Master Degree Project

UNIVERSITY OF SKÖVDE

Finding Early Signals of Emerging Trends in Text through Topic Modeling and Anomaly Detection

Master Degree Project in Informatics
with a Specialization in Data Science

Two years Level, 30 ECTS
Spring term, 2018

Sergey Redyuk

Supervisor: Alexander Karlsson
Examiner: Maria Riveiro

Abstract

Trend prediction has become an extremely popular practice in many industrial sectors and academia. It is beneficial for strategic planning and decision making, and facilitates exploring new research directions that are not yet matured. To anticipate future trends in academic environment, a researcher needs to analyze an extensive amount of literature and scientific publications, and gain expertise in the particular research domain. This approach is time-consuming and extremely complicated due to abundance of data and its diversity. Modern machine learning tools, on the other hand, are capable of processing tremendous volumes of data, reaching the real-time human-level performance for various applications. Achieving high performance in unsupervised prediction of emerging trends in text can indicate promising directions for future research and potentially lead to breakthrough discoveries in any field of science.

This thesis addresses the problem of emerging trend prediction in text in two main steps: it utilizes HDP topic model to represent latent topic space of a given temporal collection of documents, DBSCAN clustering algorithm to detect groups with high-density regions in the document space potentially leading to emerging trends, and applies KL divergence in order to capture deviating text which might indicate birth of a new not-yet-seen phenomenon. In order to empirically evaluate the effectiveness of the proposed framework and estimate its predictive capability, both synthetically generated corpora and real-world text collections from `arXiv.org`, an open-access electronic archive of scientific publications (category: `Computer Science`), and `NIPS` publications are used. For synthetic data, a text generator is designed which provides ground truth to evaluate the performance of anomaly detection algorithms.

This work contributes to the body of knowledge in the area of emerging trend prediction in several ways. First of all, the method of incorporating topic modeling and anomaly detection algorithms for emerging trend prediction is a novel approach and highlights new perspectives in the subject area. Secondly, the three-level word-document-topic topology of anomalies is formalized in order to detect anomalies in temporal text collections which might lead to emerging trends. Finally, a framework for unsupervised detection of early signals of emerging trends in text is designed. The framework captures new vocabulary, documents with deviating word/topic distribution, and drifts in latent topic space as three main indicators of a novel phenomenon to occur, in accordance with the three-level topology of anomalies. The framework is not limited by particular sources of data and can be applied to any temporal text collections in combination with any online methods for soft clustering.

Keywords: machine learning, text mining, topic modeling, emerging trend prediction, novelty detection, group anomaly detection

Acknowledgments

I want to express my gratitude to my supervisor, Alexander Karlsson, for a great idea to dig into the topic of emerging trend prediction, for all the help and guidance I received while working on this thesis. I would like to thank my examiner, Maria Riveiro, for her valuable comments and suggestions. Thank you, Alex and Maria, for the time you spent helping me out and explaining how I can improve my work. I want to thank my opponent and fellow Master's student, Jack Bennett, for his effort and contribution to the final report. I am also grateful to my program coordinator, Tove Helldin, and my colleagues at SAIL research group for their patience and desire to help.

To my family and friends, thank you for cheering me up when things were not going that shiny. I truly appreciate it, although sometimes I don't say it as often as you deserve!

Thank you! Without you, guys, this thesis project would not have been that fun!

Contents

List of Abbreviations

ACM	Association for Computing Machinery.
APA	American Psychological Associa.
BOW	Bag-Of-Words.
cDTM	Continuous-Time Dynamic Topic Model.
ciDTM	Continuous-Time Infinite Dynamic Topic Model.
DBSCAN	Density-based Spatial Clustering of Applications with Noise.
DENCLUE	DENsity-based CLUstEring.
EHDP	Hierarchical Evolving Dirichlet Process.
HDP	Hierarchical Dirichlet Process.
IR	Information Retrieval.
IT	Information Technology.
LDA	Latent Dirichlet Allocation.
LSI	Latent Semantic Indexing.
NIPS	Neural Information Processing Systems.
NLP	Natural Language Processing.
OPTICS	Ordering points to identify the clustering structure.
POS	Part-of-Speech.
SQL	Structured Query Language.
T-SNE	T-distributed Stochastic Neighbor Embedding.

1 Introduction

Trend prediction has become an extremely popular practice in many industrial sectors and scientific areas [Kontostathis et al., 2004]. It helps companies to make strategic decisions, plan budget, manage resources – to be better prepared for operating in a fast-changing world. Different fields where trend prediction is used include, but are not limited to, the analysis of emerging economics and markets, stock price prediction; monitoring of oil consumption, tourist arrivals, voting polls; trends in E-commerce and social networks, diets and food industry. By predicting emerging trends successfully, one can not only increase revenue but also improve the quality of services and reach better customer satisfaction. Researchers in academia, on the other hand, are mainly interested in foreseeing emerging directions for further research based on the analysis of scientific literature and publications. That, in turn, leads to many breakthrough innovations regardless the field of study. To mention yet another rapidly growing area, technology forecasting which is driven by various institutions and corporations such as MIT, Harvard, Google or Forbes helps practitioners anticipate customer needs and directions for technological advancement.

Trend analysis in general is highly beneficial yet extremely demanding. It requires intensive investments in human resources - reviewers, marketologists, business analysts, domain experts - who need to stay up-to-date with recent achievements and development in their field. Experts constantly browse through news articles, white papers and technical reports, scientific literature and publications, as well as posts in social networks and forums; and attend various workshops and conferences in order to keep track of the field of research as a complex, dynamically evolving ecosystem. This continuous process helps both private and public organizations shift from reactive responses to turbulent environment towards anticipation, adaptiveness and proactive approaches. Unfortunately, data abundance makes the computational complexity of trend prediction algorithms higher and has become a critical problem for both academia and industry. According to Enríquez et al. [2017], 1.8 ZB of data is generated in two days by the year 2011. This amount of data is larger than the accumulated data from the origin of civilization to 2003. Most of these data is unstructured text [Bello-Orgaz et al., 2016]. Thus, to there is a rapidly growing strive to develop new algorithms for automated prediction of emerging trends in text. This thesis project addresses the problem of detecting patterns that correspond to events that potentially precede emerging trends. In other words, the overall problem to be addressed is detection of emerging trends before they become well-established. It is believed that, in contrast to trend prediction when future changes are approximated based on historical trend, detection of signals that precede these emerging trends has a great value since it allows decision makers to be proactive and discover truly novel trends and phenomena. Thus, this project proposes a framework for capturing *early signals of emerging trends in temporal text collections*. It proposes to incorporate topic modeling and anomaly detection algorithms to achieve this goal.

The report is organized as follows. Section 2 gives an introduction to the concepts of trend prediction, topic modeling, concept drift and anomaly detection. Experienced readers might skip this section due to introductory nature of the content. Section 3 provides a detailed description of the problem of emerging trend prediction in text which is addressed in this work. Sections 4 and 5 explain the scientific method being applied in this thesis project as well as the details of experimental design. Section 6 reflects on the

main work conducted for this thesis project and shares the contribution. Sections 7 and 8 position the work with respect to related studies, discuss limitations of this project and highlight promising directions for further investigation.

2 Background

This section is aimed to make the report self-contained and gives all the information necessary for the novice reader to comprehend the content of the report. It describes a terminology used in the project, and gives a brief introduction to topic modeling, anomaly detection, and emerging trend prediction. See Section 7 for an overview of the related work.

2.1 Terminology

There are several terms and concepts used throughout the report that are the main "building blocks" in the subject area. These definitions are originally presented by Blei et al. [2003] and Wang et al. [2017], and are adapted for this project.

A *word* w is the basic unit of discrete data, defined to be an item from a vocabulary V indexed by $\{1, \ldots, N_V\}$.

A *document* d is a set of N_d words denoted by $d = (w_1, w_2, \ldots, w_{N_d})$. Each document is represented by a Bag-Of-Words (BOW) – a simplifying model used in Natural Language Processing (NLP) and Information Retrieval (IR) [Zhang et al., 2010]. This model operates with a set of distinct words that appear in the document, disregarding punctuation, grammar and the word order. It is assumed that BOW is sufficient to represent latent topics yet brings important simplifications to the topic modeling algorithms.

A *corpus* D is a collection of N documents denoted by $D = (d_1, d_2, \ldots, d_N)$. A *temporal corpus* extends the definition of a corpus. It is an ordered set of corpora X denoted by $X = (D_1, D_2, \ldots, D_T)$, where T is the number of epochs and D_t denotes a collection of documents at epoch t. D_t is denoted as $D_t = (d_{t,i})_{i=1}^{N_t}$ where $d_{t,i}$ is the ith document from collection D_t and N_t is the number of documents at epoch t.

Since BOW representation of a text collection is a high dimensional sparse matrix, it makes the task of processing and analyzing this corpus computationally demanding. Thus, there are numerous methods and algorithms designed specifically to transform textual data into the low dimensional representation.

2.2 Topic Models

Topic modeling is a family of algorithms which belong to hierarchical Bayesian models and capture the underlying semantic representation of a given document collection [Blei and Lafferty, 2009]. Topic models analyze text documents to discover the themes - topics - that run through them, how those themes are connected to each other, and how they change over time [Blei, 2012]. The most popular examples of topic models are LDA [Blei et al., 2003] and HDP [Wang et al., 2011]. LDA is a two-level hierarchical Bayesian model which assumes that every document represents a mixture over latent topics, and every topic, in turn, is a distribution over words. Every document is modeled as a Bag-Of-Words following the assumption that the word order is neglected (assumption of exchangeability [Aldous, 1985]). A classic representation theorem established by de Finetti [2016], which states that any collection of *exchangeable* random variables can, in general,

be approximated as an infinite mixture distribution, lead to the creation of LDA. HDP model is an extension of LDA that enables a nonparametric approach to the modeling of data – it finds the number of topics that is optimal for a given dataset.

2.3 Anomaly Detection

Anomaly detection is a family of algorithms which aim to find patterns in data that deviate from a commonly expected behavior [Chandola et al., 2009]. In many application domains anomaly detection is of critical importance because it highlights features of data or its subsamples that are abnormal in a given context and require agents (operator, information systems etc.) to take actions in order to handle these abnormalities. Anomaly detection is widely used in cybersecurity, fraud detection, predictive maintenance, tumor biology, and so on.

