UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Paulo Viana Bicalho

A General Framework to Expand Short Text for Topic Modeling

Belo Horizonte
2017

Paulo Viana Bicalho

A General Framework to Expand Short Text for Topic Modeling

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Gisele Lobo Pappa
Co-Advisor: Anisio Mendes Lacerda

Belo Horizonte
2017

Paulo Viana Bicalho

Um Arcabouço para Expansão de Textos Curtos em Modelagem de Tópicos

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientadora: Gisele Lobo Pappa
Coorientador: Anisio Mendes Lacerda

Belo Horizonte
2017

Bicalho, Paulo Viana

B583g A general framework to expand short text for topic modeling [manuscrito] / Paulo Viana Bicalho. — 2017.
xviii, 62 f. : il. ; 29cm

Orientador: Gisele Lobo Pappa.
Coorientador: Anisio Mendes Lacerda.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 57-62

1. Computação – Teses. 2. Sistemas de recuperação da informação - Teses. 3. Modelagem de informações – Teses. 4.– Mineração de dados (Computação) – Teses. I. Pappa, Gisele Lobo. II. Lacerda, Anisio Mendes. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da computação. IV. Título.

CDU 519.6*73(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz - CRB 6ª Região nº 819.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A general framework to expand short text for topic modeling

PAULO VIANA BICALHO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. GISELE LOBO PAPPA - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. ANISIO MENDES LACERDA - Coorientador
Departamento de Computação - CEFET

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 17 de janeiro de 2017.

Resumo

Textos curtos são frequentemente encontrados na Web, e incluem mensagens publicadas em mídias sociais, mensagens de status, comentários de blogs, etc. Descobrir os tópicos ou assuntos presentes neste tipo de mensagens é crucial para uma ampla gama de aplicações, como análise de contexto e caracterização de usuários. No entanto, extrair tópicos de textos curtos é desafiador. Isto porque existe uma dependência dos métodos convencionais, como Latent Dirichlet Allocation (LDA), da co-ocorrência de palavras, que em textos curtos são raras.

Dados os desafios dessa tarefa, esta dissertação propõe um arcabouço para modelagem de tópicos em textos curtos que trabalha expandindo os documentos originais, transformando-os em pseudo-documentos maiores e com mais informações. No arcabouço proposto, os documentos são decompostos em componentes (palavras, bigramas ou n-gramas) definidos sobre um espaço métrico, capaz de fornecer informações sobre a similaridade entre esses componentes. Apresentamos então duas especializações do nosso arcabouço que, apesar de simples, são eficazes e eficientes para a geração de pseudo-documentos a partir dos documentos de texto curto originais.

Enquanto o primeiro método, CoFE (*Co-Frequency Expansion*), considera a co-ocorrência de palavras para definir o espaço métrico, o segundo, DREx (*Distributed Representation-based Expansion*), baseia-se em representações vetoriais de palavras. Os pseudo-documentos gerados podem ser dados como entrada para qualquer algoritmo de modelagem de tópicos, o que torna nossa abordagem ainda mais genérica e flexível.

Comparamos os resultados das estratégias propostas com cinco algoritmos estado-da-arte que seguem duas estretégias: geram peduso-documentos ou modificam os métodos convencionais de extração de tópicos. Os métodos foram avaliados em sete conjuntos de dados usando a métrica de qualidade de tópico *Normalized Pointwise Mutual Information (NPMI)* e também no contexto de classificação de documentos. Resultados experimentais mostram que o DREx com a representação vetorial gerada pelo método Glove supera os métodos existentes, obtendo valores mais elevados de NPMI e melhores valores de macro-F1, com ganhos de até 15% neste último.

Palavras-chave: Modelagem de Tópicos, Expansão de Documentos, Textos Curtos, Representação Vetorial de Palavras.

Abstract

Short texts are everywhere in the Web, including messages posted in social media, status messages and blog comments, and uncovering the topics of this type of messages is crucial to a wide range of applications, e.g. context analysis and user characterization. Extracting topics from short text is challenging because of the dependence of conventional methods, such as Latent Dirichlet Allocation, in word co-occurrence, which in short text are rare and make these methods suffer from severe data sparsity.

In order to address the challenges imposed by this task, this dissertation proposes a general framework for topic modeling of short text by creating larger pseudo-document representations from the original documents. In the proposed framework, document components (e.g. words, bigrams or n-grams) are defined over a metric space, which provides information about the similarity between them. We present two simple, effective and efficient methods that specialize our general framework to create larger pseudo-documents. While the first method, CoFE (Co-Frequency Expansion), considers word co-occurrence to define the metric space, the second, DREx (Distributed Representation-based Expansion), relies on distributed word vector representations. The pseudo-documents generated can be given as input to any topic modeling algorithm.

Methods were evaluated in seven datasets using the normalized pointwise mutual information (NPMI) topic quality metric and also within the context of a text classification task. They were compared with five state-of-the-art methods for extracting topics by generating pseudo-documents or modifying current topic modeling methods for short text. Results show that DREx using the word embeddings generated by Glove significantly outperforms the baseline methods in terms of normalized pointwise mutual information and macro F1 score, with gains up to 15% in the latter.

Keywords: Topic Modeling, Document Expansion, Short Text, Word Vector Representations

Contents

Chapter 1

Introduction

The popularization of the Web and the constant production of text information has further motivated the investigation of methods capable of extracting richer information from these texts, which goes beyond syntactic relations. There is a large community of researchers that look at the semantics of the text from the point of view of ontologies and other human-readable dictionaries [Stumme et al., 2006]. However, these methods demand a set of external resources that may be language and context-dependent, and in many contexts may not be easily applied.

Topic identification methods, in contrast, work with a “loose" definition of semantics, are language independent and do not require any other external source to work. They are nowadays among the most explored tools to extract information from textual data. Topic modeling methods were conceived to find semantically meaningful topics from a document corpus, and they assume that there are hidden variables (topics) that explain the similarities between observable variables (documents). These techniques are usually based on probabilistic or non-probabilistic methods.

Probabilistic methods assume that the data was generated by a generative model that includes the hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables that allow us to infer the existing topics. Non-probabilistic methods, in contrast, are usually based on matrix factorization techniques, where the matrix of terms per document – which represents the dataset – is projected into a k-dimensional space where each dimension is a topic.

This work focuses on probabilistic methods for topic identification, where the main representative topic modeling method is Latent Dirichlet Allocation (LDA) [Blei et al., 2003], as they are considered state-of-the-art in most scenarios. LDA has been applied to many different contexts to discover topics, including text, image and biological data [Hörster et al., 2007; Pinoli et al., 2014]. However, as pointed out by Tang et al. [2014], there are scenarios where LDA models are not “data-friendly”. These scenarios include those where: (i) only a few documents are available, (ii) documents contain too many topics or (iii) documents are too short.

Since a big part of Web data is becoming shorter and shorter, e.g. messages posted in social media, status messages, blog comments, questions in Q & A websites,

advertisement texts, image captions, etc, uncovering the topics of this type of messages is crucial to a wide range of applications, including context analysis [Zhao et al., 2011; Hong Davison, 2010], user [Weng et al., 2010; Pal et al., 2016], real-time topic detection [Lin et al., 2010] , etc. For this reason, this dissertation is interested in one of the challenges in topic identification: how to uncover the topics in short text documents.

Extracting topics from short text is difficult because of the dependence of topic modeling methods in word co-occurrence, which in short text are rare and make conventional algorithms suffer from severe data sparsity [Hong Davison, 2010]. Two different approaches have been proposed to address the problem of topics extraction from short text: (i) Methods that propose new probabilistic topic models or modify the existent ones in order to deal with the high sparsity and the lack of word co-occurrence of the short text [Zhao et al., 2011; Nguyen et al., 2015; Jin et al., 2011]; (ii) Methods that create larger pseudo-documents from the original short text documents, and then apply traditional topic modeling methods to these pseudo-documents [Hong Davison, 2010; Mehrotra et al., 2013]. The latter has the main advantage of being simpler and method-independent, since it only transforms the input data.

One of the main problem with current methods that generate larger pseudo-documents is that, most of the time, they use information about the data source or the context where they are being applied, and cannot be easily generalized for other contexts. For example, in Mehrotra et al. [2013] the authors propose different tweet pooling schemes to generate pseudo-documents from tweets. They found out that grouping tweets using a common hashtag to generate larger pseudo-documents is the most effective approach to generate larger pseudo-documents. However, depending on the number of different hashtags present in the data, this approach may reduce the number of documents significantly, generating another type of problem to LDA: dealing with few documents [Tang et al., 2014]. Furthermore, in scenarios where there is not an available common element to merge the documents (e.g. hashtags), this method cannot be applied.

To overcome the aforementioned limitations, this dissertation proposes a general framework for generating pseudo-documents for topic modeling in short text that is context-independent, allows to specify the maximum desired size of the documents and creates pseudo-documents that can be given as input to any topic modeling method. These features make our framework flexible, since documents can be expanded according to user needs and the framework can be used with any textual data.

The foundation of this framework is to expand the short documents by appending similar components (e.g. words, bi-grams or n-grams) that are relevant to the documents subjects. These components are defined over a metric space, which provides information about the similarity between pairs of components and allow us to calculate the similarity between a whole document and a set of these components. We present two new methods that are specializations of this general framework to expand short text documents.

The first method, Co-Frequency Expansion (CoFE) [Pedrosa et al., 2016], exploits the co-occurrence frequency (co-frequency) of terms in the collection to define a metric space. The main idea behind CoFE is that words with high co-frequency have also high probability of belonging to the same topic, and hence can be used to expand documents. The second, Distributed Representation-based Expansion (DREx) [Bicalho et al., 2017], exploits the powerful word embedding representation to model word similarities [Mikolov et al., 2013d; Pennington et al., 2014a], taking advantage of the semantics and vector algebra captured by this type of representation.

1.1 Objectives

The main objective of this dissertation is to propose a general and easily extendable framework, capable of enriching short text documents for the topic modeling task. In order to reach that goal, five specific objectives were identified:

- Perform a literature review of the existent methods for topic modeling and their challenges on dealing with short text;
- Perform a literature review of the existent methods for topic modeling proposed specifically for short text scenarios, and an analysis of their limitations;
- Formalize a general framework for document expansion, based on metric spaces, capable of overcoming the limitations of the existent methods;
- Propose two specializations of the framework using different approaches, in order to show its generalization power;
- Evaluate the proposed methods using real datasets, and compare their results with the most relevant baselines in the literature.

1.2 Main Contributions

The main contributions of this paper are:

- A literature review on the existent methods for topic modeling on short-text.

- A new framework based on metric spaces for generating larger pseudo-documents that are more suitable for topic extraction;
- Two instances of this framework, one based on word co-occurrence and the second based on word vectors, which can be coupled to any topic model to improve topic extraction;
- The results of the expansion based on word vectors are statistically significant better than those obtained by the state-of-the art methods in the original text both using the NPMI metric and when considering the classification task.

1.3 Document organization

The remainder of this document is organized as follows. Chapter 2 introduces related work on topic models for short text. Chapter 3 describes the general framework and instantiates CoFE and DREx. Chapter 4 introduces the experimental methodology and shows the results obtained. Finally, Chapter 5 lists our conclusions and directions of future work.

Chapter 2

Background and Related Work

In this chapter we present fundamental definitions related to topic modeling, text expansion and word embedding representations. We first formalize the concept of topic and present two popular methods for topic modeling. We explain the reasons these methods are not suitable when considering the short-text scenario and present the most relevant works in this area. We also discuss and review relevant methods for text expansion and word embedding representations.

2.1 Topic Modeling

Topic identification methods are traditional machine learning tools that have become popular in many areas of knowledge, given its success in modeling and explaining real world phenomena.

The task of topic modeling consists of automatically discovering and annotating textual data with thematic information. In this section we focus on two popular methods for solving this task: Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and Non-negative Matrix Factorization (NMF). While the first is a probabilistic method that uses a generative process, the second is based on matrix factorization techniques.

2.1.1 Latent Dirichlet Allocation

LDA is a probabilistic method conceived to describe document collections using a set of topics. A topic is formally defined as a probability distribution over a fixed vocabulary used by a collection of documents, where the words with the highest probability are more likely to be selected to describe the topic. For example, considering the topic

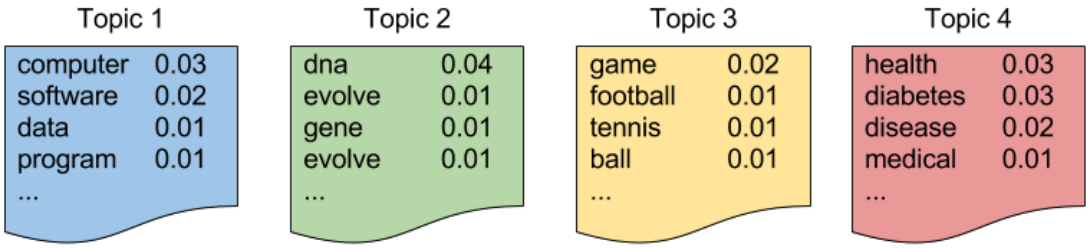

Figure 2.1: Example of topics. Topics are represented as a probability distribution over the vocabulary concentrated in some terms.

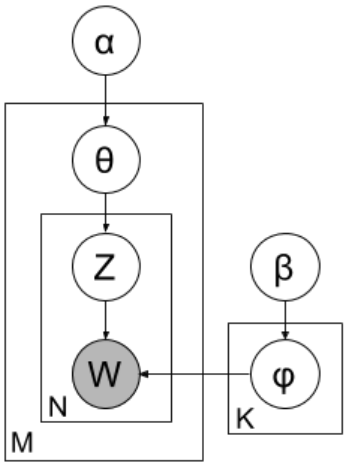

Figure 2.2: LDA graphical model in plate notation

sports, it may be described by a probability distribution concentrated in terms like *game, football, tennis, ball, and match.* Figure 2.1 shows a few examples of topics extracted using the LDA algorithm. Note that each word has a value associated to it. This value is the probability of this word appears in a document about the topic. For example, given a document about sports (Topic 3), the word *game* has a probability of 2% of appearing in this document.

The most intuitive way to understand LDA rationale is by looking at its generative process, i.e., an imaginary process assumed by the model to be responsible for creating the documents in the collection. A plate notation of LDA's generative process is illustrated in Figure 2.2. In a plate notation shaded nodes represent observed variables, while the other nodes represent latent variables. The plate notation simplifies the representation of large probabilistic graphical models by placing replicated structures into numbered rectangles, where the number represents replications (e.g. rectangles numbered with M and N refers to the replicated structures of M documents and N words, respectively, while K refers to the number of topics).

