recently developed [8, 23, 91, 196] due to the possibility to process a vast variety of data to fine-tune the models towards different domains and tasks. Among the tasks where contextualized language models has had a remarkable impact, we have question answering [143], machine reading comprehension [159], named entity recognition (NER) [46], sentiment analysis [196] and so on. Some of these techniques have been combined in the T5 model [142], an encoder-decoder transformer-based model sharing the same model, objective and training process across multiple NLP tasks, all reframed as "text-to-text" problems: document summarization, sentiment classification, question answering, machine translation and so on. Although like BERT, T5 uses a denoising approach for the masking strategy, it masks whole spans rather than single tokens.

Particular attention has been devoted to the medical domain, where different corpora and tasks still require different adaptation techniques. BioBERT [91] is a biomedical language model based on the BERT-*Base* variant [36], with additional pre-train on biomedical documents from PubMed and PMC collections, and uses the same training settings adopted in BERT. SciBERT [8] follows the BERT's masking strategy to pre-train the model from scratch using a scientific corpus composed of papers from Semantic Scholar [4]. Out of the 1.14M papers used, more than 80% belong to the biomedical domain. They both showed state-of-the-art result compared to the non-BERT SOTA on several tasks, as Named Entity Recognition, Question Answering and Relation Extraction [8, 91]. BioMed-RoBERTa [60] is instead based on RoBERTa-*Base* [104] using a corpus of 2.27M articles from the Semantic Scholar dataset [4]. While the previous models have been pre-trained on biomedical corpora, our BEM approach is a fine-tuning strategy that relies on the trained model to specialize them with fine-grained updates lead by the external biomedical knowledge encoded into the named entity recognizer.

6.3 BEM: A Biomedical Entity-Aware Masking Strategy

The fundamental principle of a masked language model (MLM) is to generate word representations that can be used to predict the missing tokens of an input text. Chosen a piece of text x from a large unlabeled corpus $\mathcal{X}$, the training is performed by masking tokens in x, so that a pair (x, y) can be generated and used to update the model (e.g. x = "`Coronaviruses [MASK] a group of RNA [MASK]`"; y= ("`are`", "`viruses`")). A good and general-purpose MLM aims at encoding syntactic and

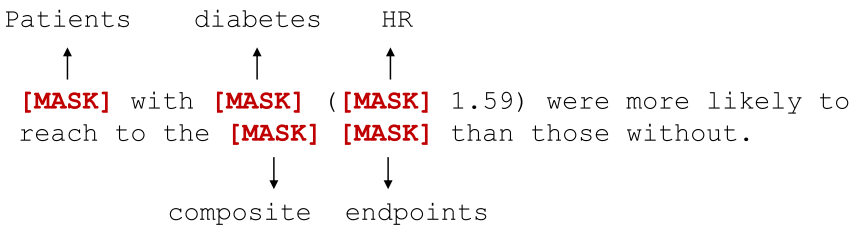

Patients
diabetes
HR
[MASK] with [MASK] ([MASK] 1.59) were more likely to
reach to the [MASK] [MASK] than those without.
composite
endpoints

Figure 6.1: Sentence from the AI2's COVID-19 Open Research Dataset [177] with entities masked based on the SciSpacy NER [126].

semantic information, along with some implicit knowledge (e.g., `viruses`), allowing it to correctly predict the missing token y using the representation of x. While this general principle is adopted in the vast majority of MLMs, the particular way in which the tokens to be masked are chosen can vary considerably. We thus proceed to analyze the random masking strategy adopted in BERT [36] which has inspired most of the existing approaches, and we then introduce the biomedical entity-aware masking strategy used to fine-tune MLMs in the biomedical domain.

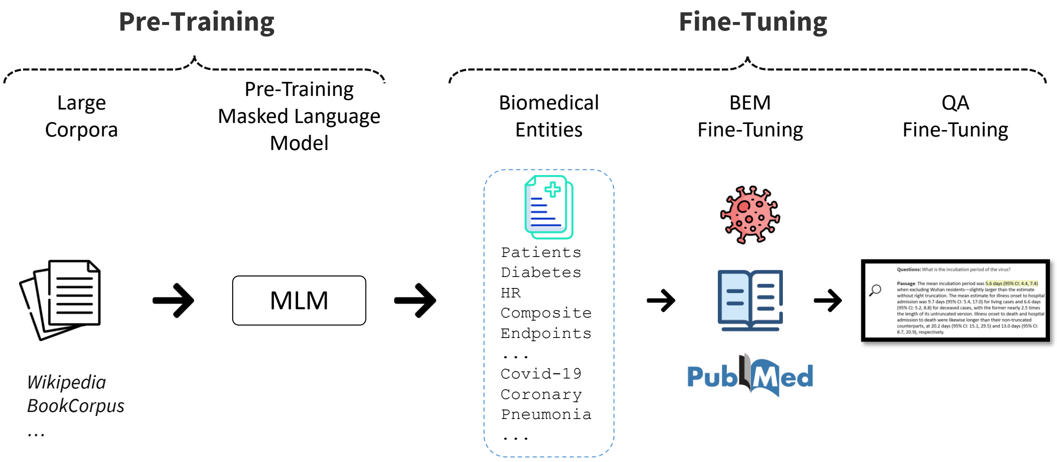
BERT Masking strategy.

The masking strategy adopted in BERT randomly replaces a predefined proportion of words with a special `[MASK]` token and the model is required to predict them. In BERT, 15% of tokens are chosen uniformly at random, 10% of them are swapped into random tokens (thus, resulting in an overall 1.5% of the tokens randomly swapped). This introduces a rather limited amount of noise with the aim of making the predictions more robust to trivial associations between the masked tokens and the context. While another 10% of the selected tokens are kept without modifications, the remaining 80% of them are replaced with the `[MASK]` token.