Novelty detection, on the other hand, strives to find emergent patterns in data. The concept of novelty is often used interchangeably with anomaly but is usually defined as an anomaly which is then integrated into the model that represents normal data features or behavior. In other words, novelty detection aims to discover new concepts or features in data flow [semi-] automatically. The concept of anomaly is used further in this work as a more general concept, to avoid confusion. Although, it is used interchangeably with the concept of novelty in the application domain of emerging trends prediction and represents emerging or novel patterns that are being discovered in data flow.

A detailed systematic overview of anomaly detection algorithms is presented by Chandola et al. [2009]. In this survey, several challenging factors are described that are relevant to this work. First of all, as anomaly detection is used to find deviating patterns, the concept of normal behavior should be formally defined and properly modeled. All the upcoming data points are then compared with the model and anomaly scores/labels are assigned.

Secondly, the model of normal behavior might be dynamic and evolve over time. In this case, there is a need of updating the model accordingly or use the one which captures dynamic changes in data points automatically.

Thirdly, labeled data is supposed to be stored for many anomaly detection algorithms before training and validating the model. In various application domains, this tasks is computationally demanding or simply infeasible. Thus, unsupervised machine learning algorithms are applied to handle this issue. In text domain, there are no such a notion of "normal" documents. A document in a given corpora is considered novel when it has some fundamentally new features or its combinations that have not been observed before. The overall task is to design an unsupervised machine learning algorithm that models the defined features and searches for documents which deviate from the rest of the text collection.

Finally, many outliers are similar enough to the normal data. It makes the task of defining boundaries (thresholds) between normal and deviating data points extremely hard and, most of the time, domain-specific. There are various assumptions in each application domain which define the methods and techniques applicable for a particular situation, and limit available tools. See section 3.1 for the assumptions specified in this work.

There is a commonly accepted topology of different anomalies – point anomalies, contextual and group anomalies. A point anomaly is an instance of data that deviates from the rest of the data set. Due to high variability of text documents within one particular topic (cluster), it is assumed that point anomaly detection algorithms will have unreasonable false alarm rate for the given application domain. Thus, only contextual and group anomalies are analyzed further.

2.3.1 Contextual Anomalies

A contextual anomaly is an instance of data which is considered an outlier in the given context only. Outside this context this instance is normal with respect to the rest of the dataset. The context is defined according to the data structure. The most common examples are spatial or temporal data when there is a clear way to aggregate or sample the data based on a time window or geoposition. If not explicitly stated, the context might be set by any means of profiling the data. As in many other application domains, contextual anomaly detection in text analysis is highly useful as there is a natural understanding of a "context" present in this field of research. Sometimes contextual anomalies are referred to as conditional anomalies. In terms of database management and SQL, `WHERE`-clause (condition) is considered to define the context. Another way to define the context is to use clustering algorithms and treat a cluster as a subsample of the data set where particular data points are anomalous while being "normal" on the data set scale.

Contextual anomaly detection algorithms, similarly to group anomaly detection, utilize groups of data points (documents) to estimate how significant one document deviates from the others. The main difference between these two families of algorithms is that an atomic item for group anomaly detection algorithms is a group of documents, whereas contextual anomaly detection algorithms consider single documents as atomic items and use groups of similar documents to define the context.

2.3.2 Group Anomalies

A group (collective) anomaly is a set of data instances which deviate with respect to the rest of the data set. This type of anomalies is interesting due to the fact that one individual data point from this group might not be an outlier by itself [Chandola et al., 2009, p. 9]. Collective anomalies can occur only in data sets where one can establish any kind of relationships between individual data points. If data points are completely independent from one another, the fact that these points are forming a group is most likely accidental.

2.4 Emerging Trend Prediction

Emerging trend prediction in text aims to anticipate topics that might be of high interest in the nearest future. Different fields where trend prediction is used include, but are not limited to, the analysis of emerging economics [Drechsel and Tenreyro, 2017] and markets [Dutta, 2018], stock price prediction [Zhang et al., 2016]; monitoring of oil consumption [Yu et al., 2018]; trends in E-commerce and social networks [Abbas et al.,

2017]. In these application domains, anticipation of the nearest future is extremely beneficial since it allows one to make better decisions. Academic interests comprise research in medical domain [Berlanga-Llavori et al., 2008] and healthcare [Hong et al., 2014], analysis of scientific literature [Mörchen et al., 2008; Liu et al., 2013]. Another rapidly growing area is technology forecasting which is run by various institutions and companies, such as MIT[1], Harvard[2], Google[3], TechCast Project[4] and Forbes[5], in order to anticipate customer needs.

[1] https://www.technologyreview.com/, MIT Technology Review, Accessed: 2018-01-23
[2] https://hbr.org/topic/technology, Harvard Business Review: Technology, Accessed: 2018-01-23
[3] https://trends.google.com/trends/, Google Trends, Accessed: 2018-01-23
[4] https://www.techcastglobal.com/, TechCast Global, Accessed:2018-01-23
[5] https://www.forbes.com, Forbes - American Business magazine, Accessed: 2018-01-23

3 Problem Formulation

Most of the state-of-the-art topic modeling algorithms have one common drawback - a few anomalies in the data do not affect the model. In other words, several individual outliers are not capable of changing statistical distribution that most of the models operate with. One of the examples is a new term presented in a particular upcoming document: the wast majority of topic modeling algorithms will discard this term since it is not a part of the model's vocabulary. The models which handle dynamic vocabulary won't be affected by this new term because it appeared in a single document only at that time. In case this term foreruns the birth of a fundamentally new trend, it will take time before the model captures this trend.

An algorithm which incorporates a particular topic model and anomaly detection techniques can serve as an early warning system by detecting changes that might lead to the discovery of new concepts before becoming an established trend. The academic interest comprises various predictive models to be integrated into the topic modeling framework. It is worth mentioning that topics in text are considered the main application domain but the idea itself is generic and can be potentially applied to other clustering algorithms.

Thus, the overall *goal* of the project is to *design a framework for detecting early signals of emerging trends in temporal text corpora, by utilizing topic modeling and anomaly detection algorithms.*

In order to achieve the goal of the project, several steps are required. First of all, literature analysis is needed which focuses on topic modeling, anomaly detection, concept drift, emerging trend prediction. It is required to review what has already been done in the research area. It helps to justify the originality of the work, limit the scope, apply best practices used in the field, and avoid common pitfalls. Literature analysis facilitates the process of positioning the work in the context of other research conducted previously in the field, and helps to identify how the project contributes to the subject area. In addition, it contributes to the assessment of the work by objectively weighting the project's advantages and drawbacks.

The next step is to define the concept of an anomaly grounded to the context of anomaly detection for topic modeling, text analysis, and emerging trend prediction. It is required in order to get a comprehensive view of what is considered to be an anomaly in temporal text collections and what anomalies can capture early signals of emerging trends. Based on literature analysis, a hierarchy of anomalies specific to the subject area is to be provided. This hierarchy itself is considered a contribution to the subject area which defines the concept of anomaly given specifically for text analysis and topic modeling. It is also aimed to elaborate on a decomposition of levels for anomalies in the hierarchical corpora-document-word representation.

As the next step, design of a framework and its prototyping are to be conducted in order to develop a piece of software that combines topic modeling and anomaly detection algorithms to detect early signals of emerging trends in text. For topic modeling, hyperparameter tuning is required since topic models depend on a specific dataset. This step also consists of choosing a metric for hyperparameter tuning applied to a topic model in order to reach greater performance on extracting topics. See section 6 for details. Choosing an anomaly detection algorithm which is capable of detecting early signals of emerging

trends is another important part of this thesis project. The concept of an early signal is defined in accordance with the proposed domain-specific hierarchy of anomalies. Combining together a topic model with anomaly detection is the main part of the project which contains experimental design and continuous improvements of the proposed solution.

The last step is to evaluate performance of the designed solution based on both synthetic and real-world datasets. That is to be done in order to estimate how good the framework performs on emerging trend prediction task by detecting anomalies in a temporal text collection. This objective contributes to the generalization of the designed solution, by estimating how stable the predictive capability of the model is depending on different settings - various text collections, time frame for each epoch, size of the corpus used for initial training process and so on. It also makes the justification of the obtained results objective and solid.

The steps described above lead to the following objectives.

- *O1.* Conducting literature analysis on topic modeling, anomaly detection, concept drift, emerging trend prediction, and the previous work which pursues similar goals;
- *O2.* Providing a definition of an outlier grounded to the context of anomaly detection for topic modeling in text, and the domain-specific hierarchy of anomalies;
- *O3.* Prototyping iteratively on a solution that combines topic modeling and anomaly detection algorithms to detect outliers that might capture early signals of emerging trends in text;
- *O4.* Evaluating performance of the designed solution (synthetic and real-world datasets), choosing a metric to estimate how good the topic model performs on extracting latent topics, and proposing a way to assess whether the framework is capable of detecting early signals of emerging trends in text by utilizing anomaly detection.

3.1 Scope

There are two main properties of the proposed model which shape the scope of the work. First, this model belongs to the class of unsupervised machine learning which implies that there are no explicit target outputs associated with each input. In other words, it is assumed that there is no prior knowledge regarding the topic space. This constraint is justified by the volume of text streams prevailing in the web at the moment [Enríquez et al., 2017]. It becomes intractable to prepare accurate labels for each document in the corpus, as supervised and semi-supervised machine learning algorithms require. The second property states that the model belongs to the class of online learning algorithms which, in contrast with batch learning, operate with upcoming data in a sequential order and update the predictor for future data at each step. In other words, given the epoch t, the prediction regarding the epoch $t+1$ is made before this epoch starts. This constraint is motivated by the temporal nature of the data being analyzed, which expands over time as a stream.

It is worth mentioning tht the key focus of this project is on early signals of emerging trends in text, not the trend prediction itself. Scientific abstracts from `arXiv.org` and

`NIPS` publications are used as real-world datasets.

The following domain-specific assumptions are made in this work. They define several limiting factors for the project and specify models and algorithms which are appropriate within a given context.