In LDA, topics are defined as a multinomial distribution over vocabulary words (φ distribution), documents are mixtures of topics (θ distribution) and words are derived,

one by one by sampling a topic Z from θ and then a word W from φ_z, where the index z indicates the topic.

This generative process reflects that fact that documents contain multiple topics. For example, a document can belong to the topics sports and politics. In other words, each document has topics in different proportions (defined by θ distribution). Only words within documents are observable variables and all priors (α and β) are defined to be Dirichlet distributions (Dir). The main objective of the algorithm is to infer the hidden (or latent) variables, such as topics proportions of documents and words distributions per topic. The generative process of LDA is detailed in Figure 2.3.

---

1. For each target topic z
   a) Draw a multinomial distribution over all terms, $\varphi_z \sim Dir(\beta)$
2. For each document d on the corpus
   a) Draw a multinomail distribution over the target topics, $\theta_d \sim Dir(\alpha)$
   b) For each word w in document d
      i. Draw a topic from the chosen distribution, $z_{w,d} \sim Multinomial(\theta_d)$
      ii. Draw a term from the topic chosen, $w \sim \varphi_{z_w,d}$

Figure 2.3: LDA generative process.

It is important to mention that topic models do not have any prior information about the topics in a collection, and that the documents do not have any annotation or keyword to define which subjects they cover. The topic distribution is generated by inferring the latent structure that, with higher probability, is responsible for generating the observed document collection. There are several techniques to perform this inference, including Variational Inference [Blei et al., 2003], Expectation Propagation [Minka Lafferty, 2002] and Markov Chain Monte Carlo (MCMC) [Griffiths Steyvers, 2004]. The latter is the most used in probabilistic topic models. Throughout this work, unless state otherwise, the described methods for topic modeling use MCMC (especifically, we use the Gibbs Sampling method) to perform the inference of the hidden variables.

2.1.2 Non-negative Matrix Factorization

In contrast with generative probabilistic methods, Non-negative Matrix Factorization (NMF) [Xu et al., 2003; Wang et al., 2012] is based on matrix factorization techniques,

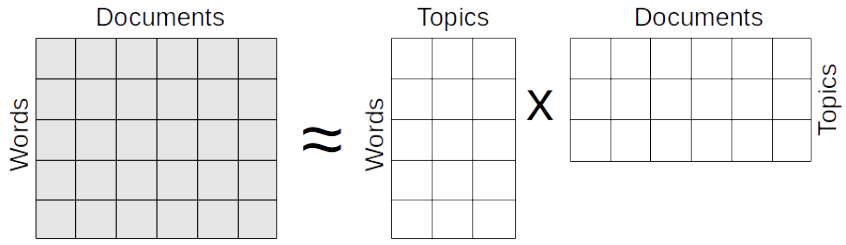

Figure 2.4: Non-negative matrix factorization (NMF). The input matrix is decomposed into two others.

where a matrix of terms per documents is projected into a k dimensional space where each dimension correspond to a topic. NMF belong to the same family of methods as Latent Semantic Indexing (LSI) [Deerwester et al., 1990]. LSI uses Singular Value Decomposition (SVD) to identify semantic latent factors with orthogonal restrictions. Recent works changed the orthogonal restriction to allow more flexibility [Wang et al., 2013]. However, these methods do not have an intuitive interpretation to the negative values found, i.e. there is no clear understanding about the topic-document relation when it is negative.

NMF introduced a non-negative constraint to the matrix decomposition. This restriction guarantees all values in the matrix decomposition to be positive, making the relation topic-document more intuitive. NMF decomposes the term-document matrix in two low-rank non-negative matrices as shown in Figure 2.4: (i) the term-topic matrix, where each column represents a topic as a convex combination of terms; and (ii) the topic-document matrix, where each column represents a document as a convex combination of topics.

Formally NMF can be briefly described as follows. Given a non-negative input matrix $V \in \mathbb{R}^{n \times m}$, where each column represents a document and each line represents a term, and given an integer $k << min\{m, n\}$, representing the number of desired latent factors, or topics in our scenario, NMF finds two non-negative matrices $W \in \mathbb{R}^{n \times k}$ and $H \in \mathbb{R}^{k \times m}$, where:

$$V \approx WH. \tag{2.1}$$

The usually approach to solve the NMF problem is to finding W and H that minimize the Frobenius norm of the difference $V - WH$ [Lee Seung, 2001]:

$$\min_{W,H} \|V - WH\|_F^2 \tag{2.2}$$

The matrices W and H are generally not unique and the designed algorithms to solve the minimization problem 2.2 generally begin by initial estimates of these matrices, followed by alternating iterations using update rules to improve the estimates.

The most used update rules are those proposed by Lee Seung [2001] based on multiplications of the matrices:

1. Initialize W and H with non-negative values and scale the columns of W to unit norm.

2. Iterate for each c,j and i (matrices indices) until convergence or after l iterations:

   a) $H_{cj} \leftarrow H_{cj} \frac{(W^T V)_{cj}}{(W^T W H)_{cj}}$

   b) $W_{ic} \leftarrow W_{ic} \frac{(V H^T)_{ic}}{(W H H^T)_{ic}}$

   c) Scale the columns of W to unit norm.

2.2 Topic Modeling on short text

Uncovering hidden topics in short texts is a major problem in topic modeling. Since short texts are all over the Web, e.g. messages posted in social media, status messages, blog comments, questions in Q & A websites, advertisement texts, image captions, etc, uncovering the topics of this type of messages is crucial to a wide range of applications, including context analysis [Ramage et al., 2010], user characterization [Weng et al., 2010], real-time topic detection [Lin et al., 2010] etc.

The conventional algorithms rely on the words co-occurrence in the document level to infer the topics and, therefore, they suffer from the severe data sparsity when dealing with short text . Aiming to solve this problem, two types of approaches have been proposed in the literature: (i) those that modify current topic modeling methods in a way that they minimize from the aforementioned problems; (ii) those that modify the input text, creating larger and richer pseudo-documents, which can then, be processed by the traditional topic modeling methods.

2.2.1 LDA Modifications for topic modeling

Following we detail proposed methods that propose to extend the LDA method and are focused on topic modeling for short texts.

DLDA: In Jin et al. [2011], the authors propose to use external information to improve the topic modeling task. They propose Dual LDA (DLDA), which enhances topic modeling for short texts via transfer learning from an auxiliary dataset of longer texts. Previous works have followed a similar approach, but they ignore the semantic and topical inconsistencies between the target and auxiliary data, [Phan et al., 2008, 2011; Xue et al., 2008]. As an example of these inconsistencies, the authors cited an advertising scenario. According to them, when merchants advertise a product using short banner Ads, the content often emphasizes on the credibility and price aspects. At the same time, in a Web page for selling the same product, merchants may focus more on the branding and product features.

In order to account for these inconsistencies, DLDA jointly learns a set of target topics on short texts and another set of auxiliary topics on long texts. As stated by the authors, a crucial step on DLDA is the selection of the auxiliary dataset with long texts. Even DLDA does not require any correspondence structures between the short-text and the auxiliary data, both datasets have to be topically-related. This limitation can restrict the application of DLDA, because it requires some knowledge of the existent topics in the short-text.

The authors proposed two different versions of the DLDA: $\alpha - DLDA$ and $\gamma - DLDA$. The first uses two asymmetric priors on the topic mixture proportions to control the relative importance of the two different topic classes (target and auxiliary) for generating short and long texts. This version of the DLDA only imposes certain settings to the hyper-parameters of the LDA, without changing the generative process.

$\gamma - DLDA$ in turn, introduces a binary switch variable into the LDA model. This switch is used for choosing between the two types of topics (target and auxiliary) when generating each term of the documents. If we compare the classical two-step process followed by LDA to $\gamma - DLDA$, it is modified as described in Figure 2.5.

BTM: In Yan et al. [2013], the authors propose a new method called Biterm Topic Model (BTM), which learns the topics by directly modeling the generation of biterms, i.e. pairs of words that co-occur in the same document considering the whole corpus [Yan et al., 2013]. The major differences between BTM and other topic models is that BTM explicitly models the word co-occurrence patterns, while other methods implicitly use this information during the inference step.

BTM also overcomes the sparsity problem at the document-level by aggregating the biterms of the whole corpus. Its generative process models the word co-occurrence patterns rather than a single word. The graphical of BTM is illustrated in figure 2.6 and its generative process is described in Figure 2.7.

The authors compared BTM to three other topic modeling methods using a public Twitter dataset: (a) standard LDA, which treats each tweet as a document; (b) LDA-

1. For each target topic z
   a) Draw a multinomial distribution over the vocabulary, $\varphi_z^{tar} \sim Dir(\beta^{tar})$
2. For each auxiliary topic z
   a) Draw a multinomial distribution over the vocabulary, $\varphi_z^{aux} \sim Dir(\beta^{aux})$
3. For each document d on the target and auxiliary corpus
   a) Draw a multinomial distribution over the target topics, $\theta_d^{tar} \sim Dir(\alpha^{tar})$
   b) Draw a multinomial distribution over the auxiliary topics, $\theta_d^{aux} \sim Dir(\alpha^{aux})$
   c) For each word w in document d
      i. Choose between using the target or auxiliary topics, $x_{w,d} \sim Bernoulli(\gamma)$
      ii. Draw a topic from the chosen distribution, $z_{d,n} \sim Multinomial(\theta_d^{x_{n,d}})$
      iii. Draw a term from the topic chosen, $w_{d,n} \sim Multinomial(\varphi_{z_{d,n}}^{x_{d,d}})$

Figure 2.5: DLDA generative process.

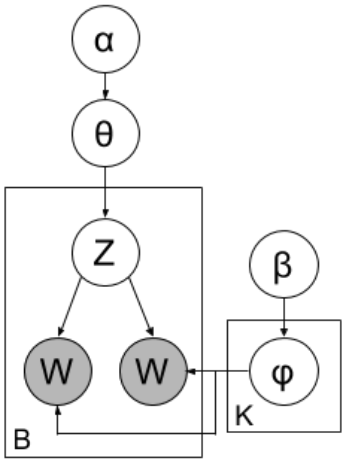

Figure 2.6: BTM graphical model in plate notation.

U, which aggregates all the tweets from a user into a single larger pseudo-document; (c) mixture of unigrams, which assumes that each tweet has a single topic. BTM outperforms the baselines in these datasets. Its main disadvantage, however, is that it does not model the topic proportions for the documents during the learning process. This has to be done after the learning phase by a different model.

LDA and word embeddings: More recently, Sridhar [2015] and Nguyen et al. [2015] tried to incorporate word embeddings to traditional topic models. A word embedding is a vector representation of the word that tries to capture the context in which that word is used. Recent approaches are based on deep neural networks, and learn the vector

1. For each target topic z
   a) Draw a multinomial distribution over all terms, $\varphi_z \sim Dir(\beta)$
2. Draw a multinomail distribution over the target topics, $\theta \sim Dir(\alpha)$
3. For each biterm b on the whole corpus
   a) draw a topic assignment, $z_b \sim Multinomial(\theta)$
   b) draw two words, $w_i, w_j \sim Multinomial(\varphi_{z_b})$

Figure 2.7: BTM generative process.

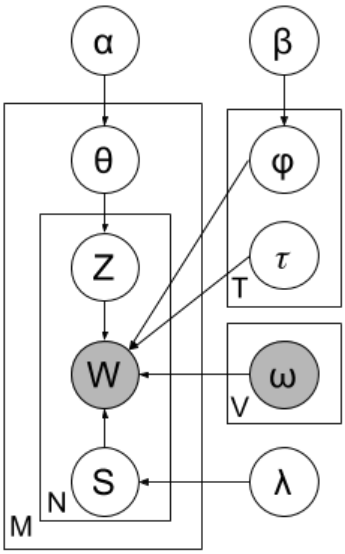

Figure 2.8: LFLDA graphical model in plate notation

representations by predicting the central word given a set of surrounding words [Mikolov et al., 2013c; Pennington et al., 2014b; Liu et al., 2015] (see section 2.4). Latent feature vectors have been used by a wide range of applications in the Natural Language Processing (NLP) community, and Glorot et al. [2011] and Cao et al. [2015] showed that it can be used to learn topics from a textual data.

In this same direction, Sridhar [2015] proposed a new topic model that performs soft clustering over latent feature word vectors. They use Gaussian mixture models (GMMs) to capture the notion of latent topics in the word vectors representations. Nguyen et al. [2015], in turn, proposed LF-LDA, a LDA modification which includes word embeddings (latent features) trained on large external corpus to enhance the topic model task in short datasets.

The graphical model of LF-LDA is illustrated in Figure 2.8, and we describe its generative process in Figure 2.9: They tested LF-LDA on three different tasks, and it

obtained better results than LDA in all scenarios.

---

1. For each word w in the vocabulary
   a) learn its vector representation, ω_w
2. For each target topic z
   a) Draw a multinomial distribution over all termos, $\varphi_z \sim Dir(\beta)$
   b) Create a vector representation of the topic, τ_z
3. For each document d on the corpus
   a) Draw a multinomail distribution over the target topics, $\theta_d \sim Dir(\alpha)$
   b) For each word w in document d
      i. Draw a topic from the chosen distribution, $z_{w,d} \sim Multinomial(\theta_d)$
      ii. Choose between using multinomial or latent feature component of the chosen topic, $s_{w,d} \sim Bernoulli(\lambda)$
      iii. Draw a term from the topic chosen, $w \sim (1 - s_{w,d})Multinomial(\varphi_{z_{w,d}}) + s_{w,d}Categorical(\tau_{z_{w,d}}, \omega_w)$

Figure 2.9: LF-LDA generative process.

2.2.2 Input Modification Methods

Following we present methods that opt to modify the input documents and after that conduct the topic modeling phase using these modified documents. In this approach any topic modeling algorithm may be considered.

Twitter-LDA: In Zhao et al. [2011] the authors propose a modification to LDA to make it more suitable for Twitter datasets. Their method, named Twitter-LDA, assumes that a single tweet is usually about a single topic, and every Twitter user has a topic distribution that defines the probabilities of this user to write a tweet related to each topic. Their assumption of a single topic per tweet simplifies the LDA model, because instead of learning a topic distribution for each document the model only needs to choose a single topic. The experimental results showed that Twitter-LDA obtained better results than the traditional LDA. The major drawback of Twitter-LDA is the fact that its needs the meta-data (the author) of the text and in some scenarios this information is not available.