Biomedical Entity-Aware Masking Strategy

We describe an entity-aware masking strategy which only masks biomedical entities detected by a domain-specific named entity recognizer (SciSpacy[1]). Compared to the random masking strategy described above, which is used to pre-train the masked language models, the introduced entity-aware masking strategy is adopted to boost the fine-tuning process for biomedical documents. In this phase, rather than randomly choosing the tokens to be masked, we inform the model of the relevant tokens to pay attention to, and encourage the model to refine its representations using the new surrounding context. Figure 6.1 shows an example of a sentence from

[1] https://scispacy.apps.allenai.org/



Questions: What is the incubation period of the virus?

Passage: The mean incubation period was 5.6 days (95% CI: 4.4, 7.4) when excluding Wuhan residents—slightly larger than the estimate without right truncation. The mean estimate for illness onset to hospital admission was 9.7 days (95% CI: 5.4, 17.0) for living cases and 6.6 days (95% CI: 5.2, 8.8) for deceased cases, with the former nearly 2.5 times the length of its untruncated version. Illness onset to death and hospital admission to death were likewise longer than their non-truncated counterparts, at 20.2 days (95% CI: 15.1, 29.5) and 13.0 days (95% CI: 8.7, 20.9), respectively.

Figure 6.2: Schematic depiction of the main steps involved in pre-training and fine-tuning a LM for the Biomedical QA task.

the AI2's COVID-19 Open Research Dataset [177] whose biomedical entities are masked based on the SciSpacy NER [126].

Replacing strategy

We decompose the BEM strategy into two steps: (1) *recognition* and (2) *sub-sampling and substitution*. During the *recognition phase*, a set of biomedical entities $\mathcal{E}$ is identified in advance over a training corpus.

Then, at the *sub-sampling and substitution* stage, we first sample a proportion ρ of biomedical entities $\mathcal{E}_f \in \mathcal{E}$. The resulting entity subsets $\mathcal{E}_f$ is thus dynamically computed at batch time, in order to introduce a diverse and flexible spectrum of masked entities during training. For consistency, we use the same tokenizer for the documents d_i in the batch and the entities $e_j \in \mathcal{E}$. Then, we substitute all the k entity mentions $w^k_{e_j}$ in d_i with the special token `[MASK]`, making sure that no consecutive entities are replaced. The substitution takes place at batch time, so that the substitution is a downstream process suitable for a wide typology of MLMs. The main steps involved in specializing a language model for the biomedical QA task are depicted in Figure 6.2.

6.4 Evaluation Design

Biomedical Reading Comprehension

We represent a document as $d_i := (s^i_0, \, . \, . \, , s^i_{j-1})$, a sequence of sentences, in turn defined as $s_j := (w^j_0, \, . \, . \, , w^j_{k-1})$, with w_k a word occurring in s_j. Given a question

q, the task is to retrieve the span $w_s^j, \ldots, w_{s+t}^j$ from a document d_j that can answer the question. We assume the extractive QA setting where the answer span to be extracted lies entirely within one, or more than one document d_i.

In addition, for consistency with the CovidQA dataset and to compare with results in Tang et al. [161], we consider a further and sightly modified setting in which the task consists of retrieving the sentence s_j^i that most likely contains the exact answer. This sentence level QA task mitigates the non-trivial ambiguities intrinsic to the definition of the exact span for an answer, an issue particularly relevant in the medical domain and well-known in the literature [171][2].

Datasets

We assess the performance of the proposed masking strategies on two biomedical datasets: CovidQA and BioASQ.

CovidQA [161] is a manually curated dataset based on the AI2's COVID-19 Open Research Dataset [177]. It consists of 127 question-answer pairs with 27 questions and 85 unique related articles. This dataset is too small for supervised training, but is a valuable resource for zero-shot evaluation to assess the unsupervised and transfer capability of models.

BioASQ [164] is one of the larger biomedical QA datasets available with over 2000 question-answer pairs. To use it within the extractive questions answering framework, we convert the questions into the SQuAD dataset format [143], consisting of question-answer pairs and the corresponding *passages,* medical articles containing the answers or clues with a length varying from a sentence to a paragraph. We used the provided passages including PubMed abstracts and snippets, associating the full abstract when available. When multiple passages are available for a single question, we form additional question-context pairs combined subsequently in a post-processing step to choose the answer with the highest probability, similarly to Yoon et al. [197]. For consistency with the CovidQA dataset, we report our evaluation exclusively on the factoid questions of the BioASQ 7b Phase B1.

Baselines

We use the following unsupervised neural models as baselines: the out-of-the-box BERT [36] and RoBERTa [104], as well as their variants BioBERT [91] and RoBERTa-

[2] Consider, for instance, the following QA pair: *"What is the incubation period of the virus?"*, *"6.4 days (95% 175 CI 5.3 to 7.6)"*, where a model returning just *"6.4 days"* would be considered wrong.

Biomed [60] fine-tuned on medical and scientific corpora.

To highlight the impact of different fine-tuning strategies, we examine several configurations depending on the data and the masking strategy adopted. We experiment using the SQuAD and BioASQ QA training pairs during the fine-tuning stage and denote the models using them with `+SQuAD` or `+BioASQ` respectively. When we fine-tune the models on the corpus consisting of PubMed articles referred to in the BioASQ and AI2's COVID-19 Open Research Dataset, we compare two masking strategies denoted as `+STM` and `+BEM`, where `+STM` indicates the standard masking strategy of the model at hand and `+BEM` is our proposed strategy.

More precisely, `+BEM+BioASQ` indicates a model that is first fine-tuned over the collection of biomedical articles (e.g., with BEM), and then further trained on a set of question-answer pairs (e.g., the BioASQ training set). The first fine-tuning step enhances the model with domain-specific information, while the second stage endows the language model with task-related capabilities. Additionally, we report the T5 [142] performance over CovidQA, which constitutes a current state-of-the-art [161].