- Text documents occur in a sequential order and have metadata to locate a document in time;
- Text documents in a corpus share vocabulary;
- Text documents in a corpus share topics;
- Topics evolve over time. The task to track topic evolution is one of the core parts of the study and is defined by the application domain;
- Main stream documents are far more frequent than the deviating ones. Violation of this assumption decrease the model accuracy and lead to high false alarm rate. In this context, false alarm is meant as a situation when one or several text documents are considered as data points which represent a new concept whereas, in reality, they belong to the main stream trend of the text corpus;
- In text analysis, the nature of data allows only scoring techniques to be applied in order to detect novelty in a text stream. Scoring techniques assign a continuous value as a score to each document in a text collection depending on the degree of certainty to which the document is considered as deviating. Data analysts can specify thresholds or select top N outlying documents for further analysis based on the list of documents ranked by the assigned anomaly score. Labeling techniques, on the other hand, assign binary markings and do not use thresholds. These techniques are applied, for instance, in rule-based systems for anomaly detection which raise an alarm when a particular pattern is found. In statistical topic modeling, where the model's parameters are represented by random variables, as well as in text analysis in general, usage of such predefined patterns is impossible due to the nature of novelties – they cannot be anticipated beforehand to set up a specific rule.
- Volumes of text corpora to analyze make the use of supervised machine learning algorithms, where main stream documents are labeled as "normal" for a model to be trained, are time-consuming implying manual labeling procedures conducted by data analysts. Thus, only unsupervised machine learning algorithms are used in this work.
- Word and topic distribution is high dimensional and, in most of the cases, extremely sparse.
- Handling high *variations* in documents which belong to the same topic and their word-topic distribution matrices is challenging [Chandola et al., 2009, p. 19].

4 Research Method

Methods of quantitative research are applied for this project. According to Oates [2006], "design and creation" research strategy is recommended to be used in applied research within the field of informatics. It focuses on developing so-called artifacts – constructs, models, methods or instantiations – that "represent a situation, aid problem understanding and solution development". This strategy was originally proposed by [March and Smith, 1995] as the design science framework which is widely used for researching in IT. In this project, the following artifacts are defined:

- Constructs: concepts and vocabulary used – formal definitions of a word, document, corpus, epoch, given the context of the project; three-level topology of anomalies; domain-specific assumptions;
- Models: a combination of constructs to represent a situation – incorporation of topic modeling and novelty detection algorithms limited by the domain-specific assumptions, to represent latent topic space, concept drift, and to capture early signals of emerging trends in text;
- Methods: guidance on the process stages and the models designed – data preprocessing methods and hyperparameter tuning algorithms;
- Instantiations: a working system that demonstrates the artifacts mentioned above – a computer-based prototype of a detection system that captures early signals of emerging trends in text; performance evaluation scenarios to demonstrate consistency of the project's main idea.

The contribution of the project is based on the literature analysis, followed by creating a framework of emerging trend prediction and evaluating it empirically in both synthetic and real-world contexts.

The following actions are to be taken with respect to the design and creation strategy – awareness, suggestion, development, evaluation and conclusion. The overall research process is formed as an iterative cycle of incremental trials and refinements [Kuechler and Vaishnavi, 2008].

- *Awareness* – Recognition of a problem, specified by practitioners or stated as part of future directions in scientific literature, *Objective 1* and *2*;
- *Suggestion* – Providing an idea of how the problem can be addressed, *Objective 1* and *2*;
- *Development* – Domain-specific implementation of an artifact that corresponds to the *Objective 3*;
- *Evaluation and Conclusion* – Performance assessment, *Objective 4*.

For the *development* stage, prototyping is used as a systems development methodology. This approach is considered as a best practice for agile development and incremental

updates. Moreover, one of the advantages of prototyping is that a researcher does not necessarily need to have a full understanding of a problem before exploring the solution space.

The artifact under evaluation – a framework for early signal detection of emerging trends in text domain – is considered to demonstrate a "proof of concept" by modeling a synthetically generated dataset where ground truth is available and the modeling process is performed under supervision, as well as "proof by demonstration" by that works in a real-life context, by analyzing real-world textual corpora. See Section 6.1 for a review of data generation and mining algorithms used for this project. Section 6 gives detailed description of the evaluation process and metrics applied for performance analysis.

5 Experimental Design

This section describes a framework that is proposed to detect early signals of emerging trends in temporal text corpora.

As mentioned in section 3, topic modeling algorithms are commonly used for document summarization and structuring of text corpora. However, when it comes to trend prediction, these models do not perform well. That happens due to several reasons. First of all, topic modeling algorithms are not designed for prediction tasks. Most of these algorithms are based on statistical inference which approximates parameters of the model given the observations. Thus, prediction tasks can be handled only if upcoming documents are similar to the ones the model was trained on. Unfortunately, this happens rarely in the real-world scenario. Secondly, topic models are "rather slow" in discovering topics due to their statistical nature. In order to capture a topic in a precise manner, there is a need to retrieve a large number of documents where this topic is present. In other words, topic models reach high performance with well-established topics which occur in text corpora often. Unfortunately, their performance drops when it comes to emerging topics, because the number of observations that capture a novel topic is relatively smaller than the number of observations that contain well-established topics. Thus, the statistical model is less likely to be affected by these novel topics.

In order to overcome these drawbacks, it is suggested to apply topic modeling algorithms together with anomaly detection. It is assumed that anomaly detection algorithms enable detection of *early signals* of novel topics or trends before they become well-established, given the context of emerging trend prediction in text.

Birth of a new trend in a text corpus can be described as follows. Applicable to the analysis of scientific literature, it starts with a single document that contains a novel term, concept, context or combination of topics that have not appeared in the text corpus before. As mentioned in the scope of this thesis project, the work is limited by the analysis of scientific abstracts. It is worth mentioning that other examples of novelties might be related to new attitudes of the writer or emotional states expressed in posts from various social networks (see Wilson et al. [2005]; Pang et al. [2008] discussing sentiment analysis). As the new trend emerges, more similar documents that mention this particular phenomenon will appear. Another setting is when the term or concept occur in a new context which also might lead to novel trends to emerge.

To address the issues stated above, the following framework is proposed (see Figure 1). Temporal text collection is taken as an input. It is represented as a list of epochs where each epoch contains several documents, and each document is a Bag-Of-Words. This data collection is then being preprocessed. First of all, data preprocessing step reduces the amount of time required for further analysis by filtering and transforming the data. Secondly, smaller vocabulary decreases variance of the data and leads to better stability of the topic model [Belford et al., 2017]. See section 6 for the description of data preprocessing techniques applied for the real-world datasets.

The proposed framework provides a novel approach on emerging trends prediction and contributes to the development of this subject area. To be more precise, the following design choices are made to limit the scope of the project. First of all, Hierarchical Dirichlet Process (HDP) is to be used as a baseline topic modeling algorithm [Teh et al., 2005]. This

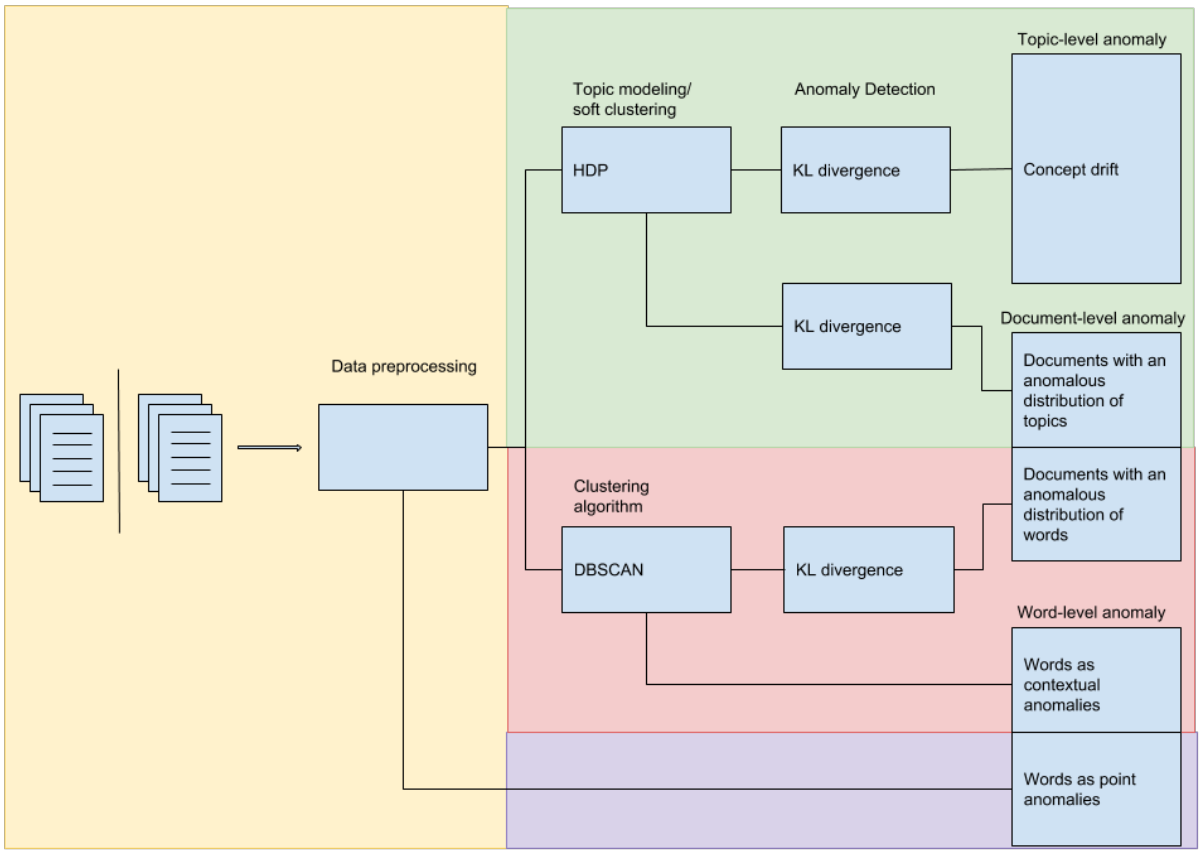

Figure 1: *Graphical representation of the proposed framework divided into 4 components. The first component (yellow, left) represents a temporal text collection and preprocessing steps to reduce dimensionality of data and improve its stability. The second part (violet, right bottom) represents a component for detection of new vocabulary. The third part (red, right center) is related to group anomaly detection based on word distribution. The last component (green, right top) focuses on group anomaly detection based on topic distribution. This figure captures a single snapshot for a particular time frame. Temporal changes are analyzed by iterative repetition of the analysis for each upcoming time frame.*

model takes into account word-document-corpus hierarchy of data and infers the number of topics represented by a given text collection automatically. Furthermore, this model is one of the most popular topic modeling algorithms that is cross-validated and serves as a baseline for many other algorithms. Anomaly detection algorithm is to be integrated with HDP in order to capture changes in both topic space topology and upcoming data.