WNTM: In Zuo et al. [2016] the authors propose a word co-occurrence network-based model named WNTM. In WNTM, an undirected weighted graph of words co-occurrence is derived from the original documents. Words that co-occur at least once in a same document are linked, and the edge is weighted by the words co-occurrence frequency. Each word w_i of this graph generates a completely new pseudo-document, which is made of the words adjacent to w_i in the graph.

These pseudo-documents are then given to LDA and the topics are extracted. WNTM explores the fact that, even in short-text scenarios, the word-word space is rather dense, making the algorithm less sensitive to the document length or heterogeneity of the topic distribution. Note that CoFE, one of our proposed methods, also uses a words co-occurrence graph to generate pseudo-documents. However, CoFE differs significantly from WNTM. While the former expands each document with vocabulary words that co-occur more frequently with each word of the document, WNTM creates entirely new documents from the graph.

2.3 General Text Expansion Approaches

As showed in the previous section, generating larger pseudo-documents or change the input representation to traditional topic modeling methods are less popular then proposing or modifying existent methods. In other areas, in turn, text expansion is commonly used to enhance the performance of standard algorithms. Good examples of other scenarios where dealing with short text is also a challenge, include mining short text [Rosso et al., 2013] and automatic query expansion (AQE) for search engines [Carpineto Romano, 2012].

Rafeeque et al. [Rafeeque Sendhilkumar, 2011] presented a survey on the challenges and open issues on mining short text. Most of the works they reviewed are applied to text classification and clustering scenarios, and several of them propose to overcome data sparsity with a document expansion approach.

In a information retrieval system, a user submit keywords (query) that are matched against the collection index to find relevant documents related to the query. As described by Carpineto Romano [2012], when a user-made query contains multiple topic-specific keywords that accurately represent the user intention, the system is likely to return good and relevant documents. However, in the great majority of cases, the query is too short and sometimes contains ambiguous words and expressions, which makes this simple retrieval model not very efficient.

The motivation behind AQE is the same of this work: user-made queries are usually

very short, and the lack of information makes it difficult to model the user intention [Carpineto Romano, 2012]. As the volume of available data has dramatically increased, and the average length of the queries has remained low, AQE techniques have received more attention from the scientific community.

In the last years, several approaches for AQE have been proposed, especially at the Text Retrieval Conference (TREC)[1], where researchers are reporting significant improvements in the document retrieval task.

Many query expansion methods are based on the information contained in the top-ranked documents retrieved by a search engine system for the original query, and thus its application in a topic extraction scenario not straight forward, as it would require the creation of an index and a search engine. Other approaches are more relevant to our work. The most significant ones are those classified as *One-to-One and One-to-Many Associations*. They are mostly focused on extracting features from the document collection that can be used to calculate similarities between every pair of *terms*. Hence, given a query, the terms more similar to the query keywords can be used in the expansion.

The framework for documents expansion proposed here is based on the same idea. We formally define a general framework that explores similarities between words or groups of words to expand short-texts. Besides the formal definition of the framework other differences between our approach and the AQE techniques are: (i) AQE wants to expand a single query, while our goal is to expand the entire set of documents, (ii) although the queries are short, the documents on the collection are usually long and many relevant features can be extracted, while in our scenario the collection is compound of short-text documents, (iii) our approach is general in the sense that it can be instantiated for several different methods depending on the exploited features.

Despite being originally designed for different tasks, one should be able to adapt any of the methods previously proposed in the literature for the topic modeling task [Hotho et al., 2003; Pinto et al., 2011; Sedding Kazakov, 2004]. Among the promising techniques proposed so far that could be adapted for topic modeling we highlight the work of Pinto et al. [2011]. They proposed Self-Term Expansion (STE), a method that does not require the use of any external source of data – while many other works rely on sources such as the WordNet or ontologies – and presents a set of similarities to one of the framework specializations presented in this paper.

STE works by replacing the terms of a document by a set co-related terms, where the correlation score between pairs of terms is computed by analyzing their co-occurrence on the dataset. It uses the PMI score [Manning Schütze, 1999], which captures semantic associations between pairs of terms and is commonly used as a quality metric in the topic modeling literature, to create a ranked co-occurrence list between every pair of terms on the dataset. It then uses this list to expand every document on the corpus.

[1] `http://trec.nist.gov/`

2.4 Word embedding representation

Most traditional natural language processing techniques consider words as atomic units of processing. While this approach has produced impressive results in many domains, it can be improved with the exploitation of words meanings and similarities, especially for rare terms. Distributed representation of words were conceived to capture semantics by coding each word and its context – an important component for assigning meaning to it – in a real vector-space embedding.

Word vectors representations are usually expected to be consistent with vector algebra, in the sense that some operations in the vector domain (such as sum or difference) should keep some degree of consistency with similar semantic manipulations. For example, Mikolov et al. [2013b] shows word vectors models in which the operation $vector(\text{“king”}) - vector(\text{“man”}) + vector(\text{“woman”})$ produces a vector that is close to $vector(\text{“queen”})$[2]. Even more abstract concepts could be derived, not only terms from a vocabulary, like the similarity between the results of operations $vector(\text{“cars”}) - vector(\text{“car”})$ and $vector(\text{“apples”}) - vector(\text{“apple”})$, which capture the concept of plurality.

Among many previous published works on fundamentals and applications of distributed words representation we cite three models: Continuous Bag of Words (CBOW) [Mikolov et al., 2013b], Skip-Gram (SG) [Mikolov et al., 2013b] and Global Vectors (GloVe) [Pennington et al., 2014a].

2.4.1 Skip-Gram and Continuous Bag of Words

The Skip-Gram (SG) model is an artificial neural network whose architecture is shown in Figure 2.10. The algorithm uses one-hot encoding to represent words in its input and output. This encoding transform categorical features, such as words, into boolean vector of size S where only one bit is high (set to 1). In our scenario the size S is equal to the vocabulary size V (number of distinct words) and each boolean value b is associated to a single word w. Hence, for each word w, its one-hot encoding is a boolean vector of size V with only the bit associated w set to one.

The input of the model is a word and the output is a set of words inside a context window of size C. Therefore, the task of the network is to infer the surrounding context for a given word. This mapping makes possible to transform any text corpus in a training

[2]Here $vector(w)$ means the vector representation of word w.

dataset for a classification problem.

All words in the output layer share the $W'_{N\times V}$ weight matrix and produce a multinomial distribution using a softmax function. Words vectors of size N (which is a parameter and defines the size of the hidden layer) are extracted from the $W_{V\times N}$ weight matrix, one vector by row. Parameters are learned through backpropagation and stochastic gradient descent. Performance is optimized by using hierarchical softmax and negative sampling [Mikolov et al., 2013a].

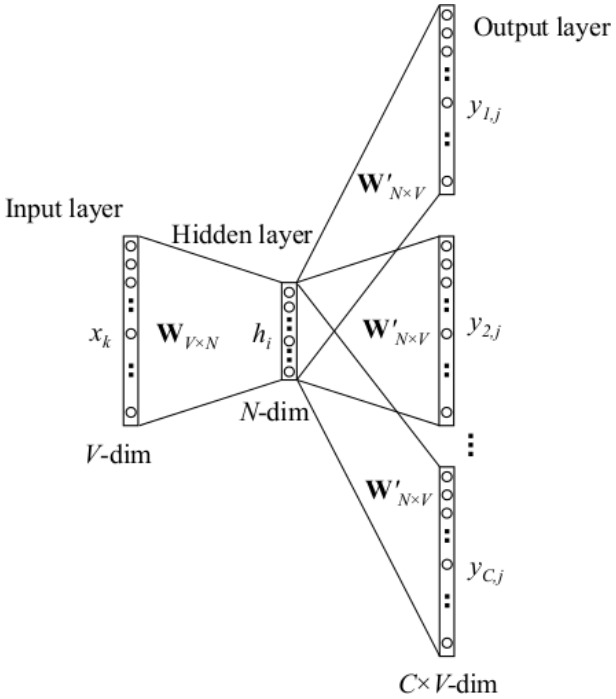

Figure 2.10: SG neural network architecture (Figure source: http://alexminnaar.com/word2vec-tutorial-part-i-the-skip-gram-model.html).

The CBOW model, presented in Figure 2.11, is also an artificial neural network that learns the inverse task of the SG model. The task of the network is, therefore, to infer a word given its surrounding context of size C. The learning process is similar to the SG model. Input and hidden layers are connected by a shared weights matrix $W_{V\times N}$, from which all words vectors are extracted, one vector by row.

2.4.2 Global Vectors

GloVe is another algorithm capable of learning vector representation of words from a textual dataset. The learning process of GloVe is based on word co-occurrence probabilities. Given a word co-occurrence matrix X, where each element X_{ij} represents how

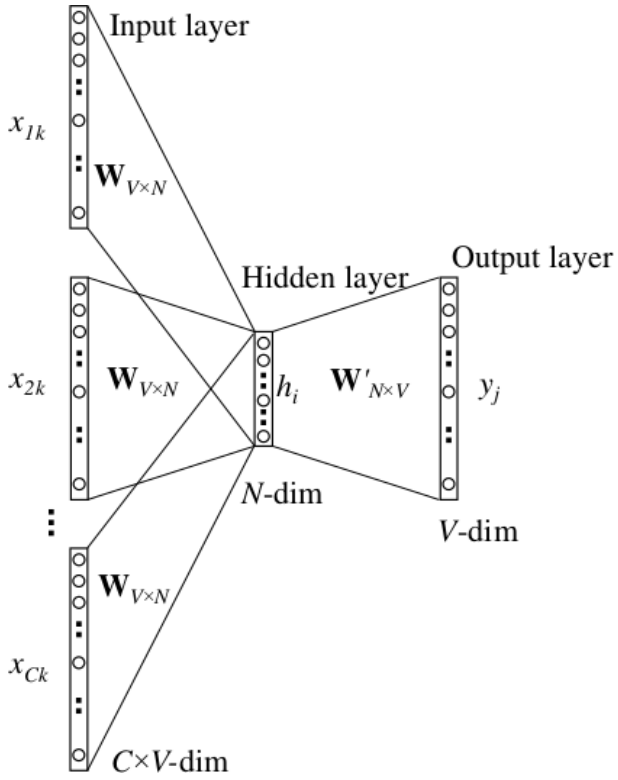

Figure 2.11: CBOW neural network architecture (Figure source: http://alexminnaar.com/word2vec-tutorial-part-i-the-skip-gram-model.html).

often a word i appears in the same context of word j, the probability $P_{i,j}$ of a word j appear in the context of another word i can be defined as $P_{i,j} = P(j|i) = X_{i,j}/X_i$, where $X_i = \sum_k X_{i,k}$ is the total frequency of the word i. The context of each word is defined as set of surrounding g words, where g is defined by a parameter called window size.

To learn the vector representation of words, GloVe explores the similarity between two words i and j, given the context of a third word k. This similarity is captured by the ratio $P_{i,k}/P_{j,k}$. Let w_i, w_j and w'_k be the vector representation of words i, j and k, respectively, where $w_i, w_j, w'_k \in \mathbb{R}$. Equation 2.3 hypothesizes that there is a function F over words vectors and context words vectors that is proportional to $P_{i,k}/P_{j,k}$.

$$F(w_i, w_j, w'_k) = P_{i,k}/P_{j,k} \tag{2.3}$$

The method learns the parameters w_i, w_j, w'_k of F, which are the word vector representations, with some restrictions over F, which impose a way of combining w_i, w_j and w'_k considering the linearity of vector space structure (for a detailed list of the restrictions, the reader is referred to). From these restrictions, a cost function J can be defined and the word vectors representations found by minimizing J through a gradient descent algorithm. Equation 2.4 presents the cost function J, where b_i and b'_k are scalar biases and f is a weighting function.

$$J = \sum_{i,j=1}^{V} f(X_{i,j})(w_i^T \cdot w'_k + b_i + b'_k - log(X_{i,k})) \tag{2.4}$$

Function $f(X_{i,j})$ alleviates the effects of extreme values of $X_{i,j}$, and it is defined as:

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \tag{2.5}$$

where x_{max} is the cutoff parameter and α the scaling power parameter.

Chapter 3

A framework for document expansion based on word embeddings

The most intuitive and natural approach for the problem of topic modeling in short-text scenario is to make the input data (short-text corpora) more suitable for the topic modeling task. The simplest way to accomplish this is by expanding each short-text document by adding to it new words that are similar to the words that already appear in the document, increasing word co-occurrences, reducing sparsity and generating larger pseudo-documents. Here, we proposed a general framework for short-text expansion based on this idea and them present two instantiations of this framework.

3.1 General Framework

The framework proposed in this dissertation works in four main steps, shown in Figure 3.1: (i) corpus preprocessing, (ii) definition of a metric space, (iii) generation of the candidate n-grams for the expansion, (iv) selection of candidates to expand the text. Following we detail and discuss each of the aforementioned steps.

3.1.1 Preprocessing

The preprocessing phase follows the traditional steps of text preprocessing in Natural Language Processing (NLP), which includes:

- Lowercase the words. Most of the text mining algorithms, including topic modeling, are case sensitive. Hence all words have to be lowercased in order to be considered

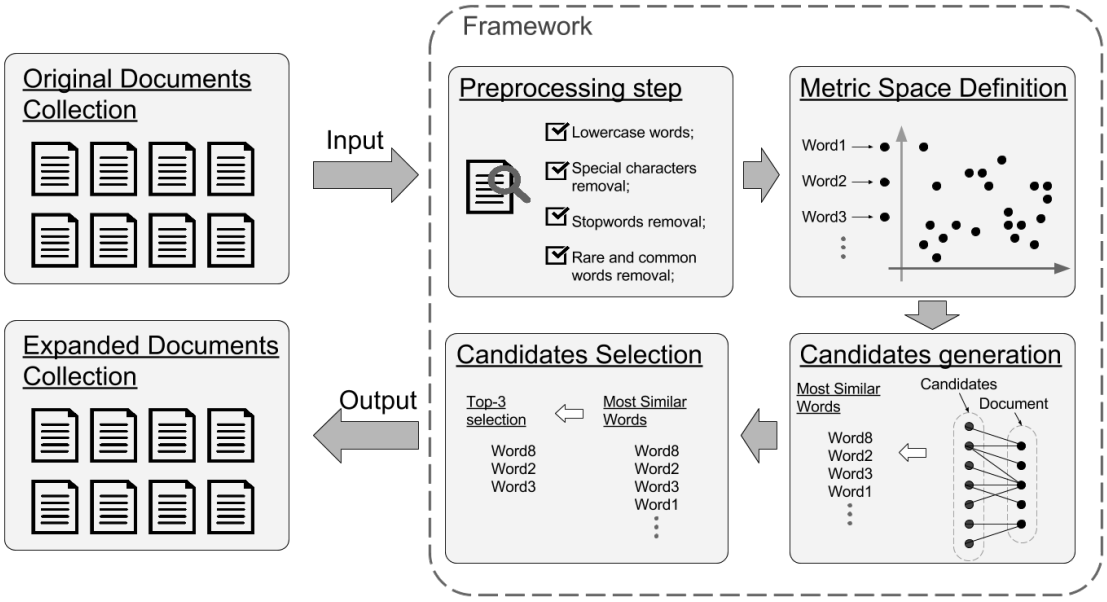

Figure 3.1: Overview of the framework proposed for short text expansion.

the same.