Metrics

To facilitate comparisons, we adopt the same evaluation scores used in Tang et al. [161] to assess the models on the CovidQA dataset, i.e. mean reciprocal rank (MRR), precision at rank one (P@1), and recall at rank three (R@3); similarly, for the BioASQ dataset, we use the strict accuracy (SAcc), lenient accuracy (LAcc) and MRR, the BioASQ challenge's official metrics [3].

6.5 Experimental Results and Discussion

We report the results on the QA tasks in Table 6.1.

Among the unsupervised models, BERT achieves slightly better performance than RoBERTa on CovidQA, yet the situation is reversed on BioASQ (rows 1,8). The low precision of the two models (especially on the BioASQ dataset) confirms the difficulties in generalizing to the biomedical domain. Specialized language models such as RoBERTa-Biomed and BioBERT show a significant improvement on the CovidQA dataset, but a rather limited one on BioASQ (rows 15,22), highlighting the importance of having larger medical corpora to assess the model's effectiveness.

[3] `http://participants-area.bioasq.org/Tasks/b/eval_meas_2018/`

#	Model	CovidQA			BioASQ 7b		
		P@1	R@3	MRR	SAcc	LAcc	MRR
1	**BERT**	0.081*	0.117*	0.159*	0.012	0.032	0.027
2	+ SQuAD	0.110	0.131	0.158	0.292	0.343	0.318
3	+ BioASQ	0.125	0.177	0.206	0.226	0.317	0.262
4	+ STM + SQuAD	0.114	0.146	0.173	0.305	0.355	0.336
5	+ STM + BioASQ	0.132	0.195	0.218	0.233	0.325	0.265
6	+ BEM + SQuAD	0.126	0.173	0.191	0.317	0.371	0.349
7	+ BEM + BioASQ	0.145	0.278	0.269	0.241	0.341	0.288
8	**RoBERTa**	0.068	0.115	0.122	0.023	0.041	0.036
9	+ SQuAD	0.098	0.134	0.160	0.353	0.365	0.328
10	+ BioASQ	0.106	0.155	0.178	0.278	0.324	0.294
11	+ STM + SQuAD	0.107	0.148	0.175	0.361	0.388	0.347
12	+ STM + BioASQ	0.112	0.167	0.194	0.282	0.333	0.300
13	+ BEM + SQuAD	0.114	0.162	0.185	0.368	0.391	0.353
14	+ BEM + BioASQ	0.125	0.198	0.236	0.323	0.374	0.325
15	**RoBERTa-Biomed**	0.104	0.163	0.192	0.028	0.044	0.037
16	+ SQuAD	0.111	0.308	0.288	0.376	0.382	0.358
17	+ BioASQ	0.128	0.355	0.315	0.415	0.398	0.376
18	+ STM + SQuAD	0.118	0.314	0.297	0.381	0.390	0.367
19	+ STM + BioASQ	0.136	0.364	0.321	0.423	0.410	0.397
20	+ BEM + SQuAD	0.121	0.331	0.323	0.385	0.397	0.378
21	+ BEM + BioASQ	0.143	0.386	0.347	**0.435**	0.443	0.398
22	**BioBERT**	0.097*	0.142*	0.170*	0.031	0.046	0.039
23	+ SQuAD	0.161*	0.403*	0.336*	0.381	0.445	0.397
24	+ BioASQ	0.166	0.419	0.348	0.410†	0.474†	0.409†
25	+ STM + SQuAD	0.161	0.411	0.339	0.387	0.447	0.401
26	+ STM + BioASQ	0.172	0.432	0.385	0.418	0.482	0.416
27	+ BEM + SQuAD	0.168	0.427	0.354	0.391	0.458	0.423
28	+ BEM + BioASQ	*0.179*	**0.458**	*0.391*	0.421	**0.497**	**0.434**
29	**T5 LM**						
30	+ MS-MARCO	**0.282***	0.404*	**0.415***	—	—	—

Table 6.1: Performance of language models on the CovidQA and BioASQ 7b1 dataset. Values referenced with * come from the Tang et al. (2020) work and with † from Yoon et al. (2020).

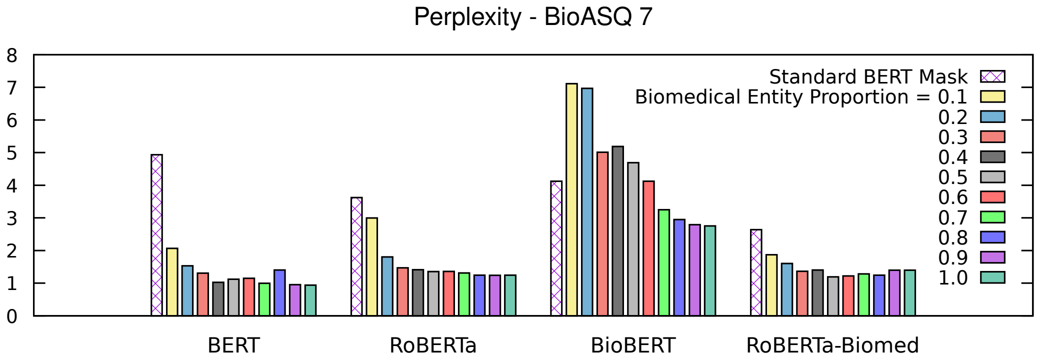


Figure 6.3: Perplexity of MLMs using different masking strategies on the collection of medical articles.

A general boost in performance is shared across models fine-tuned on the QA tasks, with a large benefit from the BioASQ QA. The performance gains obtained by the specialized models (BioBERT and RoBERTa-Biomed) suggest the importance of transferring not only the domain knowledge but also the ability to perform the QA task itself (rows 15,17; 22,24).

We further examined whether the fine-tuning of the QA pairs affects not only the model adaptation to the QA task but it further helps realign the representations for the domain at hand. The report scores point out that the vanilla LMs are the ones gaining the most when using in-domain QA pairs, such as BioASQ, compared to the SQuAD (rows 3,5; 15,17). The advantage tends to be reduced on already specialized LMs (rows 17,19; 24;26).