The core of the framework is based on the proposed three-level hierarchy of anomalies described below. It is assumed that anomalies in both upcoming data and the topic space topology belong to one of three categories – word-level, document-level and topic-level. The decision to choose this particular hierarchy follows the assumptions made by Blei et al. [2003] while working on Latent Dirichlet Allocation. LDA incorporates a two-level hierarchical Bayesian model which operates with the concepts of corpora, documents, and words representing a topic as a distribution of words and a document as a distribution of topics. This approach is widely inherited in other topic modeling algorithms such as HDP, cDTM, ciDTM Teh et al. [2005]; Wang et al. [2012]; Elshamy [2013], therefore it is

considered credible to follow.

Word-level anomalies

A word-level anomaly represents a new word which is being used in a document at epoch t that was not used before. This new word can be either added to the list of stop words or to the vocabulary. The latter case requires adjustments to the topic model. The word might also be an outlier which, in general, does not belong to stop words but is semantically meaningless in the context of the given document. It is assumed that word-level anomalies should be handled in a semi-automatic way where an analyst makes the final decision whether to add this word to the list of stop words, update the vocabulary and the topic model, or simply to ignore it.

Document-level anomalies

A document-level anomaly is a document which has abnormal distribution of words or topics, compared to the previously processed ones. On the one hand, this document might be a standalone outlier deviating due to various reasons. Some of these reasons are a document writing style, specific perspectives on how topics are described, or stringency of the language. On the other hand, this document might forerun a shift in the topic space topology where one or several topics evolve (topic evolution, birth, death, merging or branching). To differentiate these two cases, outliers should be monitored over time following the fact that standalone outliers do not have a temporal trend. If the trend remains after several epochs, it is likely to state that the outlier represent a novel trend. As mentioned earlier, a topic modeling algorithm itself cannot track these shifts in the agile way since it takes time (and more similar documents, respectively) to affect the statistical model. It is presumed that anomaly/novelty detection algorithms are able to handle this issue faster.

Topic-level anomalies

These type of outliers describe the highest level of abstraction covered by topic modeling algorithms - latent topic space representation of the given corpus. All the upcoming documents shape a topic topology which changes over time and embodies different trends occurring in the corpus. There are two main categories of anomalies which belong to this level. The first category are outliers in the topic space which do not belong to any established trends. There is a chance that these anomalies capture the birth of new trends, similar to the document-level anomalies. Another category represents an abnormal change in the established trend. One example can be that a trend which was stable and did not change over several epochs started shifting or fluctuating. In other words, this trend undergoes changes in its stability or new patterns of its evolvement are detected. For the latter category, applying anomaly detection algorithms for sequential data might be of particular interest.

In order to capture anomalies on each level, two types of anomaly detection algorithms can be potentially applied – group anomaly detection and contextual anomaly detection. The third group of algorithms - point anomaly detection - is discarded due to the following reason. High variance of the documents (diverse vocabulary, synonyms, writing styles) is assumed to cause a high false alarm rate, when an alert is raised by the system marking one particular document as a potential signal of an emerging trend to occur while this document is just a single outlier which does not lead to any future trends. In other words, one needs to detect several documents with similar deviating patterns, most

likely distributed in time, in order to highlight a group of documents that evolves and potentially leads to an emerging trend. Obviously, detection of point anomalies is not able to capture this evolving process. Thus, group anomaly detection is applied instead.

Group Anomaly Detection

A subset of data which deviates from the rest of the dataset is called as a group (collective) anomaly [Chandola et al., 2009]. It might be beneficial to capture this kind of anomalies in case individuals data points are not anomalous by themselves, but become an outlier as a group. As mentioned before, group anomaly detection algorithms are considered reasonable to use in this thesis project as they capture collective patterns of data subsets thus can incorporate temporal dependencies which are valuable in the context of trend prediction. A simple way to analyze temporal collections is to process each epoch separately. The main drawback of this approach is the difficulty to correlate detected groups over time and track their dynamics. One solution is to stack data points (documents) from neighboring epochs and analyze them together. Another alternative is to use windowing functions of particular size in order to sample documents from the temporal text collection. For this thesis project, a time window of several epochs is used with a step of one epoch. That allows neighboring time frames to have common data points that are used to detect temporal correlation and track changes over time. See section 6.5 for the illustration.

DBSCAN clustering algorithm proposed by Ester et al. [1996] is applied in this work in order to detect groups of documents with density that deviates from the rest of the dataset. It is believed that dense clusters of documents are the early signals for emerging trends to occur. There are two other alternatives of density-based clustering algorithms – OPTICS [Ankerst et al., 1999] and DENCLUE [Hinneburg and Keim, 2003] – and their variations that might be considered instead of DBSCAN. For this project, DBSCAN is chosen due to the following reasons. As reported in Ankerst et al. [1999], OPTICS has a constant slowdown in a run-time compared to DBSCAN. According to Hinneburg et al. [1998], DENCLUE performs 45 times faster than DBSCAN. Although, DENCLUE does not perform well on high-dimensional data. It works poorly on data with uniform distribution, and high-dimensional data looks uniformly distributed due to the curse of dimensionality [Aggarwal et al., 2001].

Core of the Framework

The core of the framework consists of three main parts. The first part focuses on detection of new vocabulary appearing in the text corpus. This part is rather straightforward and finds words which have not appeared in text yet. This is to detect documents that potentially describe a new concept – early signals of an emerging trend.

The second part is aimed to detect group anomalies in text. DBSCAN is applied to capture dense regions of the data space each representing a group of documents similar to one another. These dense regions are believed to be another type of early signals of trends to emerge. After all the potentially anomalous groups with deviating density are captured, Kullback-Leibler (KL) divergence is used to assign an anomaly score [Kullback and Leibler, 1951]. It is defined to be

$$D_{KL}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \tag{1}$$

where P represents the "true" data distribution, and Q is the approximation of P. KL divergence measures the amount of information lost by using Q to approximate P [Anderson, 2002].

This measure is applied to evaluate how one probability distribution diverges from another one. Based on the calculated value of KL divergence, where one probability distribution is a word distribution within the captured group and the second one is a word distribution of all the documents from the current epoch, anomaly scores are assigned to each cluster in order to conclude whether this cluster is anomalous or not.

KL divergence is chosen as a natural measure of the faithfulness with which Q models P [Maaten and Hinton, 2008]. As a statistical distance, it is preferred over similarity metrics or distance functions such as L1 or L2 norms due to the fact that statistical distributions of the data are to be compared within the framework. Compared to other statistical distances, KL divergence is chosen as a general and the most commonly used measure applied in machine learning, that naturally fits the settings of this thesis project.

The third part deals with the topic space and utilizes topic modeling algorithm to extract topics from temporal text corpora. In this thesis project, HDP model is used as one of the first topic modeling algorithms presented in the field. This model is chosen because of two main reasons. First of all, HDP is a credible model since it was cross-validated and applied in many other scientific publications. Secondly, it does not require the number of topics to be manually specified, as LDA does. In this project, HDP is used in order to capture both document- and topic-level anomalies. For document-level anomalies, documents are represented as a distribution of topics and are being analyzed then similarly to the word-level anomalies – DBSCAN clustering algorithm and KL divergence are applied to group documents and assign anomaly scores to the data. For topic-level anomalies, changes in the topic topology itself is of high interest. Since topics are represented as a distribution of words, changes in this distribution might indicate a concept drift. Generally speaking, concept drift refers to the situation when the relation between the input and output data changes over time [Gama et al., 2014]. For this thesis project, concept drift is considered to be a change in a word distribution that describes a topic. For example, a topic "text mining" in 1980s could be represented by concepts "Latent Semantic Indexing", "TF-IDF" or "non-negative matrix factorization", whereas nowadays it can be described by using concepts such as "topic models", "natural language processing", "recurrent neural networks" etc. Older concepts are still in use, but the distribution of how likely (how often) particular concepts occur in the documents has changed. In many cases, it means that a new trend has emerged. Monitoring distributions on two different time windows (see Gama et al. [2014, p. 17]) is applied in order to detect concept drift. It compares "recent" and "current" distributions of data using statistical test with the null hypothesis which states that two distributions are equal. The method presented by Dasu et al. [2006]; Sebastião and Gama [2007] utilizes KL divergence similarly to the way it is used to detect word- and document-level anomalies.

Combining these three parts altogether provides a comprehensive approach of capturing early signals of emerging trends in text.

5.1 Validity Threats

This framework has several crucial threats to validity. The first threat is related to conclusion validity (see Wohlin et al. [2012]) and represents confirmation bias or so-called fishing – the tendency to search for a specific result [Wohlin et al., 2012, p. 104]. Two harmful outcomes of this threat are the design of the framework with respect to the prior assumptions which, in the worst-case scenario, might be incomplete or simply incorrect, or the choice of evaluation metrics and data generation procedures biased to the designed framework which might demonstrate high performance on the test set but be poorly generalized to the real-world scenarios. Another threat is related to the reliability of treatment implementation. In other words, it considers the overall quality of the framework's implementation. It means that the theoretically valid framework might perform poorly or provide incorrect results due to errors in the source code. In order to handle these threats to validity, the report itself and the corresponding source code are uploaded as open-access materials to enable peer-reviewing and promote reproducible research. In this case, any parts of the work which are doubtful can be cross-validated and fixed.