- Punctuation, accentuation and special characters removal. Punctuation and accentuation are not relevant to the topic learning process and the same applies to non alphanumeric characters.

- Stopwords removal: A word is said to be a stopword when it is considered irrelevant to the problem. In the context of topic modeling, any word that does not carry information about the topic subject, can be viewed as a stopword, such as common articles and nouns. The removal of these words is an important step when preprocessing textual data for topic modeling algorithms since the frequency of such words is usually high and can misguide the learning process.

- Common and rare words removal: Another set of words that are irrelevant to topic modeling algorithms are common and rare words. The reason is the same of the stopwords removal step. For this step, we consider rare words as being words that appear in less than 5% of documents, and common words those that appear in more than 90% of documents.

Another common preprocessing step is the application of stemming. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form. We choose to not apply stemming in our textual data, because there is not a consensus in the literature of whether it is a good or bad practice for the topic modeling task.

3.1.2 Definition of a metric space

The goal of our proposed framework is to expand documents with new words that are similar to words that already appear in the document, which requires the definition of a function capable of calculating the similarity between the words in the document and the words available for the expansion.

These set of available words can be defined in several ways, all English words, a subset of relevant words, etc. In our work, we use the dataset vocabulary to start the expansion process. The dataset vocabulary is the set of unique words that appear in the document collection after the preprocessing stage. By choosing the vocabulary we do not introduce new words to the collection, avoiding future inconsistency problems (in the case of our method being used in conjunction with other algorithms).

To make the framework easily extensible, we formalize the problem of finding similar words using the concept of metric space. A metric space is a set for which distances between all members of the set are defined. More formally, a metric space is a pair $(\mathcal{V}, g)$ where $\mathcal{V}$ is a set of elements and g a metric function that defines a distance between every pair of points v_i, $v_j \in \mathcal{V}$. Metric spaces guarantee that the minimum properties of distance functions are satisfied:

- $g(v_i, v_j) \geq 0$ and $g(v_i, v_j) = 0 \iff v_i = v_j$;
- $g(v_i, v_j) = g(v_j, v_i)$;
- $g(v_i, v_j) \leq g(v_i, v_k) + g(v_k, v_j)$.

In our scenario, the elements of the metric space are n-grams (i.e., a sequence of n words), already in the document and the n-grams in the vocabulary. The distance function $g(v_i, v_j)$ measures the dissimilarity between the n-gram v_i and v_j.

3.1.3 Generation of candidates n-grams for the expansion

Many different approaches can be used to define the words that should be added to a document, but the use of the metric space allow us to define a general method. Given the set S of n-grams in the document, the n-grams candidates for the expansion are the n-grams closest to each n-gram $\in S$. Definition 1 formalize the notion of closest n-grams.

Definition 1. Let $\mathcal{V}$ be the vocabulary, $v \in \mathcal{V}$ a n-gram belonging to the documents, and g a distance function. A t-nearest neighbor n-gram function based on v, denoted as $\mathcal{NN}(v, t)$, determines the t closest n-grams with respect to v. Formally, $\mathcal{NN}(v, t) = \mathcal{A} : |\mathcal{A}| = t \wedge \forall p \in \mathcal{A}, \forall x \in (\mathcal{V} - \mathcal{A}), g(v, p) \leq g(v, x)$.

To define how similar a candidate n-gram is to a document d we create a similarity graph based on the metric space and the definition of the $\mathcal{NN}$ function. The similarity graph is presented in Definition 2.

Definition 2. Let $G_d = (L_d \cup R_d,\ E_d)$ be a bipartite graph representing the short text d, where $L_d \cup R_d \subseteq \mathcal{V}$. G_d has two types of nodes: $l \in L_d$, which represents n-grams extracted from d; and $r \in R_d$, where R_d is the set of candidate n-grams t to expand document d. An edge $e_d = (l, r, w)$ determines the relationship between a n-gram l present in d and a candidate n-gram r with weight w. Formally, $e_d = (l, r, w) : l \in d,\ \ r \in \mathcal{NN}(l, t), w = 1/g(l, r)$ and $\mathcal{NN}(l, t) \subseteq V$.

3.1.4 Selection of candidates n-grams to expand the text

For each candidate n-gram, the sum of the weights of each incoming edge determines how similar this n-gram is to the whole document. This value can be used to rank the candidates according to their similarities to the whole document. Given this rank, the process of choosing a subset of words to be added to the document can be done in several ways: top m selection, probabilistic selection, etc. For simplicity we chose the top m selection, where the m most similar words are selected. Regarding the number of words m to be added, we define a scaling parameter S which specifies the size of a document d after the expansion step as a product of the original size $|d|$ by the scaling parameter S. For example, consider a short document with an original size of 10 words, if we set $S = 2$ the final size of document should be 20, $S = 1.5$ the final size should be 15 and so on.

3.1.5 Expansion Algorithm

We present the general expansion framework in Algorithm 1. It has the following parameters: (i) D, the collection of documents to be expanded; (ii) $(\mathcal{V}, g)$, the metric space that contains the representation of document n-grams and the function to compare them;

Algorithm 1 General Expansion Framework

Require: $D, (\mathcal{V}, g), S$

1: **for** $d \in D$ **do**
2: $M \leftarrow |d| \times S$
3: $t \leftarrow M - |d|$ ▷ Number of words to be added
4: $G_d \leftarrow$ Graph $(L_d \cup R_d, E_d)$ generated from $(\mathcal{V}, g)$ and $\mathcal{NN}(t)$ ▷ *Def.* 2
5: $C_d \leftarrow \emptyset$ ▷ Candidate words
6: **for** $e_d = (l, r, w) \in E_d$ **do**
7: $C_d \leftarrow C_d \cup \{r, w\}$
8: $h \leftarrow \text{SelectionMethod}(C_d, t)$ ▷ Selected t words
9: $d \leftarrow d \cup h$

(iii) S, a scaling factor that controls the final documents length. The metric space is the basis to build a graph of candidate words used to create the new short text representation, i.e., the pseudo-document.

The expansion procedure is performed for each document d (lines 1–9 from Algorithm 1).

1. The final number of words M is calculate by the product of the actual document length and the parameter S (line 2).

2. The numbers of words t that must be selected to expand the document is calculate (line 3).

3. A similarity graph for d is generated using the given metric space, and t is passed as a parameter that represents the number of neighbors for the function $\mathcal{NN}$ (line 4).

4. Then an initially empty set C_d accumulates, for each potential expansion n-gram r, its respective weight w (lines 5–7). Next, t n-grams are selected from C_d, top t selection according to their similarities to the whole document d (lines 8–9).

The similarity of a new n-gram r to d is given by the total sum of its weights in C_d (i.e., sum of its degrees). Note that we attach the number of neighbors to be considered for each word to the number of words that must be added to the document in order to satisfy the final document length of M. By doing so, we guarantee that the set of candidates words C_d will have at least $M - |d|$ words and those words are the most similar words to the document.

Despite being simple the proposed framework has some interesting advantages. The most important one is the fact that the expanded dataset can be used as input for any topic modeling algorithm. In order to show how general is the proposed framework, we introduce two methods for documents expansion following the framework: *C*o-Frequency Expansion (CoFE) and *D*istributed Representation Expansion (DREx).

3.2 Co-Frequency Expansion (CoFE)

The CoFE method is simple and considers that similar words have a higher likelihood of occurring in the same context. That is, the conditional probability of one word to occur in a document sliding window given that a second word was observed, should be higher if the words are similar and lower otherwise. Note that, in this scenario, the document n-grams considered are the words themselves (i.e., we consider n = 1). In order to detail CoFE, we first define a metric space that exploits the co-occurrence of words.

In the defined metric space $(\mathcal{V}_{CoFE}, g_{CoFE})$, each word w of the vocabulary is represented by a set $O_w \in \mathcal{V}_{CoFE}$ that contains all documents where w occurs. We define the distance metric g_{CoFE} as:

$$g_{CoFE}(w_i, w_j) = 1 - Jaccard(w_i, w_j) = 1 - \frac{|O_i \cap O_j|}{|O_i \cup O_j|} \tag{3.1}$$

Having the metric space, the document expansion follows the steps listed in Algorithm 1. Given a document d, we first generate the bipartite graph $G_d = (L_d \cup R_d, E_d)$, where nodes in L_d are the words extracted from d, R_d is the set of all t nearest words of each word in L_d, and there is an edge $e(l, r, w) \in E_d$ between each node $l \in L_d$ and its t nearest nodes $r \in R_d$. As each edge represents the similarity between nodes l and r, the weight w is simply defined as the Jaccard of the words l and r. Figure 3.2 presents an example of a graph derived from metric space $(\mathcal{V}_{CoFE}, g_{CoFE})$. The original document contains the text "President Obama" (represented by squares in the graph), and for each word we present the neighbor candidate words (represented by circles in the graph) considering $t = 4$. Assume that the expanded pseudo-document should contain a maximum number of words $M = 6$ after expansion. Before selecting the expansion words, we add up the weights (edges) of the candidate words connected to more than one word in the original document. For example, *Barack* is connected to both *president* and *Obama*, and hence its final weight is set to 0.45. All other words remain with the same weight, as they are connected to a single word.

The last step of Algorithm 1 is the SelectionMethod, in which we use the weights of words as their probability of appearing in the final pseudo-document. Following the example, the pseudo-document contains the original words: *president* and *obama*; and the selected words: *barack*, *administration*, *house* and *michelle*.

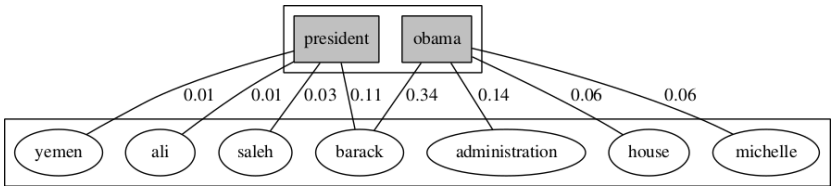

Figure 3.2: Example of document distances subgraph for CoFE ("President Obama"). Each n-gram is represented by the set of documents that contains the n-gram.

3.3 Distributed Representation-based Expansion (DREx)

DREx, the second method proposed and one of the main contributions of this work, defines a metric space $(\mathcal{V}_{DREx}, g_{DREx})$ that exploits the vector representation of words to expand short text documents. Vector representations of words allow objective comparison of document words regarding semantics. This is possible because the distance between two word vectors can be interpreted as a metric of semantic relationship between them.

Using the distributed representation of words defined in the section 2.4, we define a metric space $(\mathcal{V}_{DREx}, g_{DREx})$ that exploits these vectors to expand short text documents. For DREx, the points in the metric space, that become the nodes of the graph, can be either document bigrams or expansion words. Bigrams were chosen as they are better at capturing the document context than individual words.

The set $\mathcal{V}_{DREx}$ contains a vector representation $v_w \in \mathcal{V}_{DREx}$ for each word w of the vocabulary and each bigram $w_i - w_j$ of the documents in the original dataset. We use an external dataset with larger and richer documents to obtain the vector representation of words. For the bigrams, we exploit the arithmetic properties of vector representations and sum the word vectors for each bigram, so that each of them corresponds to an element in the metric space. Note that there are other ways to represent word vectors of bigrams, such as the average value of both vectors. However, as the arithmetic sum of vectors had already shown to be able to effectively capture document context, we opted for it.

To complete the definition of our metric space, we define the distance metric between word vectors, which is a modified version of the cosine distance [Zhang Korfhage, 1999] that satisfies all properties required by the metric distance g:

$$g_{DREx}(v_i, v_j) = 1 - \frac{\cos^{-1}(v_i \cdot v_j / \|v_i\| \, \|v_j\|)}{\pi} \tag{3.2}$$

Figure 3.3 presents a graph derived from metric space $(\mathcal{V}_{DREx}, g_{DREx})$. The original document contains the text "president obama visited cuba", corresponding to the bigrams: (i) president+obama, (ii) obama+visited, and (iii) visited+cuba, represented by rectangles in the graph. We also present, for each bigram, the neighborhood candidate

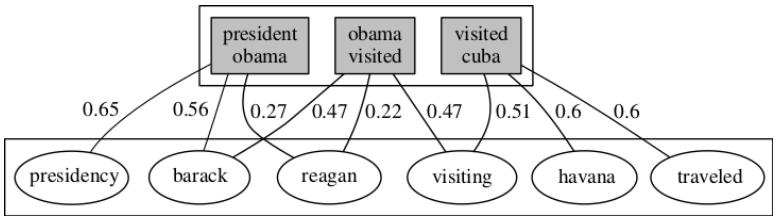

Figure 3.3: Example of expansion graph for DREx ("President Obama visited Cuba"). Note that each word and bi-gram is actually represented in DREx by their embedded vector.

words (the circles in the graph). Assume that the expanded pseudo-document considers $t = 3$ neighbor words for each word in the original document, and a maximum number of $M = 6$ (Scaling parameter $S = 2$) words in total after expansion. As in CoFE, the weights of each word are inversely proportional to their distance to the original bigrams in the metric space, and the words are probabilistic selected according to these weights. Following the example, the pseudo-document contains the original words: *president*, *obama*, *visited*, and *cuba*; and the selected words: *barack*, *presidency* and *visiting*.

3.4 Complexity Analysis

In their implementations, both CoFE and DREx have the function g of their metric spaces calculated and stored in a cache before the document expansion step starts. In terms of computational time complexity, let N be the number of documents in a dataset, V the vocabulary size, M the expected number of words in an expanded document (calculated according to the scaling factor S), T the number of closest neighbors per word stored in the cache and L the dimension of the generated word vectors.

CoFE's cache generation process creates an inverted index from the dataset considering that every document has every word in the vocabulary – $O(NV)$, and for each pair of words in the vocabulary, it calculates their Jaccard index - $O(NV^2)$. Therefore, CoFE's cache generation time complexity is of order $O(NV^2)$.