A further fine-tuning step before the training over the QA pairs has been proven beneficial for all of the models. The BEM masking strategy has significantly amplified the model's generalisability, with an increased adaptation to the biomedical themes shown by the notable improvement in R@3 and MRR; with the R@3 outperforming the state-of-the-art results of T5 fine-tuned on Ms-Marco [6] and proving the effectiveness of the BEM strategy.

In Figure 6.3, we report the LM perplexity obtained when fine-tuning the model with the standard masking strategy versus the BEM strategy with different proportions of medical entities. Vanilla LMs experienced a huge gain with just a small fraction of entities, while already specialized LMs have a lower but still significant improvement. This could be expected as the specialized LMs have already encoded a large domain knowledge with representations that need to be realigned to the new ones.

Finally, we report an excerpt of question-answer pairs from the CovidQA dataset. In particular, Table 6.2 shows questions from the CovidQA related to three

BERT with STM	BERT with BEM
*What is the **OR** for severe infection in COVID-19 patients with hypertension?*	
- There were significant correlations between COVID-19 severity and [..], diabetes [OR=2.67], coronary heart disease [OR=2.85]. **-** Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).	**-** There were significant correlations between COVID-19 severity and [..], diabetes [OR=2.67], coronary heart disease [OR=2.85]. **-** Compared with the non-severe patient, the pooled odds ratio of hypertension, respiratory system disease, cardiovascular disease in severe patients were (OR 2.36, ..), (OR 2.46, ..) and (OR 3.42, ..).
*What is the **HR** for severe infection in COVID-19 patients with hypertension?*	
- - - -	**-** After adjusting for age and smoking status, patients with COPD (HR 2.681), diabetes (HR 1.59), and malignancy (HR 3.50) were more likely to reach to the composite endpoints than those without.
*What is the **RR** for severe infection in COVID-19 patients with hypertension?*	
- - - -	**-** In univariate analyses, factors significantly associated with severe COVID-19 were male sex (14 studies; pooled RR=1.70, ...), hypertension (10 studies 2.74 ...), diabetes (11 studies ...), and CVD (..).

Table 6.2: Examples of questions and retrieved answers using BERT fine-tuned either with its original masking approach or with the biomedical entity-aware masking (BEM) strategy.

statistical indices (i.e. Odds Ratio, Hazard Ratio and Relative Risk) to assess the risk of an event occurring in a group (e.g. infections or death). We notice that even though the indices are mentioned as abbreviations, BERT fine-tuned with the STM is able to retrieve sentences with the exact answer for just one of three questions. By contrast, BERT fine-tuned with the BEM strategy succeeds in retrieving at least one correct sentence for each question. This example suggests the importance of placing the emphasis on the entities, which might be overlooked by LMs during the training process despite being available.

6.6 Summary

We presented BEM, a biomedical entity-aware masking strategy to boost LM adaptation to low-resource biomedical QA. It uses an entity-driven masking strategy to fine-tune LMs and effectively lead them in learning entity-centric knowledge based on the pivotal entities characterizing the domain at hand. Experimental results have shown the benefits of such an approach on several metrics for biomedical QA tasks.

Chapter 7

Conclusion

In this thesis, we proposed to combine topic modelling principles with neural architectures and distributional word representations for text analysis. We introduced several new models and conducted a systematic analysis showing to what extent this could be a suitable and promising combination for capturing high-level semantics and for generating topical features for downstream tasks.

In particular, we showed that by combining topic models and neural architectures it is possible to analyse user sentiments and opinions, or domain-specific concepts. With TDAM (§3) and the Context-GPU (§5), we explicitly took into account the local and global context of words to represent user sentiments and biomedical concepts, respectively, while simultaneously exploiting the large knowledge implicitly encoded in these distributional representations of text. Then, with DIATOM (§4), we combined topic models with neural techniques to generate topics conveying different types of information (i.e. user opinions or plot/factual descriptions). Finally, inspired by the recent advancements in topic models [85], we proposed BEM (§6), a simple yet effective masking strategy to enhance contextualised language models by leveraging the entities of interest in the biomedical domain.

We conclude by summarising the aforementioned contributions with regards to each chapter, highlighting the current limitations and some suggestions for future research directions.

7.1 Summary of Contributions

In the following, we describe the contributions (**C**) of each chapter referencing the corresponding research objectives (**RO**) outlined in CHAPTER 1.

CHAPTER 3 provided a thorough analysis of TDAM, introducing a new GRU cell to capture the topical information in a hierarchical neural architecture with auxiliary memory (**C. 1**). We showed that the resulting architecture is a promising approach to combine global and local context of words in the multi-task learning setting (**RO 1**). The experimental assessment on sentiment classification, topic coherence (**RO 2**) and aspect-based analysis (**RO 5**) demonstrated that the attention mechanism could be a viable solution to simultaneously control and integrate the sentiment (**RO 3**) and topic information.

CHAPTER 4 described DIATOM, a novel neural topic model to generate disentangled topics through the combination of variational autoencoders and adversarial learning (**C. 3**). Employing just the ratings of user reviews, the model was able to generate and separate topics describing user opinions from topics conveying plots or factual descriptions, without being misled by the sentiment expressions within them (e.g., plots with sentiment lexicons). We were able to assess the model capability to consistently disentangle opinion-bearing topics from plot/neutral ones measuring the introduced disentangling rate (**RO 5**). The experimental assessment was based on a newly introduced dataset, namely the MOBO dataset (**C. 4**), pairing movie and book reviews with their plots, along with human-annotated sentences used for topic labelling. The results on the novel dataset showed an overall improvement of the quality of topics (**RO 2**) using several metrics, and better sentiment classification compared to other supervised topic models (**RO 5**).