As part of external validity threats, generalization issues are of high importance. There might be a chance that the designed framework is not suitable for general population of data and works effectively only for a limited subsample of data (selection bias). Scientific abstracts are one specific example that uses stringent language and rather short form of representation which might simplify the analysis. Having longer documents or text with higher vocabulary might decrease the performance or lead to unexpected artifacts. To handle this threat, two actions are taken. First of all, randomization principles are used in data generation procedures to avoid bias and make the results as generic as possible. Secondly, two real-world datasets are used to cross-validate the performance of this framework. It is worth mentioning that both of the temporal text collection belong to the category of scientific publications thus this method does not guarantee generalization properly.

6 Empirical Evaluation

This section describes evaluation procedures for the designed solution. It demonstrates how synthetic data is generated, as well as the way how real-world datasets – abstracts from `arXiv` (selected, category: `Computer Science`) and `NIPS` publications – are collected and preprocessed. It also discusses hyperparameter tuning methods applied in this work and demonstrates the results.

6.1 Datasets

In this work, two approaches are used to evaluate performance of the model. The first approach is focused on the generation of synthetic data which provides ground truth for performance analysis. Another approach utilizes real-world datasets in order to demonstrate predictive capability of the model for several applications. These approaches are discussed below.

6.1.1 Synthetic data

Synthetic temporal text collection is generated with the following parameters specified.

- the number of epochs;
- the number of topics;
- the number of words that describe a single topic;
- mean and standard deviation for the number of documents generated per epoch, including linear trend to simulate slight increase in the number of publications over time;
- mean and standard deviation for the number of words generated per document;
- weights for topic distribution (popularity) on the corpus level.

These parameters are believed to be sufficient in order to generate a temporal text collection. To be exact, the following values are chosen for the empirical evaluation purposes in this project: 10 topics are generated, 10 words each; 40 epochs are chosen based on the trade-off between the computational time required for HDP model to extract topics and the realistic representation of real-world settings (taking one epoch as one week, 40 epochs correspond to the time frame of 10 months - the amount of time sufficient for trends to emerge in real world).

Mean and standard deviation for the number of documents documents generated per epoch and words generated per document add variation to the synthetic corpus and make the setting similar to stochastic processes underlying generation of the real textual document.

For simplicity, each word is generated by following the pattern "*word_i_j*" where i represents the topic identifier and j – word identifier. In other words, a term "*word*_3_4"

corresponds to the fourth word describing topic #3. The document generation procedure looks as follows.

```
for each epoch in epochs:
 for each document in epoch:
  for each word in document:
   draw topic from Multinomial(topic_distribution)
   draw word from Uniform(topic)
```

Here a text collection is a list of epochs where each epoch is a list of documents and each document is a list of words. For each word, topic identifier is sampled first from the topic distribution specified for the given document. Multinomial distribution is used to describe a discrete mixture of topics. A word is sampled from the uniform distribution of words that describe a given topic i.

Below is one example of the synthetically generated text:

```
['word_0_5', 'word_9_5', 'word_8_1', ..., 'word_0_2']
['word_0_6', 'word_8_9', 'word_9_9', ..., 'word_0_7']
```

Other more complicated ways to generate the data can be applied. For instance, topic distribution can be specified manually for each epoch. Yet another example is that dynamic changes of each topic might follow a particular trend that can be set by the analyst. However, these approaches are time-consuming and require manual crafting. Thus, they are kept for future work.

Anomaly Injection

In order to evaluate the framework and estimate how well anomalies in text can be detected, injection of anomalies is used. It is executed on the word- and the document-level by injecting new words and simulating document generation in several dense regions. It is worth mentioning that a straight-forward way of detecting word-level anomalies makes evaluation procedures unnecessary because the task of comparing upcoming words with existing vocabulary is trivial. For application domain where the size of vocabulary is rather big, data structures such as search trees might be beneficial to use in order to store the vocabulary and accelerate the comparison.

When it comes to the detection of dense regions of similar documents, DBSCAN is used together with T-SNE visualization algorithm [Maaten and Hinton, 2008]. T-SNE is used to facilitate visual inspection and provides two-dimensional embeddings of high-dimensional data which preserve data similarity. DBSCAN clustering algorithm depends mainly on two hyperparameters – the maximum distance between two samples for them to be considered as in the same neighborhood, and the number of samples in a neighborhood for a point to be considered as a core point. Several combinations of these hyperparameters are used to cluster the data. Visual inspection is applied in order to choose the best suitable candidate that is capable of detecting all the dense regions of documents present on the scatter plot. For an illustration, see Figure 4.

In order to detect concept drift, KL divergence is used to compare how similar "recent" word distribution that describes a particular topic to the "current" one. Based on this entropy estimation, one can capture a time frame where the distribution is changed and thus the concept (topic) is drifting. Since there is no semantic meaning lying behind the synthetically generated data, performance evaluation of the detection of topic-level

anomalies becomes unreasonable. As an outcome for this stage, it is worth mentioning that manual inspection of the topics extracted over time shows that a constantly updating HDP model is able to capture concept drift.

6.1.2 arXiv

`arXiv.org` is a "highly-automated electronic archive and distribution server for research articles", providing e-print service in various fields such as physics, mathematics, computer science, quantitative biology, statistics, and so on. It is maintained by the Cornell University Library, supports open access and allows automated downloads and machine access [6].

Scientific abstracts from a subject area of computer science were used as a main data source. Scientific publications are chosen mainly because of the language stringency and availability of subcategories. Another motivation is the author's expertise in the domain which facilitates manual validation of the quality of extracted topics.

112 epochs (weekly basis) are downloaded starting with January 1st, 2016. This text collection includes around 200000 abstracts with the average length of 80 (+- 40) words after preprocessing steps. The dictionary contains 80000 terms. Data preprocessing techniques are applied in order to reduce the size of vocabulary (2000 terms instead of 80000), time complexity of topic extraction and remove noise – semantically irrelevant words. These techniques are discussed further.

6.1.3 NIPS

NIPS conference papers 1987-2015 is the second dataset used for performance evaluation. It is an open-access dataset stored in the UCI Machine Learning Repository [7]. It contains an occurrence matrix for over 11000 terms used in 5800 articles. Unlike the `arXiv` dataset, NIPS dataset is already preprocessed and does not require any additional steps. Furthermore, it contains terms extracted from the full text of a publication, whereas `arXiv` dataset is collected based on the abstracts only.

6.2 Data Preprocessing

Since effective feature selection is of high importance, as it reduces dimensionality of the data and improves its stability, several data preprocessing techniques are applied to the `arXiv` dataset. First, only abstracts from the publications are used because they are assumed to be a summary of the article sufficient to capture topics and key concept highlighted in a particular research paper. Secondly, mathematical formulas and numbers are removed since they do not add any semantic value to the text. Thirdly, frequent and rare words are removed as well. When it comes to topic modeling algorithms or Latent Semantic Indexing, it is recommended to remove frequent words because they appear in

[6] Indiscriminate automated downloads from this site are not permitted due to the limited server capacity, https://arxiv.org/help/robots, accessed: 2018-05-12.

[7] https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015, accessed: 2018-05-12.

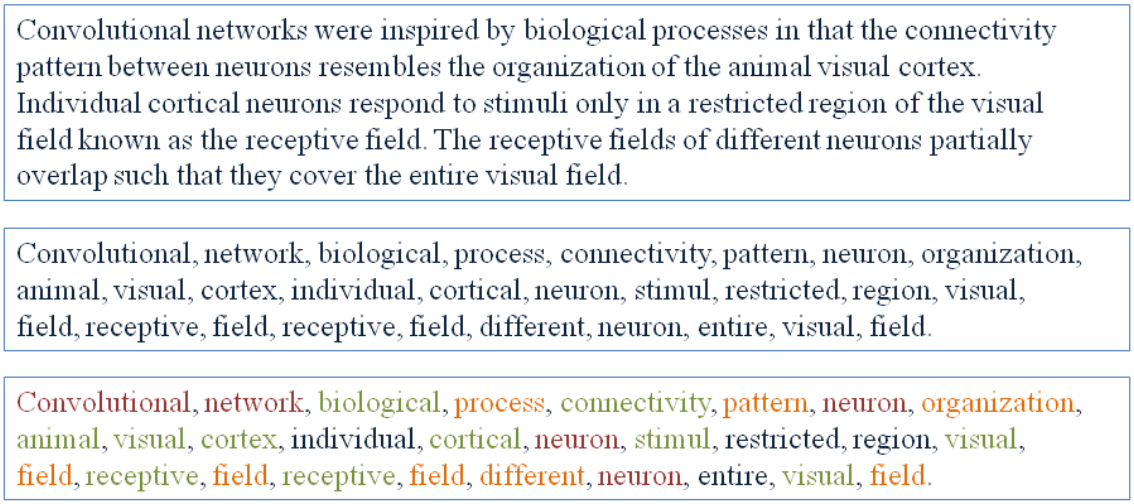

Figure 2: *Example of the preprocessing chain of a given text document from the* `arXiv.org` *data set. An excerpt from the original document (top); filtered Bag-Of-Words after frequent and rare words removal, word tokenization, punctuation, mathematical formulas, stop words removal, short words removal, POS tagging and filtering, stemming and lemmatization (center); processed with HDP - different colors represent different topics (bottom).*

most of the documents thus cannot be used to cluster the data efficiently. Rare words, on the other hand, do not affect underlying statistical model of the topic modeling algorithm.

Several techniques for textual data preprocessing applied in this work are described further.

6.2.1 Word Tokenization

Tokenization is the first step in data preprocessing which takes a sentence, paragraph or text in general as a string and splits into chunks, called tokens. Word tokenization takes a string and splits it into tokens-words by following specific rules. A basic tokenizer usually uses space as a delimiter and removes punctuation. As an outcome, word tokenizer returns a list of tokens – words. In case splitting a string by using space as a delimiter is applied instead of word tokenization, punctuation removal is supposed to be done manually.

6.2.2 Collocation Detection

Collocation detection is applied to capture concepts which are described with more than one word. This provides an additional level of abstraction – from words as tokens to collocations as concepts. This is usually done by detecting neighboring words with high co-occurrence [Mikolov et al., 2013]. In case of a design choice to filter out all parts of speech (see POS Tagging and Filtering) except nouns as carriers of semantics in text, applying a collocation detection algorithm beforehand will preserve some meaningful adjectives as part of a concept. For example, using collocation detection will preserve a

concept “convolutional neural network”, whereas without this step only a single noun “network” will be kept. It is reasonable to conclude that collocation detection helps to preserve semantics during data preprocessing steps.