DREx's cache generation process, in turn, creates a vector representation for every bigram present in the dataset, which in the worst case is equals to $O(V^2)$ bigrams. For each bigram, its distance to the other V word vectors is calculated - $O(VL)$ calculations - and the T closest vectors are retrieved using a partial sorting of time complexity $O(V + T \log T)$. Therefore, DREx's cache generation time complexity is of order $O(V^3L + V^3 + V^2T \log T)$.

For the document expansion step, both methods are of order $O(NMT + NM) =$

$O(NMT)$. For each document, T candidate words are retrieved for every word in the document, which has length $M-1$ in the worst case, and at most $M-1$ words are selected to be part of the new pseudo-document generated. In general, the time complexity of both algorithms are dominated by the cost of the cache generation procedure, which is highly dependent on the vocabulary size.

Chapter 4

Experiments and Results

This chapter describes the set of experiments proposed to measure the effectiveness of the method for topic modeling. We first introduce the set of six datasets used to measure the effectiveness of the proposed method and describe the experimental setup, where methods parameters and the experimental methodology were defined. We then introduce the evaluation metrics used and the four phases of experimental analysis.

These four phases were: (i) the impact of the parameters to the method when used together with the traditional LDA; (ii) comparisons with baselines; (iii) performance of the method when used with other methods that not the LDA; (iv) a more indirect evaluation of the topic representation of documents in a classification task.

In the first phase, we assessed the impact of varying the expected lengths of documents (S), an important parameter of the system, for both CoFE and DREx run with LDA. We also looked at different numbers of topics and the performance of different word embedding methods. We then compared the results obtained in the first phase with other methods that also generate pseudo-documents to improve topic modeling for short text, namely WNTM, LDA-# and STE, introduced in Section 2.1. As LDA-# uses a tweet pooling scheme based on common hashtags to generate the pseudo-documents, it was run only for datasets where hashtags were available. STE expands documents with terms correlations based on PMI. Originally, the authors propose a threshold scheme based on the PMI score to determine which words should be added to the documents. Here, to make a fair comparison with CoFE and DREx, we changed the method to continually add new words to the document until it reaches a target size, controlled by the same parameter S.

In a third phase, we explored the results of the proposed expansion methods with other topic modeling besides LDA, namely LF-LDA and BTM. Note that although these methods were conceived to deal with short text, both authors argue they should perform well in datasets with larger text [Nguyen et al., 2015; Yan et al., 2013]. With that in mind, we compared the results obtained by the methods using the original short documents and those obtained when extracting the topics from the pseudo-documents.

Besides evaluating the methods using topic assessment metrics, we finally looked at the quality of topical document representation through a document classification task, where the short text comments had classes associated.

Table 4.1: Average and standard deviation for dataset features.

Dataset	N. of Docs	Vocab. Size	N. of Classes	Words per Document	Unique words per Document
TMN	30376	6314	7	4.9 (±1.5)	4.9 (±1.5)
NBA	70707	12504	-	8.6 (±3.0)	8.4 (±3.0)
Politics	70712	15029	-	8.1 (±2.6)	8.0 (±2.5)
20Nshort	1723	964	20	8.2 (±3.5)	7.1 (±2.9)
Sanders	3770	1311	4	6.1 (±2.7)	5.8 (±2.5)
Snippets	12117	4677	8	14.3 (±4.4)	10.3 (±3.1)

4.1 Datasets

To evaluate our proposed framework and the existent baselines, we compile a set of six real short-text document corpus, namely:

1. Tweets NBA (NBA): A sample of tweets about two NBA teams, Golden State Warrior and Los Angeles Lakers, collected from June to August 2015, using the hashtags #warriors and #lakers.

2. Tweets Politics (Politics): A sample of tweets mentioning Democrats and Republicans, collected from June to August 2015, using the hashtags #democrats and #republicans.

3. Tweets Sanders (Sanders): Tweets related to four different companies: Apple, Google, Microsoft, Twitter.[1]

4. 20 Newsgroups (20Nshort): A collection of newsgroup documents, partitioned across 20 different public newsgroups. We use only the documents with less than 21 words, as done in Nguyen et al. [2015].

5. Tag My News (TMN): A collection of English RSS news items grouped into 7 categories, where only the news titles are considered [Vitale et al., 2012].

6. Web Snippets (Snippets): A collection of web search snippets, which are summaries of documents presented as results of a query by a search engine [Phan et al., 2008]. The queries used are related to 8 different domains.

All datasets were preprocessed before the expansion step by making all the text lower-case, removing non-alphabetic characters and stop words. We also removed words shorter than 3 characters, and words appearing less than 10 times in 20Nshort and under 5 times in TMN and Twitter datasets.

[1] Available at `http://www.sananalytics.com/lab`.

Table 4.1 shows statistics for the datasets. Note that we have few words per document for all datasets (column w/doc). This is also true when considering only unique words per document (column unique w/doc), which ranges from 5.82 (Sanders) to 10.27 (Snippets).

4.2 Experimental Setup

This section describes all the parameter configurations used by the methods considered in our experiments, namely the word embeddings for DREx, the topic modeling algorithms and their parameters. Note that a in-depth analysis of parameters was only performed for the proposed method. For all other cases, we used parameters previously defined in the literature.

Word embeddings algorithms: For DREx, we first need to obtain the vector representation of words. As mentioned before, we use three different methods to extract the word vectors from the English Wikipedia dump from 06/02/1015: SG, CBoW and GloVe [Mikolov et al., 2013c; Pennington et al., 2014b]. Text data extracted from the Wikipedia XML produced a dataset with $8,102,107$ articles and a vocabulary of size $2,120,659$ (also the number of word vectors). Experiments used the original implementations of SG, CBoW and GloVe. All methods consider a context window of size 10. Both SG and CBOW used word vectors of size 300 and negative sampling (5 negative examples). Initial learning rate was set to 0.025 for SG and 0.05 for CBoW. For GloVe, we used the values suggested by the authors [Pennington et al., 2014a], fixing $x_{max} = 100$ and $\alpha = 0.75$.

Topic Modeling algorithms: Regarding the topic models, LDA, LF-LDA and BTM share four main parameters: the number of topics (k), the hyper-parameters α and β for the Dirichlet distribution and the number of sampling iterations. The values of α and β for LDA were estimated using Minka's fixed point iteration technique [Minka, 2000], and LDA was run for 2000 iterations. The number of topics assumed values 20, 50 and 100. LF-LDA has two extra parameters than the other methods: the word vectors representations and a mixture factor λ, which controls whether to use the Dirichlet or the latent feature component of the method. We use the default value of λ suggested by the authors (0.6) [Nguyen et al., 2015], and the word vectors learned from Wikipedia.

All experiments involving intrinsic evaluation of topics (i.e. all but the documents classification experiment) were repeated 5 times. In order to verify the statistical validity

of our conclusions and perform comparison between different methods, we used the non-parametric Wilcoxon signed-rank test with 0.05 of significance level over the means.

4.3 Evaluation Metrics

Topic evaluation is still an open problem and a large area of study in the topic modeling community [Wallach et al., 2009; Chuang et al., 2013; Morstatter et al., 2015]. Defining a quantitative metric to evaluate whether the topics represent the semantics of the documents considered is a challenge, and performing a qualitative analysis is too time-consuming. For these reason, we performed two types of evaluation: a direct one, that considers the Normalized Pointwise Mutual Information (NPMI)-score [Bouma, 2009] to evaluate the quality of the topics extracted by the methods; and an indirect one, where the topical representation of documents is considered within a classification framework.

Concerning direct evaluation of topics, two metrics are popular in the literature: topic coherence [Mimno et al., 2011] and the PMI-score. While the first performs the evaluation considering only the co-occurrence of words in the datasets were the topics are extracted from, the NPMI-score uses an external source for validation. Following recent works in the area, which give preference to PMI, our analysis focuses on this metric.

The PMI-score [Newman et al., 2010] verifies if the semantic relation between a pair of words suggested by a topic model is also found in an external dataset by evaluating the pointwise mutual information (PMI) of all pairs of its most probable words. The probabilities are evaluated by counting word co-occurrence frequencies in a 10-words sliding window in a large external dataset. Its normalized version was proposed by Bouma [2009], and removes the score sensibility to frequency and provides more intuitive score values: when w_i and w_j only occur together, $\text{NPMI}(w_i, w_j) = 1$; when they never occur together, $\text{NPMI}(w_i, w_j)$ is defined as -1.

Given a topic t and its ten most probable words W_{10}, NPMI-score is calculated as:

$$\text{NPMI-Score}(t; W_{10}) = mean\{\text{NPMI}(w_i, w_j), i, j \in 1...10, i \neq j\} \tag{4.1}$$

$$\text{NPMI}(w_i, w_j) = \left(\ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}\right) \Big/ - \ln p(w_i, w_j) \tag{4.2}$$

The external dataset used for evaluation consisted of a randomly generated sample of 15M documents in English from the WMT11 news corpus[2]. We used Palmetto's NPMI implementation [Röder et al., 2015].

[2]Available at `http://www.statmt.org/wmt11/training-monolingual.tgz`.

For the document classification evaluation we used the macro-average F1 score [Yang Liu, 1999], which is the mean F1 score of all classes. The F1 score for a class is the harmonic mean between the class precision and recall. The precision of a class c, $Precision_c$, is defined as the fraction of correct predictions for that class, and its recall, $Recall_c$, the fraction of instances of c that were correctly predicted, as shown by the following equations:

$$Precision_c = \frac{tpr_c}{tpr_c + fpr_c} \tag{4.3}$$

$$Recall_c = \frac{tpr_c}{tpr_c + fnr_c} \tag{4.4}$$

where tpr_c is the *true positive rate* of class c, fpr_c its *false positive rate* and fnr_c its *false negative rate*.

The F1 score of class c, $F1_c$, is defined according to the following equation:

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{4.5}$$

4.4 Impact of Parameters in CoFE and DREx

This section evaluates the performance of CoFE and DREx when the pseudo-documents generated are given as input to LDA, and analyzes the impact of the parameters k – the number of topics, and S – a scaling parameter that defines the final document length. For k, we tested the methods with 20, 50 and 100 topics. For S, we tested maximum document scaling sizes of 2, 3, 5, 7 and 10, meaning that the final length of the documents after the expansion phase is S times the original length. Due to the large number of experiments, Table 4.2 summarizes the values of NPMI by showing the average results over 20 experiments (4 values of M $\times$ 5 replications) for different numbers of topics. The complete tables are available in Appendix A. For DREx, we also analyzed the impact of different word vector representations (CBOW, SG and GloVe), which in the table presented in this section follow the name of the method. Values in bold indicate methods that are statistically significantly better than all other methods in the same column for that dataset according to a Wilcoxon signed-rank test.

Note that, with the exception of NBA and $N20short$, where DREx with GloVe and SG present no statistical difference, for all other datasets and configurations the results are consistent: DREx with GloVe is always superior regardless of the number of topics. These results indicate that the word vector representations generated by GloVe

Table 4.2: Average results of NPMI for CoFE and DREx run with LDA (mean of different values of expected document length).

	20 Topics	50 Topics	100 Topics	20 Topics	50 Topics	100 Topics
Topic Model	Tweets NBA			Tweets Politics		
CoFE	-0.165	-0.168	-0.175	-0.141	-0.156	-0.168
DREx-CBOW	-0.065	-0.091	-0.104	-0.043	-0.076	-0.089
DREx-GloVe	**-0.027**	**-0.03**	**-0.04**	**0.023**	**0.009**	**0.001**
DREx-SG	**-0.024**	**-0.024**	**-0.037**	-0.025	-0.044	-0.056
	Tweets Sanders			N20 Short		
CoFE	-0.098	-0.144	-0.152	-0.198	-0.209	-0.22
DREx-CBOW	-0.019	-0.046	-0.069	-0.072	-0.099	-0.126
DREx-GloVe	**0.022**	**0.003**	**-0.021**	**-0.003**	**-0.054**	**-0.097**
DREx-SG	0.002	-0.016	-0.041	**-0.02**	**-0.059**	**-0.097**
	TMN Title			Web Snippets		
CoFE	-0.116	-0.141	-0.165	-0.041	-0.063	-0.08
DREx-CBOW	-0.018	-0.028	-0.056	-0.002	-0.017	-0.037
DREx-GloVe	**0.032**	**0.036**	**0.0271**	**0.05**	**0.046**	**0.029**
DREx-SG	-0.018	-0.027	-0.04	0	0.006	-0.013

are more robust for generating pseudo-documents than those obtained by CBoW and SG for the task of topic modeling.

Comparing DREx's performance with CoFE's, DREx always performs better, no matter which word vector representation is used. We believe this happens because the word vectors were trained in Wikipedia, which has articles in several different subjects. This causes the word relations found to be richer and less context-specific when compared to CoFE, which obtains word similarity information only from the original dataset.

For the number of topics, we observe that NPMI values decrease when we increase the number of topics in all configurations. This behavior may be related to the nature of the datasets and, when available, their number of classes or contexts. Since the number of contexts varies from 2 (NBA, Politics) to 20 (N20short), a very high number of topics may indeed worsen the quality of inferred topics. Regarding the results of S, no clear pattern appears, but the best overall results were obtained by CoFE and DREx-GloVe with a scaling factor of 7, meaning that the final length of the documents after the expansion phase is seven times the original length (see Appendix 1).

In order to provide a better insight on how our proposed methods influence LDA, we performed a comparison with topics from models trained with the Web Snippets dataset expanded by DREx-GloVe – as it presented the best values of NPMI, and CoFE. Table 4.3 shows the most representative words from learned topics when the number of topics is set to 20. Topics were paired using cosine similarity in a greedy strategy. A good topic should be interpretable and reflect the dataset eight categories: (1) business, (2) computers, (3) culture, arts and entertainment (4) education and science, (5) sports, (6) politics and society, (7) engineering, and (8) health.

Note that topics for both methods can be easily labeled considering the dataset

Table 4.3: Topics discovered by LDA for Web snippets expanded by CoFE and DREx-GloVe. Column *class* indicates the respective dataset class the topic refers to. Column *Sim* indicates the cosine similarity for topics in the same row.