CHAPTER 5 presented the Context-aware Pólya urn model (Context-GPU). It expanded the Generalised Pólya urn model leveraging the combination of the LSI and the FastText embeddings (**RO 1**) to drive the weighting scheme for topic generation (**C. 2**). This implied a combination of word representations based on their local and global contexts using windows-based and corpus-based embeddings, respectively. As a result, it mitigated the impact of the topic shifting issue on the final coherence of the generated topics (**RO 2**). Additionally, the pretrained FastText embedding had the advantage of implicitly incorporating external knowledge used to analyses the technical documents (**RO 3**), while its character-oriented design was proven suitable to identify medical phrases, which often share morphological similarities, such as prefixes or suffixes (**RO 4**). The experimental assessment demonstrated a cost-effective process, generating more accurate and expressive topics.

CHAPTER 6 introduced the biomedical entity-aware masking strategy (BEM), which inspired by the promotion of entities in topic models [85], leveraged the biomedical entities (**RO 4**) to efficiently fine-tune masked language models (**C. 2**). Results on both the BioASQ dataset [164] and the CovidQA collection [161] showed the BEM strategy outperforming the existing methodologies to fine-tune MLMs (**RO 5**).

7.2 Current Limitations and Future Directions

We conclude by highlighting the limitations and still open research questions for each work presented, along with possible future research directions.

Effective pretraining of hierarchical neural networks. Although TDAM (§3) is able to provide contextualised features analogously to the recently developed transformer architectures, its sentiment classification performance is still not on par with the state-of-the-art language models [36]. One of the main reasons is that the pretraining of the hierarchical architecture has a much more limited impact on the final performance [47] than the benefit gained from most of the transformer-based architectures. To mitigate this issue, a possible extension of TDAM and its hierarchical architecture would consist of pairing a neural topic model with the GRU unit to inject further topical information directly into the recurrent analysis of the corpus. Additionally, an intermediate layer in between the word and sentence layers could be added, defining a *discourse-level* layer [44, 189]. This discourse-level layer would process the so-called *elementary discourse units* (EDUs), which are the minimal text units of a discourse tree as defined in the Rhetorical Structure Theory [111], and can be recognised with the appropriate text-level discourse parsers. As a result, such a discourse layer would add an intermediate level of abstractions for a more fine-grained resolution in detecting multi-word semantic units within sentences.

Topic with consistent polarities and different priors. Even though the adversarial mechanism implemented in DIATOM (§4) is rather effective in disentangling opinion and neutral/plot topics, at times, the opinion topics could exhibit terms of mixed polarities. An additional adversarial mechanism can be a viable solution at the cost of increasing the model's complexity. Also, in our current model, the latent plot topics z_a extracted from reviews are encouraged to have similar discriminative power as the latent topic z_d learned from plots directly for predicting the plots. Instead, it would also be possible to impose a Gaussian prior centred on z_d for the latent plot topics in reviews rather than using the Gaussian prior of zero mean and unit variance. Another approach would consist of replacing the plot classifier with a discriminator, as typically used in GAN training, where the learned plot topics from different sources (reviews and plots) would compete with each other to confuse the discriminator. Considering the augmented set of neutral/plot topics generated by DIATOM, further analysis could leverage them to show the impact of different types of neutral topics on the user opinions. Finally, while we focus on separating opinion topics from plot or neutral ones in movie and book reviews, our proposed framework can be applicable in other scenarios. For example, in veracity classification of Twitter rumours, we want to disentangle latent factors which are indicative of the veracity of tweets from those which are event-related. Our proposed framework provides a potential solution to it.

Language models for medical abbreviations and medical topic coherence. The Context-GPU model combines the windows-based and the corpus-based embeddings to generated topics composed of topical phrases. However, when analysing clinical notes, many topical phrases results from the expansion of medical abbreviations. The clinical annotation tool adopted so far, i.e. *MedTagger*, does not provide any mechanism to detect the best way to expand the abbreviations based on their surrounding context. A promising approach would consist of adopting contextualised language models to perform a context-aware expansion [82] at the cost of a slightly more complex pipeline. General-purpose contextualised language models [36, 104] or domain-specific biomedical LMs [3, 91] could be used to substitute FastText [79]. These language models are not only strictly context-dependent, but they also rely on the WordPiece tokeniser [185] that combines sub-words by following a precomputed likelihood, making such a tokeniser particularly suitable to deal with medical phrases and jargon. Additionally, another approach to induce more informative topics could leverage the named entity recogniser, i.e. SciSpacy [126], presented for the BEM strategy (§6), to identify and then promote the medical entities within topics, analogously to what suggested in Krasnashchok et al. 2018. Although the current evaluation is based on the analysis of clinical notes, the quality of the concepts expressed by the generated topics is solely based on the general-purpose topic coherence [123, 148]. Although this provides some guarantees about the meaningfulness of the topics, it does not assess whether they are consistent from a medical point of view. A first viable research direction could be focused on the human evaluation of such topics by medical practitioners, and on how to consequently adjust and modify the models to increase the *medical topic coherence*. In turn, this would pave the way to automatic approaches checking the medical consistency of the topics by matching the expert judgement.

Filtering the biomedical entities of interest. The introduced biomedical entity-aware masking strategy relies on SciSpacy to identify the biomedical entities in text. However, so far there is no mechanism leading the model to detect the most relevant entities. A straightforward extension would consist of substituting the uniform random sampling strategy that during each iteration chooses the biomedical entities to mask with two possible alternative criteria reducing the original pool of biomedical entities. In particular, a fine-grained approach could be based on the *Unified Medical Language System* (UMLS) [18] meta-information associated with each entity. One viable solution could be leveraging the *Type Unique Identifier*s (TUI) to filter and keep only those entities belonging to specific type groups. Alternatively, we could use the *Concept Unique Identifier*s (CUI), which group together entities referring to similar concepts. Consequently, we would have fewer but more relevant entities to be masked, and thus a more efficient and domain-driven process.