By default, collocation detection finds bi-grams – collocations that consist of two tokens. Applying this algorithm iteratively allows one to get collocations of a greater length. Otherwise, more complicated algorithms of association mining are supposed to be used.

6.2.3 Punctuation and Stop Words Removal

Punctuation removal task is rather straight-forward and is used to remove punctuation marks that do not carry any semantical meaning. This step should be performed in case work tokenization is not used.

Stop words removal utilizes various lists of tokens that are meaningless such as words “they”, “another”, “each” etc. These lists are created for most of the languages and facilitate dimensionality reduction. There are no unified lists of stop words thus, for this thesis project, several of them are merged together in order to reduce more irrelevant tokens. It is worth mentioning that some additional rules should be used for texts written in English in order to remove short forms of verbs that use apostrophes such as “’ll” or “’ve”.

6.2.4 Short Words Removal

As an optional step in data preprocessing, short words (up to 4 letters) removal might be applied. It is recommended to use due to the high number of short words that do not convey semantics. Although, this step is applicable to English language and is domain specific thus is discarded in this work.

6.2.5 POS Tagging and Filtering

Part-of-Speech tagging is a task of assigning part-of-speech tags to tokens. It belongs to the field of computational linguistics and has a fixed list of tags used in English grammar. Since a lot of words can take more than one POS tag, modern tagging tools utilize context and grammatical structure of the sentence in order to assign a tag to a given token. After the process of tagging is complete, tokens with irrelevant POS tags such as pronouns, possessive endings, wh-adverbs etc. are removed.

6.2.6 Stemming and Lemmatization

Stemming and lemmatization are used to handle terms of different grammatical forms. According to the common sense, endings of the words in English are functional parts of a word but do not carry semantic meaning. Words “introduce”, “introduction” and “introductory” have similar meaning but different forms in writing thus are supposed to be reduced to represent a single concept instead. Porter’s algorithm is the most popular stemming technique so far which uses an extensive set of rules for suffix reduction that

cut out meaningless endings of words and preserves roots that convey semantics [Porter, 1980].

Lemmatizers apply full morphological analysis of words in order to identify and extract their lemmas, or dictionary forms [Müller et al., 2015]. This allows one to analyze various forms of a word as a single item. In order to achieve this, POS tagging and context analysis are used. Using either stemming or lemmatization reduce dimensionality of the given textual corpus drastically by merging all the inflected forms of a word into its root or into a morphological base form.

For this project, data preprocessing techniques mentioned above are applied in a sequential order to clean the data and prepare it for further analysis.

6.3 HDP Hyperparameter Tuning

Since HDP model's performance depends on hyperparameters, tuning is required to achieve the highest results possible. For this scenario, three hyperparameters are being tuned – the first- and the second-level concentration factors, and the top truncation level which limits the maximal number of topics being extracted from the corpora. See Wang et al. [2011] for the detailed description.

Greedy search is used for hyperparameter tuning. This algorithm takes various combinations of hyperparameters and computes a predefined metric. A model that corresponds to the best value of the metric is then chosen as the most promising candidate. The choice of metrics is described below.

Synthetic Dataset

Since most of the topic modeling algorithms are statistical models, it does not matter whether a word has any semantic meaning. Thus, topic models can be applied to the synthetically generated text – topics are to be extracted based on the word frequencies and co-occurrence. To evaluate how well the model performs on the synthetic dataset, the following procedure is designed. First of all, HDP model is applied to the data. Then, extracted topics are analyzed as a word-topic matrix. Since words can be easily identified as belonging to a particular topic or not (the first digit is a topic identifier), that is used to evaluate how well the model captures different topics. Below is an example of topics extracted by HDP.

```
1 Topic #0: word_4_6, word_4_1, word_4_2, word_4_9, word_4_3,
    word_4_4, word_4_5, word_4_0, word_4_7, word_4_8
2 Topic #1: word_4_9, word_4_6, word_4_2, word_4_4, word_4_5,
    word_4_0, word_4_3, word_6_2, word_4_7, word_4_8
3 Topic #2: word_4_7, word_4_2, word_4_9, word_2_6, word_0_9,
    word_4_1, word_4_3, word_6_9, word_2_2, word_4_8
4 Topic #3: word_6_1, word_5_8, word_4_2, word_3_2, word_1_8,
    word_4_9, word_4_3, word_6_7, word_2_9, word_6_0
5 Topic #4: word_4_4, word_4_5, word_4_1, word_6_9, word_4_2,
    word_4_9, word_2_6, word_4_8, word_4_6, word_4_3
6 Topic #5: word_5_6, word_9_3, word_6_8, word_4_8, word_4_9,
    word_4_3, word_4_6, word_4_7, word_4_0, word_9_4
```

```
7  Topic #6: word_4_1, word_7_3, word_8_9, word_5_4, word_3_7,
       word_9_7, word_6_9, word_6_5, word_4_5, word_6_6
8  Topic #7: word_6_3, word_8_0, word_6_9, word_2_5, word_6_1,
       word_1_3, word_4_0, word_2_9, word_8_3, word_6_5
9  Topic #8: word_7_2, word_9_7, word_2_9, word_1_2, word_3_9,
       word_4_2, word_4_8, word_1_3, word_6_8, word_0_4
10 Topic #9: word_2_4, word_6_6, word_3_0, word_1_6, word_5_3,
       word_6_4, word_0_5, word_1_3, word_1_8, word_6_9
```

This representation can be reconstructed as a matrix of topic identifiers. Then, this word-topic matrix can be used to calculate the occurrence matrix. For this new matrix, jth item in the line i is equal to the number of words that correspond to the topic j.

```
1  0, 0, 0, 0, 10, 0, 0, 0, 0, 0
2  0, 0, 0, 0, 9, 0, 1, 0, 0, 0
3  1, 0, 2, 0, 6, 0, 1, 0, 0, 0
4  0, 1, 1, 1, 3, 1, 3, 0, 0, 0
5  0, 0, 1, 0, 8, 0, 1, 0, 0, 0
6  0, 0, 0, 0, 6, 1, 1, 0, 0, 2
7  0, 0, 0, 1, 2, 1, 3, 1, 1, 1
8  0, 1, 2, 0, 1, 0, 4, 0, 2, 0
9  1, 2, 1, 1, 2, 0, 1, 1, 0, 1
10 1, 3, 1, 1, 0, 1, 3, 0, 0, 0
```

To measure how well the topics are captured, the following metric is defined: a total sum of maximum values for each column divided by 100. For this example, this metric is equal to $(1+3+2+1+10+1+3+1+2+2)/100 = 0.26$. The best-case scenario is when the occurrence matrix looks the following way:

```
1   0, 10, 0, 0, 0, 0, 0, 0, 0, 0
2  10, 0, 0, 0, 0, 0, 0, 0, 0, 0
3   0, 0, 0, 0, 0, 0, 0, 10, 0, 0
4   0, 0, 0, 0, 0, 0, 10, 0, 0, 0
5   0, 0, 0, 0, 0, 0, 0, 0, 10, 0
6   0, 0, 0, 0, 0, 10, 0, 0, 0, 0
7   0, 0, 10, 0, 0, 0, 0, 0, 0, 0
8   0, 0, 0, 0, 10, 0, 0, 0, 0, 0
9   0, 0, 0, 10, 0, 0, 0, 0, 0, 0
10  0, 0, 0, 0, 0, 0, 0, 0, 0, 10
```

In this case, all 10 words that describe a topic are captured by the model correctly, and there is no overlapping in between the topics. The total score is $(10+10+10+10+10+10+10+10+10+10)/100 = 1.0$. The worst-case scenario is when all the cells of the occurrence matrix are equal to 1 (a uniform distribution of words). In this case, the total score is 0.1 which is the least possible value in this setting.

As the result, the model with total score of 0.72 is chosen as the best candidate.

Real-world Dataset

For the real-world data, hyperparameters of the HDP model are supposed to be tuned to achieve higher performance. That is done similarly to the synthetically generated

data. The only difference is the choice of the cost function to be optimized. In this case, topic coherence metric is applied to evaluate how good the topics extracted by the model are [Newman et al., 2010]. This metric computes a pairwise similarity for top K words describing a topic. Word similarity is computed by using WordNet lexical database [Miller, 1990].

As the result, the model with topic coherence of 0.76 is chosen as the best candidate. It should be stated explicitly that metrics used for both synthetic and real-world scenarios have different natures and cannot be compared to each other.

6.4 Evaluation Procedures

In a general context, there are several objective ways to evaluate how well anomaly detection algorithms perform on the test set. They are usually based on various assumptions that are used to define the concept of anomaly itself. For instance, clustering-based anomaly detection algorithms are based on the assumption that normal data points belong to a cluster, whereas data point which do not belong to any cluster are considered anomalous. Following this assumption, several objective performance metrics can be designed. One metric can utilize cluster labels provided by synthetically generated data as a ground truth to estimate how many anomalies the algorithm detects, and apply standard performance metrics such as accuracy, precision, recall or F-score. Another metric might be a silhouette score to estimate how well the algorithm clusters the data. Unfortunately, in the context of this thesis project and the designed framework, the task of defining objective performance metrics becomes highly unlikely to achieve. That is due to the absence of the definition of an "anomalous document". In other words, there are no means of defining a document as wrong or incorrectly written. All documents in the real-world text corpus are valid pieces of scientific literature.

Three-level hierarchy of anomalies introduced in this work, on the other hand, is used to define factors that indicate potential *early signals* of emerging trends. But the final decision whether a particular signal leads to a new emerging trend is on analysts and domain experts. Another problem is that the amount of time required for a specific topic to become a well-established trend is unknown in most of the cases. For instance, the concept of neural networks as mathematical models is known for almost 50 years. However, this research topic has become a hot trend only during the last decade. This transition occurred because of the reasons (technological advances) that cannot be anticipated automatically within the scope of the data available. See the black swan theory for more details on this topic [Nicholas Taleb, 2015]. With this being said, performance evaluation of the designed framework based on the real-world data is considered subjective and being estimated on a conceptual level. Detecting group anomalies based on high-density regions in the data space and concept drift are inspected visually.