#	CoFE	DREx	class	sim
1	car engine electrical motor wheels electric cars gear fuel automatic	car engine cars equipment manufacturing electrical vehicle motor components	7	0.75
2	health cancer medical disease healthy nutrition information diet treatment hiv	health medical care treatment cancer patients disease patient medicine clinical	8	0.74
3	sports football games news game soccer com league team scores	sports football soccer league teams game basketball games sport team	5	0.74
4	intel computer memory chip processor device cpu cache core pentium	computer hardware computers intel software processor memory computing	2	0.72
5	political democracy party democratic social politics parties communist	political government politics party democratic election democracy	6	0.71
6	research edu science university school department graduate program students	university graduate edu faculty education college student students school harvard	4	0.70
7	news information online yahoo web directory com search sites links	information web online internet links external google search websites blog	2,4	0.65
8	business trade services management marketing gov development international	business industry financial market companies company investment finance	1	0.65
9	movie movies film imdb awards actor video director academy tom	music movie film video movies feature best films shows released	3	0.64
10	software programming computer web data java systems linux code parallel	computer software systems application applications internet information based	2	0.61
11	wikipedia encyclopedia wiki culture history article American ancient category	wikipedia articles article wiki https pages org page encyclopedia doesn	3,4	0.61
12	theory physics quantum philosophy theorem mathematical newton	theory theoretical analysis methods mathematical instance physics concepts	4	0.55
13	journal theoretical journals biology natural paper papers research theory evolution	research science study scientific technology studies institute development	4	0.40
14	music art rock band pop classical artists lyrics arts album	art work works gallery museum arts photo collection artist painting	3	0.40
15	amazon com books fashion online selection design book shopping manga	published book books publications journal publication literature work	4	0.35
16	system gov house government president presidential republic united congress	development public business government education economic information	6	0.27
17	war military navy force air army nuclear revolution civil weapons	culture history american world part united europe modern america first	3,4	0.17
18	tickets tennis golf ski buy chicago grand diego maradona woods	news media coverage cnn chicago broadcast bbc interview york washington	5/3	0.17
19	market stock finance financial exchange bank investment income quotes money	food health healthy diet nutrition calorie eating fitness eat foods	1/8	0.05
20	network internet security wireless bandwidth test mobile speed access	education students teaching learning school learn work help experience	2/4	0.04

document categories. For instance, topic 1 is related to engineering, topic 2 to health and topic 3 to sports. Exceptions are topics 11 and 15, where one of the methods generated topics hard to label. For topic 11, CoFE presents words related to education and culture, while it is hard to correctly define what major concepts DREx's words are related to. In topic 15, the opposite happens. DREx presents an easier topic to categorize.

Overall, words added by DREx in the document expansion step tend to be less context-specific when compared to CoFE's. Comparing two topics related to the same subject, one may find DREx's to have more nouns that represent concepts, such as 'hardware' in topic 4, while CoFE's topic would have more specific words, such as 'cpu' and 'intel'. Similarly, in topic 10 CoFE's words include 'java' and 'code', while DREx's includes 'application', 'internet' and 'information'.

4.5 Evaluating the Expanded Documents

Strategies that create pseudo-documents can potentially change the original meaning of the document when generating its expanded version. This may occur when the document becomes skewed or even random. The proposed methods can change the meaning of the documents by adding random or non-related words to it.

We performed a manual analysis of the pseudo-documents generated. Here we show the results for datasets *TMN* and *Web Snippets*. As documents in these datasets are labeled according to their subjects, one would expect that CoFE and DREx would add words related to these documents categories. Table 4.4 shows the list of words more frequently added to the documents of the respective dataset with category label *health* and *business*. Each word is followed by the frequency it was added to documents in that dataset.

Table 4.4: List of words most frequently added to documents labeled as *health* and *business*. For both CoFE and DREx, the scaling factor used was 7. Each word is followed by the number of times it was added to that class documents.

CoFE-7x	DREx-Glove-7x	CoFE-7x	DREx-Glove-7x
Web Snippets			
business		health	
business (1669)	business (1991)	health (1598)	health (1634)
services (1144)	information (1427)	information (979)	information (1187)
finance (1004)	financial (1261)	medical (961)	medical (1011)
financial (1002)	work (1253)	cancer (893)	life (988)
information (982)	market (1245)	gov (881)	care (951)
market (944)	development (1245)	disease (814)	research (910)
news (926)	industry (1228)	nih (729)	treatment (776)
trade (802)	part (1195)	treatment (728)	public (752)
resources (790)	management (1115)	news (721)	help (738)
gov (766)	finance (1086)	prevention (717)	medicine (654)
TMN			
business		health	
stocks (914)	brought (1636)	risk (414)	health (430)
prices (823)	due (1595)	cancer (393)	care (345)
sales (634)	including (1288)	study (357)	patients (266)
oil (604)	business (1142)	heart (291)	things (264)
profit (575)	financial (1002)	drug (271)	cancer (253)
rise (562)	money (960)	diabetes (260)	treatment (252)
euro (544)	market (872)	fda (226)	disease (239)
shares (513)	increase (845)	drugs (223)	working (230)
crisis (498)	investment (713)	health (219)	medical (229)
inflation (496)	working (659)	recipes (198)	provide (218)

As expected, both methods added words highly correlated with the categories, weakening the hypothesis of changing the original meaning of the documents. One major difference between CoFE and DREx regards the specificity of the words they add to documents.

The set of words added by DREx is also usually more general than the words added by CoFE. For example, considering the class of documents about *business*, for both datasets DREx add general words such as 'business', 'market' and 'financial' more often while CoFE diversified the words according to the dataset, adding words like 'stocks', 'oil', 'sales' more often for the dataset TMN and 'business', 'service', 'trade' for the Web Snippets dataset.

Something similar occurs to the *health* category where one can note more specific terms like 'cancer', 'nih', 'heart' and 'diabetes' added frequently by CoFE and more general words like 'public', 'life', 'care' and 'disease' added frequently by DREx. Regarding the frequency with which words where added, both methods behave similar with DREx adding the top words with a slightly higher frequency.

Finally, Table 4.5 also presents a few examples of the original TMN documents and the pseudo-documents generated by DREx-GloVe to illustrate the semantic agreement between the words in the original and expanded documents. We present one document for each of the eight categories of TMN. For all documents most of the words added by DREx were related to the document category and one can clearly distinguish the subject of the document by looking at the expansion words selected by DREx.

These results reinforce the hypothesis that both methods were able to expand the documents by adding relevant words to them and without changing their subjects, and thus their topics.

4.6 Comparison With Baselines

Results previously show that, when comparing the results of CoFE and DREx run with LDA, DREx presents the best results of NPMI while expanding the documents using the GloVe vector representation. We also showed that using DREx with GloVe and a scaling factor of 7 has led to pseudo-documents and topics characterized by less context-specific words. This section compares this configuration of our method with LDA with other approaches previously proposed to generate pseudo-documents for topic modeling in short text, and also state-of-the-art methods for topic modeling in short text. The first comparison is the one we consider our true baseline, while the second, besides showing the method generality, also shows its superior performance regardless of the topic model considered.

Table 4.5: Examples of Tag My News documents expanded by DREx-GloVe-7x. The last column shows the selected words during the document expansion step.

Class	**Original text**	**Words added by DREx**
Business	delta air lines q1 loss grows to $318 million	due line force growing grow flight lost base operations service grown caused losing forces light aircraft reaches connection
Entertainment	”like a rolling stone" dylan’s best song	album rock songs bob written tribute track found back love singer rolling recording band music cover released tune live inspired beatles version girl release
Health	fda to regulate e-cigarettes as tobacco products	food consumer drugs sell sale trade goods regulation smoking alcohol brand brands coffee sugar marijuana export produce foods sold selling regulators smoke restricting affect reduce consumers control drug increase products
Science & Technology	apple co-founder wozniak: computers can teach kids	computer working learn students teaching lessons ceo founders learning teachers technology early company bring continue make program teacher based named programs education things business science skills dedicated modern time chairman
Sports	nadal cruises past ljubicic into quarters	semis final round federer losing lost finals berdych open quarterfinals draw tournament madrid recent future back time day current coming including year years days made present earlier taking brought continue make making cruise djokovic spain trip roddick
US	arizona supreme court stays execution	case judge appeals courts law stay justice states takes appeal leave rest cases staying months remain united time finally makes attorney put days close longer leaving held decides federal ruling

4.6.1 Comparison with document expansion methods

We compare the performance of DREx with LDA-# [Mehrotra et al., 2013], WNTM [Zuo et al., 2015] and STE [Pinto et al., 2011], all introduced in Section 2.2. Recall that LDA-# generates pseudo-documents by grouping Twitter hashtags when they are available. WNTM, in turn, generates pseudo-documents using the word co-occurrence network and STE expands documents with terms correlations based on PMI. As previously mentioned, we adapted STE to expand the documents until they achieved a target size controlled by a scaling factor S (same used by CoFE and DREx), to make comparisons fair with the proposed methods. The value of S used is the same chosen for CoFE and DREx: $S = 7$. We also show the results of LDA when run with the original documents,

Table 4.6: Results of NPMI for methods that generate pseudo-documents.

	20 topics	50 topics	100 topics	20 topics	50 topics	100 topics
Topic Model	**Tweets NBA**			**Tweets Politics**		
LDA - Original	-0.158	-0.156	-0.154	-0.072	-0.090	-0.095
LDA-DREx-GloVe-7x	**-0.014**	**-0.021**	**-0.023**	**0.040**	**0.024**	**0.012**
LDA-Hashtag	-0.158	-0.151	-0.153	-0.124	-0.116	-0.117
WNTM	-0.135	-0.141	-0.135	-0.086	-0.089	-0.099
STE	-0.087	-0.070	-0.069	-0.043	-0.045	-0.055
	Tweets Sanders			**20-News Short**		
LDA - Original	-0.087	-0.099	-0.116	-0.184	-0.188	-0.193
LDA-DREx-GloVe-7x	**0.031**	**0.011**	**-0.012**	**0.012**	**-0.046**	**-0.091**
WNTM	-0.085	-0.113	-0.125	-0.194	-0.194	-0.198
STE	-0.156	-0.164	-0.170	-0.232	-0.241	-0.239
	TMN			**Web Snippets**		
LDA - Original	-0.062	-0.056	-0.085	-0.061	-0.102	-0.106
LDA-DREx-GloVe-7x	**0.032**	**0.046**	**0.044**	**0.050**	**0.051**	**0.034**
WNTM	-0.026	-0.047	-0.067	0.004	-0.034	-0.064
STE	-0.245	-0.217	-0.220	-0.115	-0.141	-0.153

as a reference for comparison[3].

Table 4.6 shows the values of NPMI obtained by each method. Notice that the results of LDA-# are only available for the Twitter datasets *Politics* and *NBA*. For *Politics*, the 70712 original documents were grouped into 4184 pseudo-documents and the 70702 documents of *NBA* into 3924 pseudo-documents. Note that, after the pseudo-document generation, the total number of documents in the collections decreased drastically. Since the number of documents is as important as their size to the success of topic modeling techniques Tang et al. [2014], this reduction may impact negatively on the results found by LDA-#. For *Sanders*, information about hashtags was not available.

Observe that the results obtained by LDA-DREx were statistically significantly better than those obtained by all baselines in all datasets. Note that LDA-# does not even perform better than LDA with the original documents, while STE and WNTM showed improvements over LDA for two and three out of the six datasets, respectively.

In summary, considering the datasets used in our experiments, previously proposed methods that generate pseudo-documents did not even improve the results of LDA with the original datasets in a large number of cases, while the results obtained by LDA-DREX were statistically significant better than those obtained by both LDA and the two baselines in all cases. This shows the robustness of combining word vector representations trained in external datasets to generate improved larger pseudo-documents.

[3]We implemented the referred LDA-#, WNTM, STE and LDA (with hyperparameter optimization) methods according to their description in the original papers

Table 4.7: NPMI values for LDA, LF-LDA and BTM methods with 20 topics considering both the original and expanded versions (DREx-GloVe) of the dataset. Improvements of expansion are in parenthesis.

Dataset	Original	DREx-GloVe	Original	DREx-GloVe
	Tweets NBA		**Tweets Politics**	
LDA	-0.158	**-0.037 (76.58%)**	-0.072	**0.024 (133.33%)**
LF-LDA	-0.149	**-0.014 (90.60%)**	-0.059	**0.027 (145.76%)**
BTM	-0.168	**-0.036 (78.57%)**	-0.085	**0.022 (125.88%)**
	Tweets Sanders		**20-News short**	
LDA	-0.087	**0.047 (154.02%)**	-0.184	**0.009 (104.89%)**
LF-LDA	-0.079	**0.055 (169.62%)**	-0.179	**0.019 (110.61%)**
BTM	-0.085	**0.038 (144.71%)**	-0.202	**0.005 (102.48%)**
	TMN		**Web Snippets**	
LDA	-0.062	**0.056 (190.32%)**	-0.061	**0.061 (200.00%)**
LF-LDA	-0.039	**0.055 (241.03%)**	-0.061	**0.069 (213.11%)**
BTM	-0.048	**0.070 (245.83%)**	-0.042	**0.082 (295.24%)**

4.6.2 Comparison with other topic models developed for short text

The previous section showed DREx's pseudo-documents generate better topics than those created by other expansion methods. This section, in contrast, compares the results of DREx with methods that have changed the LDA model to overcome the problems of short-text scenarios. Two of the main representatives of this category are LF-LDA[4] and BTM[5] (see Section 2 for details). It is important to emphasize that their authors claim these methods can also be used to learn topics in datasets with larger text. Considering this and the fact that our expansion framework can be used by any topic modeling algorithm, we compare the performance of these methods using the original version of the datasets and the expanded version generated by DREx-GloVe.

Table 4.7 shows the values of NPMI obtained by each method followed by the percentage of improvement over the use of the original dataset. Because of the high time complexity of these methods, we only conduct experiments with 20 topics, as previous experiments showed NPMI degraded as we increased the number of topics.

We observed that, for all topic models and datasets, the pseudo-documents generated by DREx were able to improve the quality of the topics learned. The improvements range from 76% (LDA on Tweets NBA) to 295% (BTM on Web Snippets). However, note that these very high values of improvement occur because the values of NPMI in the original dataset were really low.

We also highlight the fact that LDA-DREx performs better than BTM and LF-

[4]Implementation available at: `https://github.com/datquocnguyen/LFTM` (2016/12/16)
[5]Implementation available at: `https://github.com/xiaohuiyan/BTM` (2016/12/16)

LDA with the original datasets in all cases, showing that our expansion framework can be used to overcome the sparsity problem of the short-text scenarios without the need of a specific method. However, note that the best overall results were obtained when using the expanded dataset with either BTM or LF-LDA. BTM presented the best results for the datasets *TMN* and *Web Snippets*, while LF-LDA was the best in the four remaining datasets.