Bibliography

[1] Asad Abdi, Siti Mariyam Shamsuddin, Shafaatunnur Hasan, and Jalil Piran. Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, 56(4):1245 – 1259, 2019. ISSN 0306-4573.

[2] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19 (1):1947–1980, 2018.

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (NAACL)*, pages 72–78, June 2019.

[4] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, June 2018.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA*, 2015.

[6] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated machine reading comprehension dataset, 2018.

[7] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting

semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, November 2019.

[9] Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *CoRR*, abs/1703.07718, 2017.

[10] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[11] Yoshua Bengio. The consciousness prior. *CoRR*, abs/1709.08568, 2017.

[12] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3: 1137–1155, 2003.

[13] Sonia Bergamaschi and Laura Po. Comparing lda and lsa topic models for content-based movie recommendation systems. In *Web Information Systems and Technologies: 10th International Conference, WEBIST 2014, Barcelona, Spain, April 3-5, 2014*, 2015.

[14] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, April 2012. ISSN 0001-0782.

[15] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, (ICML), pages 113–120, New York, NY, USA, 2006. ACM.

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003. ISSN 1532-4435.

[17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, Apr 2017.

[18] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research Journal*, 32, Database issue, 2004.

[19] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[20] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, August 2016.

[21] John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, Aug 2007.

[22] Dallas Card, Chenhao Tan, and Noah A. Smith. Neural Models for Documents with Metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 2031–2040, Melbourne, Australia, July 2018.

[23] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2898–2904, November 2020.

[24] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296, 2009.

[25] Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. TopicBERT for energy efficient document classification. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1682–1690, November 2020.

[26] Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 1650–1659, Austin, Texas, USA., 2016.

[27] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information

maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29*, pages 2172–2180, 2016.

[28] Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 1226–1240, New Orleans, Louisiana, June 2018.

[29] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014.

[30] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, ACL, pages 76–83, Stroudsburg, PA, USA, 1989. Association for Computational Linguistics.

[31] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, (ICML), pages 160–167, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4.

[32] Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. Recurrent batch normalization. In *Proceedings of the 2017 International Conference for Learning Representations, ICLR 2017*, Touloun, France, 2017.

[33] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP*, 2015.

[34] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 41(6):391–407, 1990.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of The Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL)*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[37] Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. Topicrnn: A recurrent neural network with long-range semantic dependency. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[38] Adji B. Dieng, Chong Wang, Jianfeng Gao, and John W. Paisley. Topicrnn: A recurrent neural network with long-range semantic dependency. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France.*, 2017.

[39] Adji B. Dieng, Chong Wang, Jianfeng Gao, and John William Paisley. TopicRNN: A recurrent neural network with long-range semantic dependency. In *Proceedings of the 2017 International Conference for Learning Representations, ICLR*, Touloun, France, 2017.

[40] Ran Ding, Ramesh Nallapati, and Bing Xiang. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 830–836, Brussels, Belgium, 2018. Association for Computational Linguistics.

[41] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, Workshop Track Proceedings, ICLR*, 2015.

[42] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *In Proceedings of Machine Learning Research (PMLR)*, pages 2525–2534, 16–18 Apr 2019.

[43] Geli Fei, Zhiyuan Chen, and Bing Liu. Review topic discovery with phrases using the pólya urn model. In *COLING 2014, 25th International Conference on Computational Linguistics, Dublin, Ireland*, pages 667–676, 2014.

[44] Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich

linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 60–68, Jeju Island, Korea, 2012.

[45] J. R. Firth. A synopsis of linguistic theory 1930-55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, volume 1952-59, pages 1–32, 1957.

[46] Jinlan Fu, Pengfei Liu, and Graham Neubig. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, November 2020.

[47] Shang Gao, M. T. Young, John X. Qiu, Hong-Jun Yoon, J. B. Christian, P. Fearn, G. Tourassi, and Arvind Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association : JAMIA*, 25:321 – 330, 2018.

[48] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Autoencoding total correlation explanation. *Proceedings of Machine Learning Research*, 89:1157–1166, Apr 2019.

[49] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[50] S. Gershman and Noah D. Goodman. Amortized inference in probabilistic reasoning. *Cognitive Science*, 36, 2014.

[51] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics, AISTATS*, 2010.

[52] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://skylion007.github.io/openwebtextcorpus, 2019.

[53] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014.

[54] Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. Parallel Gibbs Sampling: From colored fields to thin junction trees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 324–332, 2011.

[55] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27, NIPS*, pages 2672–2680, 2014.

[56] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 1462–1471, 2015.

[57] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[58] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation, 2002.

[59] Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. Neural topic model with reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3478–3483, Hong Kong, China, November 2019.

[60] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL)*, 2020.

[61] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[62] Yulan He. Extracting topical phrases from clinical documents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2957–2963, United States, February 2016. AAAI.

[63] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28, NIPS*, pages 1693–1701, Montreal, Canada, 2015.

[64] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[65] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018.

[66] Geoffrey E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504 – 507, 2006.

[67] Tai Hoang, Huy Le, and Tho Quan. Towards autoencoding variational inference for aspect-based opinion summary. *Applied Artificial Intelligence (API)*, 33: 796–816, 2019.

[68] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667.

[69] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, volume 23 of *NIPS*, pages 856–864, 2010.

[70] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems 31, NIPS*, pages 517–526, 2018.

[71] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 95:423–469, 2014.

[72] Furong Huang and Animashree Anandkumar. Unsupervised learning of word-sequence representations from scratch via convolutional tensor decomposition. *arXiv preprint arXiv:1606.03153*, 2016.

[73] Mingmin Jin, Xin Luo, Huiling Zhu, and Hankz Hankui Zhuo. Combining deep learning and topic modeling for review understanding in context-aware recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 1605–1614, New Orleans, Louisiana, 2018.

[74] Yohan Jo and Alice H Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining, WSDM 2011*, pages 815–824, 2011.