Performance evaluation of the framework based on synthetically generated data, where ground truth is available, is done by repeating simulations and calculating precision of the model to estimate a false alarm rate.

6.5 Results

As the result, a prototype of the designed framework is implemented which works in the following way. As the new batch of documents becomes available, it is being automatically processed. First, word-level anomalies are detected which represent new vocabulary. In order to detect this type of anomalies, all the upcoming terms are compared with the exysting one. In case a new term arrives, a notification is being raised. Secondly, if metadata which adds context to the dataset is available (such as user profile for posts in social media), contextual anomalies are being captured. The context specifies a subsample of the dataset which is analyzed further (reduced dimensionality). For this scenario, terms that occur rarely in the given context are detected.

For document-level anomalies, DBSCAN is used to detect high-density groups of doc-

uments with similar word distribution as potential early signals of emerging trends. KL divergence is then applied to assign anomaly score to each detected group. If the value is greater than the threshold, a notification is raised. The same approach is used to detect high-density groups of documents with similar topic-distribution. The only difference is the input data – for the first case documents as a Bag-Of-Words are taken, whereas the second case takes documents as a distribution of topics extracted by HDP. Since text collections have temporal structure, documents from several neighboring epochs are analyzed together (in this project the 4-epoch long window function is used). This approach allows two neighboring frames to have overlapping documents which facilitate tracking of all the changes over time.

To make it clear, two neighboring frames – documents from week 4-8 and week 5-9 – will have the same documents from week 5-8. These documents allow to correlate labels of different clusters over time and monitor dynamics of each particular cluster.

As the last step of this solution, KL divergence is applied to each topic (word distribution that describes each extracted topic) from two neighboring frames. If the value is greater than the threshold, new notification is raised. That is used to capture temporal changes in topic description which allows to track concept drifts of each topic.

It is worth stating explicitly that thresholds mentioned above are hyperparameters for this framework and used to tune its sensitivity. Tuning of each threshold is corpus-specific and requires manual analysis done by the operator. Figure 3 represents several cases when particular high-density groups of documents are marked as potential emerging trends or not, depending on the value of the threshold. Another point is that the final decision whether the notification corresponds to the future emerging trend is done by the operator.

Illustration

Several examples are given below as illustration of what cases the designed framework is supposed to detect. The first example is related to the concept of generative adversarial networks. By analyzing the original article, the designed solution will detect an abbreviation “GAN” as a new term, as well as the concept “generative_adversarial_network” extracted by phrase detector. Note that each separate word - “generative”, “adversarial”, and “network” - is already present in vocabulary thus no notifications will be raised.

When it comes to document-level anomalies, a group of documents that describe topic modeling algorithms in application to single-cell analysis and RNA sequencing will be detected as potential early signal of an emerging trend due to the deviating combination of topics – topic modeling and text analysis combined with single-cell analysis and biology.

At the last level of the hierarchy, burst of scientific articles that describe convolutional and recurrent networks will shift a word distribution for the topic “neural networks” from the description of primarily feed-forward neural networks towards a more diverse and informative topic. Over time, there might be a topic split to occur due to the high popularity of different variations and applications of convolutional neural networks.

Evaluation on Synthetic Data

As was mentioned earlier in the report, evaluation of the word-level anomalies is unnecessary because the process of detecting new vocabulary is trivial. Detection of concept

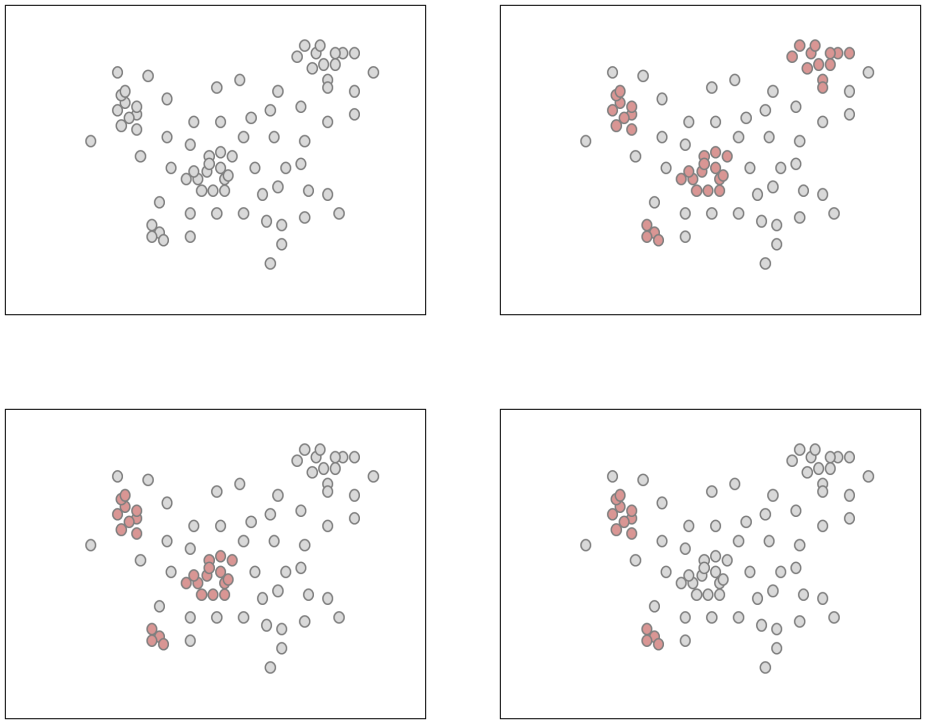

Figure 3: *Graphical representation of group anomalies captured by DBSCAN and marked by using KL divergence, depending on the threshold applied to KL divergence.*

drift is also trivial since it compares changes in the word distribution for a given topic. Evaluation of how meaningful these changes are cannot be done on synthetic data because the "words" do not convey any semantic meaning. Detection of documents that belong to high-density regions is performed with mean precision of 85% (+- 7%). It is worth mentioning that all the dense regions are correctly identified. Precision is lower than 100% due to the fct that the clustering algorithm misclassifies data points which are located on the edge of the cluster. The conclusion can be made that the designed framework performs sufficiently well on detection of high-density clusters with respect to the potential confirmation bias.

Evaluation on arXiv and NIPS Datasets

Similar results are achieved for the real-world data. Word-level anomalies and concept drift are detected because the algorithms are rather trivial. Document-level anomalies are captured in a sufficient way based on the visual inspection. It is worth mentioning that the quality of DBSCAN algorithm depends on hyperparameters which have to be tuned each particular dataset in order to achieve higher performance. Applied to `arXiv` and `NIPS` datasets, clustering algorithm detects groups starting with 8-10 documents in it. Figure 4 represents a snapshot of the `arXiv` dataset (a time frame of 4 weeks) used to demonstrate high-density clusters of documents which might capture early signals of emerging trends. The minimal size of a cluster is claimed to be low enough to consider the framework rather sensitive for the given settings. On the other hand, high sensitivity leads to a high false alarm rate. This conclusion is made based on visual inspection of the results produced by the designed framework, and on the limited expertise of the

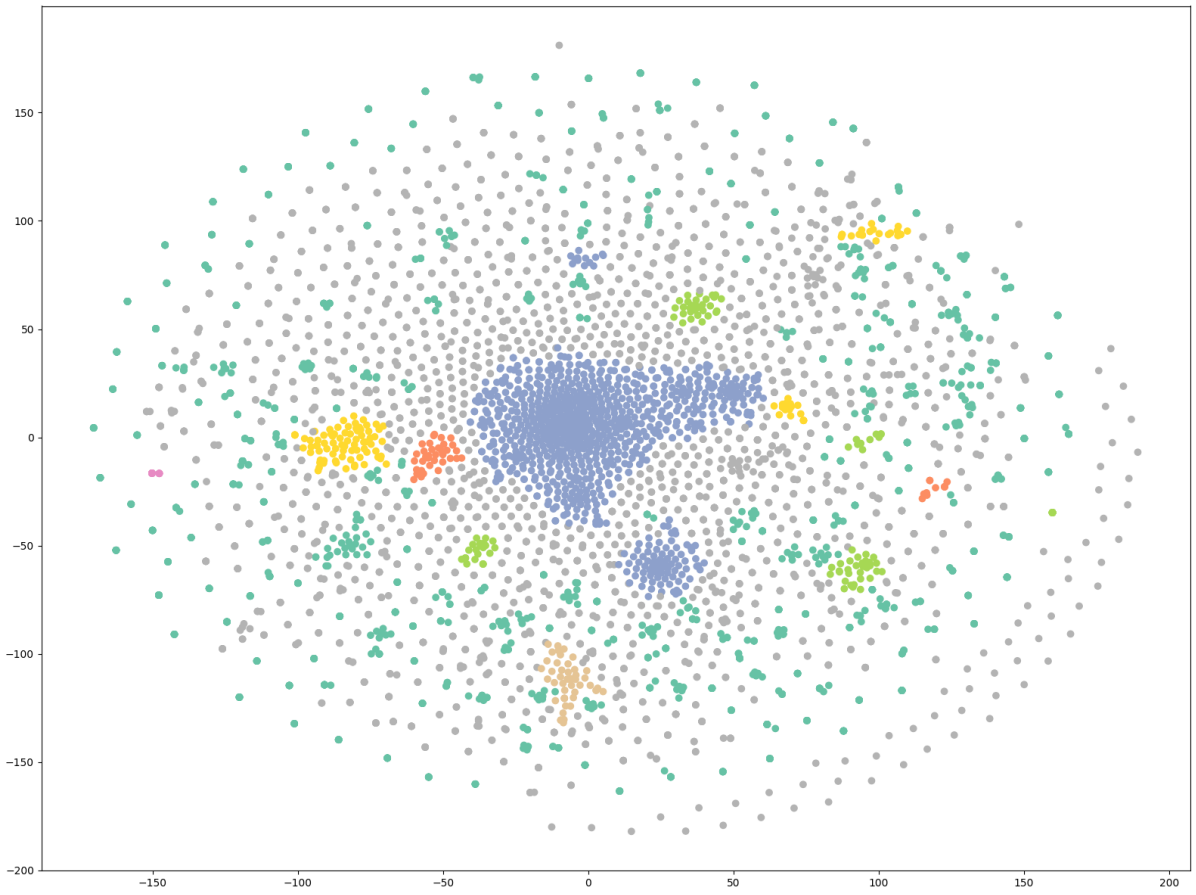

Figure 4: *T-SNE visualization of one snapshot of data from* `arXiv` *dataset (time frame - 4 weeks) used for visual inspection.*

author in the subject area of computer science that the articles are published in. Since KL divergence is used to assign anomaly scores on each level of the proposed hierarchy, this sensitivity can be tuned for a particular dataset by leveraging the threshold used to decide whether an alert should be raised. An operator, analyst or domain expert can adjust this threshold in order to filter out the alert that lead to a false alarm.