4.7 Evaluation Under a Classification Task

So far all experiments evaluated the topics generated with the original and expanded datasets according to NPMI. This section, in contrast, evaluates the representative power of topics to classify documents into different categories. From the six datasets considered, four were manually labeled and took part of this experiment, namely *2*0-News short, *T*MN, *S*anders and *W*eb Snippets. The number of categories of each dataset is shown in Table 4.1.

In this experiment, each document was represented by its posterior topical distribution instead of their words. Therefore, the feature set f_i of a document d_i is defined as:

$$f_i = [p(z_1|d_i), p(z_2|d_i), ..., p(z_k|d_i)]$$

Two different datasets were generated for each of the topic models considered: the first includes the topics extracted by the method from the original dataset, and the second the topics generated from the expanded dataset generated by DREx -GloVe. In order to compare the results of the classification task with those obtained when using the NPMI metric, both LDA, LF-LDA and BTM were used as topic models. All experiments used 20 topics as document features (i.e. $|f_i| = 20$). These datasets were given as input to a SVM classifier with a Gaussian kernel[6] to classify the documents. The documents classification experiment was performed using 5 executions of a 5-fold cross-validation (1 fold for tuning SVM parameters with a grid search, 3 folds for training and 1 fold for test), in a total of 25 repetitions for each configuration. Multiclass classification was performed using the one-against-all strategy.

Table 4.8 shows the results of the mean macro-average F1 score comparing pairs of original and expanded datasets for each topic modeling technique, i.e., LDA, BTM and LF-LDA. Bold values indicate statistically significant best versions for each pair comparison.

[6]We used the R wrapper (package "e1071") for the libSVM library (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`).

Results of F1 show that the proposed expansion method improves the quality of document representation (i.e. topics) from the perspective of document classification. With the exception of 20Nshort using LF-LDA, all other mean results revealed improvements of up to 15.4%. When compared to the original dataset, a total of 8 out of 12 results (in bold) show that classification results with expanded datasets are statistically significantly better than those obtained with the topics extracted from the original datasets. The other four results show no evidence of statistical difference.

Table 4.8: Classification results of f-measure when representing documents by topics extracted from original and expanded documents. Improvements of expansion are in parenthesis.

	LDA		BTM		LF-LDA	
	Original	DREx-GloVe	Original	DREx-GloVe	Original	DREx-GloVe
20Nshort	0.216	**0.24 (+11.1%)**	0.252	0.267 (+5.8%)	0.239	0.235 (-1.6%)
TMN	0.599	0.62 (+3.4%)	0.652	**0.689 (+5.7%)**	0.618	0.624 (+0.9%)
Sanders	0.842	**0.901 (+6.9%)**	0.88	**0.924 (+4.9%)**	0.852	**0.899 (+5.5%)**
Snippets	0.757	**0.836 (+10.4%)**	0.857	**0.872 (+1.7%)**	0.729	**0.841 (+15.4%)**

Distributions of topics extracted from BTM and LF-LDA with the original dataset and used as features for short text classification revealed to be natural good predictors of classes. These results are consistent with the obtained in Yan et al. [2013]. However, DREx further enhances classification performance, with statistically significant improvements that vary from 1.7% to 15.4%.

These results also reinforce the independence of the expansion method from the subjacent topic modeling techniques and the potential for consistent improvement. Furthermore, intrinsic evaluation of topics, like NPMI, could raise the hypothesis of eventual misrepresentation of the original latent topic structures by the expansion procedure (even when good results are obtained). The results shown undermine this hypothesis, since if it were true this would worsen classification performance, once classes carry information about topics. Instead, classification performance was systematically improved.

In summary, we observed that for intrinsic evaluation of topics with NPMI, the general best method in our results is LF-LDA-DREx using the GloVe representation. On the other hand, for the task of document classification, BTM-DREx with GloVe showed to be the best method for all datasets. This means that although there are more appropriate topic modeling methods for particular tasks (e.g. BTM for document classification), our proposed expansion procedure can generally be applied to improve the quality of topics extracted from short text.

Chapter 5

Conclusion and Future Work

In this work we addressed the problem of topic modeling on short-text scenarios and its challenges. A literature review was conducted, where we presented the most representative works on this field and discussed their limitations. Knowing the existent works and their limitations, we proposed a framework for generating large pseudo-documents based on the definition of metric spaces. Metric spaces are powerful tools to quantitatively measure the relationship between two elements (e.g., n-grams of words) and we explored this property to propose a general framework that can be instantiated in different ways, according to the metric space used. Here, we presented two instances, namely CoFE and DREx, and showed their robustness and effectiveness in topic modeling for short text. While CoFE uses a simple word co-occurrence to expand the original text, DREx relies on distributed representation of word vectors calculated on an external dataset.

The methods were evaluated in six datasets and compared to other state-of-the-art approaches for topic modeling in short text documents. We first compared CoFE and DREx against each other and with other algorithms for pseudo-documents generation. The results showed that DREx outperforms all methods, achieving higher values of NPMI in all datasets. In a second phase, the collections expanded with DREx were given as inputs to LDA, BTM and LF-LDA, the last two being methods specifically designed to learn topics from short-texts. Finally, the methods were also evaluated in a document classification task.

Overall, the experiments performed indicated that datasets expanded with DREx using the GloVe word representation improved significantly the results of all topic modeling methods, both in terms of NPMI and when performing text classification, showing that the proposed expansion framework can be used to enhance the quality of the topics found by any topic modeling algorithm. For this reason, we recommend the use of this approach when dealing with short text and topic modeling.

Following, we discuss some of directions of promising future directions.

Hybrid DREx and CoFE metric space: Although DREx achieved better results of NPMI than CoFE, Tables 4.4 and 4.3 indicate that the words added by CoFE are also relevant to the documents, with the difference that DREx added more general words

while CoFE focused on more specific ones. So far, we only use NPMI to evaluate the quality of the topics found by our methods, and as discussed in Section 4.3, NPMI uses the co-occurrence of the topic most relevant words on an external and larger dataset, to asses its quality. Hence, since DREx uses a metric space based on distributed vector representations of words learned on an external dataset, it is natural that DREx outperform CoFE for this metric. However, note that the dataset use to generate the word vectors is different from the one used to evaluate NPMI.

Another metric well used to evaluate the quality of topics is topic coherence [Mimno et al., 2011]. The topic coherence metric evaluates the topic quality by looking at the co-occurrence of the most probable words for the topic in the original dataset. Given a topic t, its 10 most probable words W_{10}, $D(w_i, w_j)$ being the co-document frequency of words w_i and w_j and $D(w_j)$ the document frequency of word w_j, the coherence score for t is:

$$coherence(t; W_{10}) = \sum_{i=2}^{10} \sum_{j=1}^{i-1} log \frac{D(w_i, w_j) + 1}{D(w_j)} \tag{5.1}$$

Note that coherence is similar to NPMI (see Equation 4.2), but is calculated using the same dataset used by the topic model algorithm, which leads to an evaluation that takes into account the specific content of each dataset. A preliminary result of both CoFE and DREx using the topic coherence metric is presented in Table 5.1 and shows that, as expected, CoFE outperform DREx for this metric. The intent of this section is not to argue that NPMI is better then Coherence or otherwise. As previously mentioned, the automatic evaluation of topic modeling methods is an open problem on the literature [Wallach et al., 2009; Chuang et al., 2013; Morstatter et al., 2015].

Table 5.1: Average results of Topic Coherence for CoFE and DREx run with LDA (mean of different values of expected document length).

	20 Topics	50 Topics	100 Topics	20 Topics	50 Topics	100 Topics
	Tweets NBA			Tweets Politics		
CoFE	-107.8	-101.5	**-94.7**	**-106.0**	**-96.0**	**-88.8**
DREx-CBOW	-140.8	-129.4	-118.9	-140.1	-130.0	-122.6
DREx-GloVe	-134.7	-132.0	-126.6	-149.6	-142.4	-135.4
DREx-SG	**-88.0**	**-95.2**	**-94.9**	**-108.5**	-108.0	-103.9
	Tweets Sanders			N20 Short		
CoFE	**-107.5**	**-91.7**	**-85.6**	**-89.1**	**-82.6**	**-79.9**
DREx-CBOW	-124.1	-119.5	-114.9	-113.4	-107.5	-102.5
DREx-GloVe	-132.1	-126.1	-120.6	-119.9	-111.4	-104.1
DREx-SG	-114.9	-113.1	-111.0	-112.4	-107.7	-103.3
	TMN Title			Web Snippets		
CoFE	**-160.6**	**-149.7**	**-140.6**	**-132.1**	**-124.3**	**-116.2**
DREx-CBOW	-170.8	-162.3	-156.9	-148.5	-143.6	-140.4
DREx-GloVe	-181.8	-175.7	-171.0	-149.3	-143.9	-142.0
DREx-SG	-167.0	-161.3	-156.0	-145.1	-139.9	-136.8

Given this disagreement between the two metrics, and the fact that the framework proposed can be instantiated in several different ways, the most natural sequence of this

work would be explore other metric spaces that combine the metric spaces defined by DREx and CoFE.

As detailed in Section 2.3 both metric spaces are based on vector representations of words, with CoFE vectors been calculated on the original dataset using the Jaccard coefficient and DREx vectors been calculated on an external richer dataset using modern techniques of distributed representation of words.

One way to easily combine both metric spaces is to define a distance function g that considers both vector representations using, for example, a convex combination of the distances functions already defined by DREx and CoFE. The idea is to provide a new metric space which has a good commitment with the generalization of DREx and specificity of CoFE.

Non-probabilistic topic modeling: Another future work we intent to perform is regarding the nature of the topic modeling algorithms. As discussed in Chapter 2, there are two categories of topic modeling algorithms: probabilistic and the non-probabilistic. So far, we only present the results of probabilistic methods, since they are considered state-of-the art for this problem, but the framework proposed here can also be used by the non-probabilistic methods.

An advantage of non-probabilistic methods might be that they can explore directly the similarity between a document and the candidate terms. In its fourth step, our framework selects the top m most similar n-grams to expand a given document. This n-grams are appended to the end of the document, and the similarity score calculated in the previous step ignored. This is necessary because probabilistic topic modeling algorithms works by counting the frequency of the n-grams in the documents and their relations, and cannot deal with decimal scores. Non-probabilistic methods, in contrast, can use the similarity score directly in their input. This is possible because, non-probabilistic methods such as NMF, use as input a term-document matrix V (as showed in Figure 2.4) where the value in each cell $V_{i,j}$ is the weight of term i to the document j. This weight can represented by any value, integer or decimal (for NMF the only restriction is that it has to be positive).

We conduct a preliminary experiment based on this idea. We follow the same setup used in Section 4.6, using a scaling factor of 2, 3, 5, 7 and 10 but instead of adding the top n-grams to a document we use the similarity score of these n-grams in the input matrix V. The preliminary results are presented in Table 5.2, where the results obtained by NMF using the original version of the datasets are compared to the results obtained by NMF using the matrix V expanded by DREx. The last column shows the average improvement of NMF with our expanded input matrix over the original one. The results are quite promising, showing this as another interesting direction of future work.

Table 5.2: Preliminary results of DREx-Glove run with NMF

3

Scaling Factor	Expanded Matrix with DREx					Original Matrix	Average Improvement
	2x	3x	5x	7x	10x		
Tweets Sanders	0.018	0.017	0.021	0.014	0.003	-0.126	101.47%
Web Snippets	0.019	0.015	0.010	-0.004	-0.014	-0.127	100.52%
N20 Short	0.001	-0.007	-0.006	-0.008	-0.021	-0.193	99.20%
TMN Title	0.018	0.027	0.028	0.033	0.036	-0.145	102.86%
Tweets NBA	-0.011	-0.007	-0.006	0.001	-0.006	-0.167	99.44%
Tweets Politics	0.020	0.016	0.018	0.011	0.001	-0.082	101.33%

LDA modification: Although the modification of the input data is a simpler and general approach than directly modify the existent methods for topic modeling, another idea is to incorporate the ideas of the framework proposed here into an already existing method. For example, LDA infers the topics through the generative process describe in Section 2.1. This generative process is an imaginary process that LDA considers to be responsible for creating each document in the corpora. The only assumption made by LDA during the document creation is that each word in the document is associated to a topic and is draw from the topic multinomial distribution over the vocabulary. A simple modification on this process could be the inclusion of DREx or CoFE during the selection of the next word. This can be done by looking at the words already included in the document, and calculating their similarities with other candidate words.

Application in different tasks: The designed framework and its specializations proposed here can also be used to expand documents for other tasks that are also sensitive to short text or that can be improved over longer texts. Some example of such tasks, presented in section 2.3, are automatic query expansion and general text mining tasks, such as document clustering and document classification. In Section 4.7 we have already showed that DREx was able to improve the quality of a SVM classifier when compared to the original dataset, and further investigation on this matter is also left for future work.

Appendix A

Complete Results of DREx and CoFE

In Section 4.4 we discuss the impact of the parameters of DREx and CoFE. Due to the large number of experiments we only present in Table 4.2 the average NPMI results for these methods when we vary the scaling factor parameter S. This parameter controls the final size of the documents after the expansion, and larger values means larger final documents. We tested scaling factors of 2, 3, 5, 7 and 10.

In this appendix we present the complete results for each value of S in Tables A.1, A.2,A.3 and A.4. Each table shows the results for one method, CoFE, DREx-CBOW, DREx-GloVe and DREx-SG, respectively. The last line shows the average value (considering all datasets and number of topics) achieved by each method for each scaling factor.

During the analysis of the values, we could not identify a pattern able to explain the variation of the results according to the scaling factor. The best values differs for each method, each dataset and each number of topics.

Considering only the average values, we highlight the fact that DREx-GloVe with a scaling factor of *7x* achieved the best overall results. Hence, this value was chosen to be the value used in the other experiments of this dissertation.

Table A.1: Complete results of NPMI for CoFE run with LDA.