[75] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (ACL)*, Florence, Italy, July 2019.

[76] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[77] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An Introduction to Variational Methods for Graphical Models*, volume 37. Machine Learning, USA, 1999.

[78] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[79] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[80] Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA).

[81] Zenun Kastrati, Ali Shariq Imran, and Sule Yildirim Yayilgan. The impact of deep learning on document classification using semantically rich representations. *Information Processing & Management*, 56(5):1618 – 1632, 2019. ISSN 0306-4573.

[82] Juyong Kim, Linyuan Gong, Justin Khim, Jeremy C. Weiss, and Pradeep Ravikumar. Improved clinical abbreviation expansion via non-sense-based approaches. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 161–178. PMLR, 11 Dec 2020.

[83] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations, ICLR*, 2015.

[84] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations ICLR*, 2014.

[85] Katsiaryna Krasnashchok and Salim Jouili. Improving topic quality by promoting named entities in topic modeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL)*, pages 247–253, Melbourne, Australia, July 2018.

[86] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *Proceedings of the 5th International Conference on Learning Representations, ICLR*, 2017.

[87] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Apr 2014.

[88] Jey Han Lau, Timothy Baldwin, and Trevor Cohn. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[89] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1188–1196, 2014.

[90] Jieh-Sheng Lee and Jieh Hsiang. PatentBERT: Patent classification with fine-tuning a pre-trained BERT model. *World Patent Information*, 61(101965), 2020.

[91] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 2019*, 09 2019. ISSN 1367-4803.

[92] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics (TACL)*, 3:211–225, 2015.

[93] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 165–174, 2016.

[94] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 666–675, Berlin, Germany, August 2016. Association for Computational Linguistics.

[95] Y. Li, P. Nair, Xing Han Lu, Z. Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Y. Miao, Weiqi Liu, T. Ordog, J. Biernacka, E. Ryu, J. Olson, Mark A Frye, A. Liu, Liming Guo, A. Marelli, Yuri Ahuja, J. Davila-Velderrain, and Manolis Kellis. Inferring multimodal latent topics from electronic health records. *Nature Communications*, 11, 2020.

[96] Percy Liang and Christopher Potts. Bringing machine learning and compositional semantics together. *Annual Review of Linguistics*, 1(1):355–376, 2015.

[97] C. Lin, Y. He, R. Everson, and S. Ruger. Weakly supervised joint sentiment-topic detection from text, 2012. *IEEE Transactions on Knowledge and Data Engineering*, 5(6):1134–1145, 2012.

[98] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM, pages 375–384, 2009.

[99] Alexander H. Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31, (NIPS)*, pages 2590–2599, 2018.

[100] Hongfang Liu, Stephen T Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J Haug, Stanley M Huff, and Christopher G Chute. Towards a semantic lexicon for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, page 568, 2012.

[101] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Deep multi-task learning with shared memory for text classification. In *Proceedings of the 2016 Conference on*

Empirical Methods in Natural Language Processing, (EMNLP), pages 118–127, Austin, Texas, USA, 2016.

[102] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 1–10, Vancouver, Canada, 2017.

[103] Yang Liu and Mirella Lapata. Learning structured text representations. *Transactions of the Association for Computational Linguistics (TACL)*, 6:63–75, 2018.

[104] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach, 2019.

[105] Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, and Yulan He. CHIME: Cross-passage hierarchical memory network for generative review question answering. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 2547–2560, Barcelona, Spain (Online), December 2020.

[106] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (EMNLP)*, pages 1412–1421, Lisbon, Portugal, 2015.

[107] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJACI 2017*, pages 4068–4074, Melbourne, Australia, 2017.

[108] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011.

[109] Hosam Mahmoud. *Polya Urn Models*. Chapman & Hall/CRC, 1 edition, 2008. ISBN 1420059831, 9781420059830.

[110] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. In *4th International Conference on Learning Representations, ICLR*, 2016.

[111] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[112] James Martens. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML, pages 735–742, 2010.

[113] Tomonari Masada and Atsuhiro Takasu. Adversarial learning for topic models. In Guojun Gan, Bohan Li, Xue Li, and Shuliang Wang, editors, *Advanced Data Mining and Applications, ADMA*, 2018.

[114] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 5040–5048, 2016.

[115] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2015*, Santiago, Chile, 2015.

[116] Jon D. Mcauliffe and David M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems 20, (NIPS)*, pages 121–128, Vancouver, Canada, 2008.

[117] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[118] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*, 2016.

[119] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. *In Proceedings of The 34th International Conference on Machine Learning, (ICML)*, 2017.

[120] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR*, 2013.

[121] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS, pages 3111–3119, 2013.

[122] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2013.

[123] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 262–272, 2011.

[124] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, ICML, 2014.

[125] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics.

[126] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy, August 2019.

[127] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, (TACL)*, 3:299–313, 2015.

[128] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, October 2020.

[129] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd*

Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain, July 2004.

[130] Nicole Peinelt, Dong Nguyen, and Maria Liakata. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7047–7055, July 2020.

[131] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, pages 5102–5112, 2019.

[132] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task (ACL)*, pages 58–65, Florence, Italy, August 2019.

[133] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2014.

[134] Gabriele Pergola, Yulan He, and David Lowe. Topical phrase extraction from clinical reports by incorporating both local and global context. In *The 2nd AAAI Workshop on Health Intelligence (AAAI18)*, pages 499–506, June 2018.

[135] Gabriele Pergola, Lin Gui, and Yulan He. TDAM: A topic-dependent attention model for sentiment analysis. *Information Processing & Management*, 56(6): 102084, 2019. ISSN 0306-4573.

[136] Gabriele Pergola, Lin Gui, and Yulan He. A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Online, June 2021. Association for Computational Linguistics.

[137] Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. Boosting low-resource biomedical QA via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Main Volume*, pages 1977–1985, Online, April 2021. Association for Computational Linguistics.

[138] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[139] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, November 2019.

[140] Domenico M. Pisanelli, Aldo Gangemi, Massimo Battaglia, and Carola Catenacci. Coping with medical polysemy in the semantic web: the role of ontologies. *Studies in health technology and informatics*, 107 Pt 1:416–9, 2004.

[141] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[142] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

[143] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP16)*, pages 2383–2392, November 2016.

[144] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, (EMNLP-IJCNLP)*, 2019.

[145] Navid Rekabsaz, Mihai Lupu, Allan Hanbury, and Hamed Zamani. Word embedding causes topic shifting; exploit global context! In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 1105–1108, 2017. ISBN 978-1-4503-5022-8.

[146] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In

Proceedings of the 31st International Conference on Machine Learning (ICML), volume 32, pages 1278–1286, 2014.

[147] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, ICML, pages 833–840, 2011.

[148] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM*, pages 399–408, Shanghai, China, 2015.

[149] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 487–494, 2004.

[150] David E. Rumelhart, Richard Durbin, Richard Golden, and Yves Chauvin. *Backpropagation: The Basic Theory*, page 1–34. Lawrence Erlbaum Associates, Inc., USA, 1995.

[151] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226, Lille, France, 07–09 Jul 2015. PMLR.

[152] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the International Conference for Learning Representations, ICLR 2014*, Banff, Canada, 2014.

[153] Alex Smola and Shravan M. Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3:703 – 710, 2010.

[154] David Sontag and Dan Roy. Complexity of inference in latent dirichlet allocation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 1008–1016, 2011.

[155] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[156] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3664–3674, Brussels, Belgium, 2018.

[157] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2020.

[158] Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter. Code and named entity recognition in StackOverflow. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4913–4926, July 2020.

[159] Alon Talmor and Jonathan Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4911–4921, Florence, Italy, July 2019.

[160] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1422–1432, Lisbon, Portugal, 2015.

[161] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. Rapidly bootstrapping a question answering dataset for COVID-19, 2020.

[162] Valentin Thomas, Emmanuel Bengio, William Fedus, Jules Pondard, Philippe Beaudoin, Hugo Larochelle, Joelle Pineau, Doina Precup, and Yoshua Bengio. Disentangling the independently controllable factors of variation by interacting with the world. In *Learning Disentangled Representations, Workshop NIPS 2017.*, pages 2590–2599, 2017.

[163] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning, 2019.

[164] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of

the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1), 2015.

[165] Ozlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523, 2010.

[166] Laurens Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research (JMLR)*, 15(1), January 2014.

[167] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[168] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

[169] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 5998–6008, 2017.

[170] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML, page 1096–1103, 2008.

[171] Ellen M. Voorhees and Dawn M. Tice. The trec-8 question answering track evaluation. In *In Text Retrieval Conference TREC-8*, pages 83–105, 1999.

[172] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems (NIPS09)*, volume 22, pages 1973–1981, 2009.

[173] Hanna M. Wallach. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, (ICML), pages 977–984, 2006. ISBN 1-59593-383-2.

[174] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings*

of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 2605–2610, 2019.

[175] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19, 2019.

[176] Chong Wang, David M. Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, Florida, USA*, pages 1903–1910, 2009.

[177] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[178] Rui Wang, Deyu Zhou, and Yulan He. ATM: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098, 2019.

[179] Weichao Wang, Shi Feng, Wei Gao, Daling Wang, and Yifei Zhang. Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Brussels, Belgium, 2018.

[180] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–198, Vancouver, Canada, 2017.

[181] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.

[182] Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. Disentangling topic models: A cross-cultural analysis of personal values through words. In *Proceedings of the First Workshop on NLP and Computational Social Science*, November 2016.

[183] J. Wolfowitz. On the stochastic approximation method of robbins and monro. *Annals of Mathematical Statistics*, 23(3):457–461, 09 1952.

[184] Fangzhao Wu and Yongfeng Huang. Personalized microblog sentiment classification via multi-task learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI*, pages 3059–3065, Phoenix, Arizona, 2016.

[185] Y. Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, G. Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. S. Corrado, Macduff Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144, 2016.

[186] Patrick Xia, Shijie Wu, and Benjamin Van Durme. Which *BERT? A survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, November 2020.

[187] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning, (ICML)*, pages 2048–2057, Lille, France, 2015.

[188] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. Variational autoencoder for semi-supervised text classification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.

[189] Hanqi Yan, Lin Gui, Gabriele Pergola, and Yulan He. Position bias mitigation: A knowledge-aware graph model for emotion cause extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online, August 2021. Association for Computational Linguistics.

[190] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW, page 1445–1456, 2013.

[191] Chao Yang, Hefeng Zhang, Bin Jiang, and Keqin Li. Aspect-based sentiment analysis with alternating coattention networks. *Information Processing & Management*, 56(3):463 – 478, 2019. ISSN 0306-4573.

[192] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020.

[193] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 5753–5763, 2019.

[194] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL*, pages 1480–1489, San Diego, California, 2016.

[195] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pages 3881–3890, 2017.

[196] Da Yin, Tao Meng, and Kai-Wei Chang. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, July 2020.

[197] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases*, 2020.

[198] Mo Yu and Mark Dredze. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics, (TACL)*, 3: 227–242, 2015.

[199] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, April 2004. ISSN 1046-8188.

[200] Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. Multi-task label embedding for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4545–4553, Brussels, Belgium, 2018.

[201] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 1441–1451, Florence, Italy, July 2019.

[202] Renjie Zheng, Junkun Chen, and Xipeng Qiu. Same representation, different attentions: Shareable sentence representation learning from multiple tasks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 4616–4622, Stockholm, Sweden, 2018.

[203] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. Jec-qa: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9701–9708, 2020.

[204] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online, August 2021. Association for Computational Linguistics.

[205] Qile Zhu, Zheng Feng, and Xiaolin Li. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4663–4672, October-November 2018.

[206] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.