Scaling Factor	2x	3x	5x	7x	10x
20 Topics					
Tweets Sanders	-0.023	-0.125	-0.116	-0.118	-0.111
Web Snippets	-0.083	-0.057	-0.032	-0.024	-0.008
N20 Short	-0.216	-0.207	-0.191	-0.189	-0.186
TMN Title	-0.169	-0.136	-0.104	-0.086	-0.087
Tweets NBA	-0.167	-0.165	-0.164	-0.162	-0.167
Tweets Politics	0.129	-0.132	-0.148	-0.152	-0.146
50 Topics					
Tweets Sanders	-0.155	-0.148	-0.142	-0.137	-0.137
Web Snippets	-0.101	-0.081	-0.057	-0.045	-0.031
N20 Short	-0.217	-0.212	-0.209	-0.206	-0.200
TMN Title	-0.170	-0.152	-0.142	-0.130	-0.111
Tweets NBA	-0.170	-0.170	-0.167	-0.167	-0.169
Tweets Politics	-0.154	-0.158	-0.165	-0.151	-0.153
100 Topics					
Tweets Sanders	-0.158	-0.154	-0.151	-0.148	-0.149
Web Snippets	-0.107	-0.085	-0.074	-0.067	-0.069
N20 Short	-0.219	-0.222	-0.221	-0.220	-0.218
TMN Title	-0.195	-0.173	-0.157	-0.151	-0.147
Tweets NBA	-0.170	-0.173	-0.177	-0.174	-0.180
Tweets Politics	-0.169	-0.170	-0.170	-0.169	-0.162
Average	-0.154	-0.151	-0.144	-0.139	-0.135

Table A.2: Complete results of NPMI for DREx run with LDA and CBOW vectors.

Scaling Factor	2x	3x	5x	7x	10x
20 Topics					
Tweets Sanders	-0.023	-0.021	-0.019	-0.012	-0.023
Web Snippets	-0.005	0.002	-0.005	0.006	-0.007
N20 Short	-0.070	-0.055	-0.075	-0.072	-0.086
TMN Title	-0.027	-0.004	-0.009	-0.021	-0.029
Tweets NBA	-0.079	-0.066	-0.055	-0.059	-0.065
Tweets Politics	-0.040	-0.033	-0.033	-0.051	-0.060
50 Topics					
Tweets Sanders	-0.053	-0.043	-0.042	-0.045	-0.045
Web Snippets	-0.024	-0.013	-0.015	-0.017	-0.018
N20 Short	-0.104	-0.095	-0.094	-0.100	-0.104
TMN Title	-0.034	-0.024	-0.029	-0.027	-0.027
Tweets NBA	-0.105	-0.096	-0.083	-0.082	-0.089
Tweets Politics	-0.074	-0.079	-0.076	-0.079	-0.070
100 Topics					
Tweets Sanders	-0.075	-0.069	-0.067	-0.069	-0.065
Web Snippets	-0.038	-0.030	-0.034	-0.039	-0.045
N20 Short	-0.135	-0.123	-0.121	-0.126	-0.124
TMN Title	-0.074	-0.051	-0.048	-0.051	-0.054
Tweets NBA	-0.118	-0.104	-0.099	-0.099	-0.100
Tweets Politics	-0.098	-0.092	-0.085	-0.086	-0.083
Average	-0.065	-0.055	-0.055	-0.057	-0.061

Table A.3: Complete results of NPMI for DREx run with LDA and GloVe vectors.

Scaling Factor	2x	3x	5x	7x	10x
20 Topics					
Tweets Sanders	0.000	0.028	0.025	0.031	0.026
Web Snippets	0.039	0.061	0.051	0.050	0.048
N20 Short	-0.035	-0.017	0.009	0.012	0.013
TMN Title	0.028	0.027	0.039	0.032	0.036
Tweets NBA	-0.046	-0.033	-0.018	-0.014	-0.024
Tweets Politics	-0.005	0.014	0.031	0.040	0.038
50 Topics					
Tweets Sanders	-0.016	-0.001	0.010	0.011	0.012
Web Snippets	0.030	0.050	0.055	0.051	0.046
N20 Short	-0.070	-0.054	-0.049	-0.046	-0.050
TMN Title	0.005	0.039	0.045	0.046	0.044
Tweets NBA	-0.063	-0.030	-0.021	-0.021	-0.015
Tweets Politics	-0.020	0.004	0.021	0.024	0.019
100 Topics					
Tweets Sanders	-0.047	-0.020	-0.016	-0.012	-0.010
Web Snippets	0.020	0.031	0.035	0.034	0.028
N20 Short	-0.113	-0.094	-0.090	-0.091	-0.097
TMN Title	-0.011	0.017	0.039	0.044	0.046
Tweets NBA	-0.072	-0.049	-0.030	-0.023	-0.023
Tweets Politics	-0.027	-0.004	0.010	0.012	0.011
Average	-0.023	-0.002	0.008	0.010	0.008

Table A.4: Complete results of NPMI for DREx run with LDA and Skip-Gram vectors.

Scaling Factor	2x	3x	5x	7x	10x
20 Topics					
Tweets Sanders	0.001	0.002	0.006	0.002	-0.001
Web Snippets	-0.010	-0.008	0.010	0.006	0.005
N20 Short	-0.033	-0.015	-0.014	-0.015	-0.024
TMN Title	-0.035	-0.026	-0.020	-0.009	0.002
Tweets NBA	-0.056	-0.025	-0.009	-0.013	-0.016
Tweets Politics	-0.031	-0.018	-0.013	-0.034	-0.032
50 Topics					
Tweets Sanders	-0.027	-0.012	-0.015	-0.017	-0.012
Web Snippets	-0.009	0.011	0.010	0.013	0.005
N20 Short	-0.074	-0.065	-0.050	-0.054	-0.054
TMN Title	-0.061	-0.039	-0.018	-0.008	-0.009
Tweets NBA	-0.064	-0.029	-0.013	-0.008	-0.004
Tweets Politics	-0.070	-0.056	-0.036	-0.030	-0.027
100 Topics					
Tweets Sanders	-0.057	-0.036	-0.040	-0.040	-0.033
Web Snippets	-0.015	-0.010	-0.006	-0.012	-0.019
N20 Short	-0.112	-0.100	-0.090	-0.091	-0.092
TMN Title	-0.081	-0.042	-0.030	-0.023	-0.024
Tweets NBA	-0.073	-0.050	-0.027	-0.020	-0.015
Tweets Politics	-0.090	-0.066	-0.048	-0.041	-0.034
Average	-0.050	-0.032	-0.022	-0.022	-0.021

Bibliography

Bicalho, P. Pita, M. Pedrosa, G. Lacerda, A. Pappa, G. L. (2017). General framework to expand short text for topic modeling. .

Blei, D. M. Ng, A. Y. Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3:993--1022.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. GSCL, 31--40.

Cao, Z. Li, S. Liu, Y. Li, W. Ji, H. (2015). A novel neural topic model and its supervised extension. Bonet, B. Koenig, S., , *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 2210--2216. AAAI Press.

Carpineto, C. Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Comput. Surv., 44(1):1:1--1:50. ISSN 0360-0300.

Chuang, J. Gupta, S. Manning, C. D. Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. ICML (3), 612--620.

Deerwester, S. C. Dumais, S. T. Landauer, T. K. Furnas, G. W. Harshman, R. A. (1990). Indexing by latent semantic analysis. JASIS, 41(6):391--407.

Glorot, X. Bordes, A. Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. Getoor, L. Scheffer, T., , *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 513--520. Omnipress.

Griffiths, T. L. Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1):5228--5235.

Hong, L. Davison, B. D. (2010). Empirical study of topic modeling in twitter. Proceedings of the first workshop on social media analytics, 80--88. ACM.

Hörster, E. Lienhart, R. Slaney, M. (2007). Image retrieval on large-scale image databases. CIVR, 17--24.

Hotho, A. Staab, S. Stumme, G. (2003). Ontologies improve text document clustering. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, 541--544. IEEE.

Jin, O. Liu, N. N. Zhao, K. Yu, Y. Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, 775--784, New York, NY, USA. ACM.

Lee, D. D. Seung, H. S. (2001). Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 556--562.

Lin, C. X. Zhao, B. Mei, Q. Han, J. (2010). Pet: A statistical model for popular events tracking in social communities. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, 929--938, New York, NY, USA. ACM.

Liu, Y. Liu, Z. Chua, T. Sun, M. (2015). Topical word embeddings. Bonet, B. Koenig, S., , *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 2418--2424. AAAI Press.

Manning, C. D. Schütze, H. (1999). Foundations of statistical natural language processing, volume 999. MIT Press.

Mehrotra, R. Sanner, S. Buntine, W. Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. SIGIR, 889--892.

Mikolov, T. Chen, K. Corrado, G. Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. NIPS, 1--9.

Mikolov, T. Corrado, G. Chen, K. Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. ICLR, 1--12.

Mikolov, T. Sutskever, I. Chen, K. Corrado, G. S. Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. Burges, C. J. C. Bottou, L. Ghahramani, Z. Weinberger, K. Q., , *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, 3111--3119.

Mikolov, T. tau Yih, W. Zweig, G. (2013d). Linguistic Regularities in Continuous Space Word Representations. PNAACL-HLT.

Mimno, D. Wallach, H. M. Talley, E. Leenders, M. McCallum, A. (2011). Optimizing semantic coherence in topic models. EMNLP, 262--272.

Minka, T. (2000). Estimating a dirichlet distribution.

Minka, T. Lafferty, J. (2002). Expectation-propagation for the generative aspect model. Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, 352--359. Morgan Kaufmann Publishers Inc.

Morstatter, F. Pfeffer, J. Mayer, K. Liu, H. (2015). Text, topics, and turkers: A consensus measure for statistical topics. Proceedings of the 26th ACM Conference on Hypertext & Social Media, 123--131. ACM.

Newman, D. Noh, Y. Talley, E. Karimi, S. Baldwin, T. (2010). Evaluating topic models for digital libraries. JCDL, 215--224.

Nguyen, D. Q. Billingsley, R. Du, L. Johnson, M. (2015). Improving topic models with latent feature word representations. TACL, 3:299--313.

Pal, A. Herdagdelen, A. Chatterji, S. Taank, S. Chakrabarti, D. (2016). Discovery of topical authorities in instagram. Proceedings of the 25th International Conference on World Wide Web, 1203--1213. International World Wide Web Conferences Steering Committee.

Pedrosa, G. Pita, M. Bicalho, P. Lacerda, A. Pappa, G. L. (2016). Topic modeling for short texts with co-occurrence frequency-based expansion. Proc. of the Brazilian Conference on Intelligent Systems (BRACIS). IEEE.

Pennington, J. Socher, R. Manning, C. D. (2014a). GloVe: Global Vectors for Word Representation. EMNLP, 1532--1543.

Pennington, J. Socher, R. Manning, C. D. (2014b). Glove: Global vectors for word representation. Moschitti, A. Pang, B. Daelemans, W., , *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1532--1543. ACL.

Phan, X. H. Nguyen, C. Le, D. Nguyen, M. L. Horiguchi, S. Ha, Q. (2011). A hidden topic-based framework toward building applications with short web documents. IEEE Trans. Knowl. Data Eng., 23(7):961--976.

Phan, X.-H. Nguyen, L.-M. Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceedings of the 17th International Conference on World Wide Web, WWW '08, 91--100, New York, NY, USA. ACM.

Pinoli, P. Chicco, D. Masseroli, M. (2014). Latent dirichlet allocation based on gibbs sampling for gene function prediction. IEEE Conf. Computational Intelligence in Bioinformatics and Computational Biology, 2014, 1--8. IEEE.

Pinto, D. Rosso, P. Jiménez-Salazar, H. (2011). A self-enriching methodology for clustering narrow domain short texts. The Computer Journal, 54(7):1148--1165.

Rafeeque, P. Sendhilkumar, S. (2011). A survey on short text analysis in web. 2011 Third International Conference on Advanced Computing, 365--371. IEEE.

Ramage, D. Dumais, S. T. Liebling, D. J. (2010). Characterizing microblogs with topic models. Cohen, W. W. Gosling, S., , *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

Röder, M. Both, A. Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining, 399--408. ACM.

Rosso, P. Errecalde, M. Pinto, D. (2013). Analysis of short texts on the web: introduction to special issue. Language Resources and Evaluation, 47(1):123.

Sedding, J. Kazakov, D. (2004). Wordnet-based text document clustering. proceedings of the 3rd workshop on robust methods in analysis of natural language data, 104--113. Association for Computational Linguistics.

Sridhar, V. K. R. (2015). Unsupervised topic modeling for short texts using distributed representations of words. Proceedings of NAACL-HLT, 192--200.

Stumme, G. Hotho, A. Berendt, B. (2006). Semantic web mining: State of the art and future directions. Web semantics: Science, services and agents on the world wide web, 4(2):124--143.

Tang, J. Meng, Z. Nguyen, X. Mei, Q. Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. ICML, 190--198.

Vitale, D. Ferragina, P. Scaiella, U. (2012). Classification of short texts by deploying topical annotations. Advances in Information Retrieval. Springer.

Wallach, H. M. Murray, I. Salakhutdinov, R. Mimno, D. (2009). Evaluation methods for topic models. Proceedings of the 26th Annual International Conference on Machine Learning, 1105--1112. ACM.

Wang, Q. Cao, Z. Xu, J. Li, H. (2012). Group matrix factorization for scalable topic modeling. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, 375--384, New York, NY, USA. ACM.

Wang, Q. Xu, J. Li, H. Craswell, N. (2013). Regularized latent semantic indexing: A new approach to large-scale topic modeling. ACM Trans. Inf. Syst., 31(1):5:1--5:44. ISSN 1046-8188.

Weng, J. Lim, E.-P. Jiang, J. He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, 261--270, New York, NY, USA. ACM.

Xu, W. Liu, X. Gong, Y. (2003). Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 267--273. ACM.

Xue, G.-R. Dai, W. Yang, Q. Yu, Y. (2008). Topic-bridged plsa for cross-domain text classification. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, 627--634, New York, NY, USA. ACM.

Yan, X. Guo, J. Lan, Y. Cheng, X. (2013). A biterm topic model for short texts. Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, 1445--1456, New York, NY, USA. ACM.

Yang, Y. Liu, X. (1999). A re-examination of text categorization methods. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 42--49. ACM.

Zhang, J. Korfhage, R. R. (1999). A distance and angle similarity measure method. Journal of the Association for Information Science and Technology, 50(9):772.

Zhao, W. X. Jiang, J. Weng, J. He, J. Lim, E. Yan, H. Li, X. (2011). Comparing twitter and traditional media using topic models. Clough, P. D. Foley, C. Gurrin, C. Jones, G. J. F. Kraaij, W. Lee, H. Murdock, V., , *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, volume 6611 of *Lecture Notes in Computer Science*, 338--349. Springer.

Zuo, Y. Zhao, J. Xu, K. (2015). Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 1--20. ISSN 0219-3116.

Zuo, Y. Zhao, J. Xu, K. (2016). Word network topic model: a simple but general solution for short and imbalanced texts. Knowl. Inf. Syst., 48(2):379--398.