7 Related Work

To the best of the author's knowledge, currently there is few research being conducted in the subject area of trend prediction in text that focuses on early signals of emerging trends. Huang et al. [2017] proposed a model for emerging topic tracking based on local weighted linear regression (LWLR) which is used to estimate word novelty and fading. Three challenges are identified in this work – detecting emerging trends as early as possible, track its evolution, and present topics with coherent words. A new emerging topic tracking algorithm (ETT) is proposed that generates topics based on the word co-occurence and estimates words novelty. Similarly to the statements made in this thesis project, Huang et al. [2017] highlights the difficulties of evaluating the model empirically based on the real-world streaming data without any labeling. Synthetic stream is designed by using a crawling system that samples event-relevant feed that can be labeled. It also utilizes a time frame of the fixed length (1 day) and applies similar data preprocessing techniques. It uses topic coherence as a performance indicator.

Hurtado et al. [2015, 2016] proposed a model for topic discovery and future trend forecasting by combining association rule mining and ensemble forecasting for trend prediction. This work focuses on time-series analysis and fits a regression model to predict a short-term estimation of a future trend. Compared to Hurtado et al. [2016], the designed framework is claimed to provide alerts regarding emerging trends earlier than association rule mining. The reason behind this claim is that for the short-term prediction the topic (trend) has to be already established. This work provides a more comprehensive performance evaluation and uses 4 different datasets to validate the results.

There is an extensive amount of research done in the area of temporal topic modeling [Lafferty and Blei, 2006; Wang et al., 2012; Elshamy, 2013]. The state-of-the-art model is called Hierarchical Evolving Dirichlet Process (EHDP) [Wang et al., 2017]. It supports hierarchical dependencies in text corpora and tracks dynamic changes in the topic space over time. Similarly to Hurtado et al. [2016], this model requires a topic to be well-established in order to capture and process it further. It also provides a comprehensive performance evaluation of the model based on 4 datasets as well as a graphical representation of the collected results.

There are several research articles that focus on detection of deviating documents in text corpora. Xiong et al. [2011] proposed a hierarchical probabilistic model for group anomaly detection. It uses a standard approach for group anomaly detection – to model the data first and to mark the data which is unlikely to be generated from that model. In this case, kNN algorithm is applied to assign anomaly scores for both synthetic and real-world astronomical data. Zhuang et al. [2017] introduced penalization of lexically general words in order to identify semantically deviating documents. A case study was conducted to evaluate how the proposed model outperformed all the baselines. Compared to these models, the designed framework has lower reliability of the empirical evaluation process but is more detailed and complete on the conceptual level.

8 Discussion

This section briefly describes limitations of the work, related ethical issues and promising future directions.

8.1 Limitations

There are several limitations of this work. First of all, there are no objective performance measures defined for empirical evaluation of the designed framework on the real-world data. Although performance evaluation on a conceptual level is provided, there is a lack of objective estimation of how well the design framework performs on emerging trend prediction. That would make the outcomes of the evaluation process more credible. Secondly, topic-level anomalies and signals of concept drift are limited to the simple shifts in a word distribution that describes a particular topic. However, concept drift might have more complex nature including nonlinear and hierarchical dependencies. These aspects are not covered within the scope of this work. Thirdly, applying a window function of a particular size in order to segment the data works under one assumption that changes in patterns are gradual and trends are emerging consistently. However, there might be a chance to face documents of a particular word and topic distribution that appear in the corpus with a pulsing pattern. In case pulsation frequency is lower than the size of the window function, similar documents of this kind will not be stacked together and thus will not lead to an alert. The next problem is related to selection bias – specificity of the real-world data does not allow one to claim regarding the generalization of the framework [Wohlin et al., 2012]. There might be application domain or settings where the framework achieves lower performance or simply does not work. A comprehensive case study is required to draw conclusions of this sort. Finally, there are no analytical way of justifying that the three-level hierarchy of anomalies and the proposed framework itself are complete. Thus, all claims regarding predictive capabilities of the framework are rather subjective and supposed to be treated with a certain level of scepticism.

8.2 Future Directions

This thesis project has various promising directions for future work. Since the designed framework, in theory, is not limited by any particular topic modeling or anomaly detection algorithm, it is worth applying more complicated models in order to achieve higher predictive capability or incorporate more meta data that describe the documents. As an example, Author Topic-Model can be used to integrate data about authors into the framework [Rosen-Zvi et al., 2012].

As was mentioned in section 8.1, a part of the framework that is related to topic-level anomalies and concept drift might be extended to more complicated topic modeling algorithms. This can enable capturing of early signals that are related to hierarchical changes in the topic space such as topic merge, split, or death.

Yet another direction for future research is related to generalization of the model. High-abstract idea of incorporating algorithms for data structuring and anomaly detection might benefit the application areas of fraud and malicious activity detection, patent

monitoring, predictive maintenance, emerging tools in pharmaceutics and single-cell research where DNA sequences are used as text.

Finally, threats to validity can be addressed in a more comprehensive way. A case study might be required in order to validate whether there are selection or confirmation biases (choice of the datasets and evaluation metrics), or whether this study might be generalized to broader populations of textual data.

8.3 Ethical Issues

When it comes to ethical considerations in research, Ethical Principles of Psychologists and Code of Conduct prepared by American Psychological Associa are the most credible guidelines to follow (see Hobbs [1948]; Canter et al. [1994]). Although APA mainly focuses on research with human participants, there are several problems presented which are related to this work – protection of intellectual property and data privacy.

Copyright grants authors and producers protection for their creations, allowing them to hold exclusive right to authorize or prohibit translation, adaptation, broadcasting and other forms of reproduction of their intellectual property [of WIPO, 2003]. Transferring these rights is usually done in return for compensation of some sort – most likely, the monetary one. There is a need to verify that the process of data collection does not violate copyright laws. Thus, for this thesis project, scientific articles from the open-access data sources are being used. However, for any further usage of the designed framework, data is supposed to be collected under compliance of the research principles and copyright laws.

When it comes to data privacy, Solove [2005] provides a brief discussion on applicable ethical issues, and formulates a taxonomy of privacy based on various types of harmful activities:

- data collection – surveillance and interrogation;
- data processing – aggregation, identification, insecurity, exclusion, secondary use;
- information dissemination – breach of confidentiality, disclosure, increased accessibility, appropriation, distortion, blackmail; and
- invasion – intrusion and decisional interference [Solove, 2005, p. 491].

Several of these harmful activities might be applicable to this thesis project. For data collection, which does not involve interpersonal communication, such as data mining techniques, there is a risk of this process to be considered surveillance. This risk is neglected given the source of data used for empirical evaluation, but can be present when it comes to the analysis of emails or posts from various social networks. Automated web crawling is prohibited by many web services including ACM digital library and different publishing agencies. Choosing data sources for further analysis has to be made taking these notes into account.

Aggregation and processing of collected data can potentially violate data privacy. Harmful activities such as insecurity and breach of confidentiality are neglected because of the open access nature of the data. However, it should be mentioned that "[even if] a

piece of information here and there is not very telling [...], aggregated information can reveal new facts about a person that she did not expect would be known about her when the original data was collected" Solove [2005]. Activities such as aggregation, identification and secondary use are potentially dangerous for the given case. Therefore, it is vital to analyze hypothetical ethical problems while collecting, storing, sharing or processing data of any kind in order to protect privacy.

9 Conclusions

This thesis project addresses the problem of emerging trend prediction in text. It presents a theoretical framework and a prototype solution that is capable of detecting early signals of emerging trends in text. The framework utilizes topic modeling algorithms (Hierarchical Dirichlet Process) to represent latent topic space of a given temporal collection of documents, and applies several anomaly detection algorithms in order to capture deviating text which might indicate birth of a new not-yet-seen phenomenon. The prototype is empirically evaluated on both synthetically generated corpora and real-world text collections from `arXiv.org` and `NIPS` publications. For synthetic data, a text generator is designed which provides ground truth to evaluate the performance of anomaly detection algorithms.

This work contributes to the body of knowledge in the area of emerging trend prediction in several ways. First of all, the method of incorporating topic modeling and anomaly detection algorithms for emerging trend prediction is a novel approach and highlights new perspectives in the subject area. Secondly, the three-level word-document-topic hierarchy of anomalies is formalized in order to detect anomalies in temporal text collections which might lead to emerging trends. Finally, a framework for unsupervised detection of early signals of emerging trends in text is designed. The framework captures new vocabulary, documents with deviating word/topic distribution, and drifts in latent topic space as three main indicators of a novel phenomenon to occur, in accordance with the three-level hierarchy of anomalies. The framework is not limited by particular sources of data and can be applied to any temporal text collections in combination with any online methods for soft clustering.

As for the future directions, it is necessary to conduct performance evaluation of the designed framework based on objective metrics and descriptive statistics. Another direction is to apply more complicated topic modeling algorithms, and mitigate validity threats related to generalization issues and confirmation bias.

Achieving high performance in unsupervised prediction of emerging trends in text can indicate promising directions for future research and potentially lead to breakthrough discoveries in any field of science.

A Literature Search

In order to contribute to the reproducibility of this work, several notes on literature search are presented. First of all, below is the list of search queries used in order to retrieve scientific articles related to the subject area.

- "(emerging | emergent) & trend";
- "trend & (prediction | forecasting)";
- "trend & (prediction | forecasting) & text";
- "future & trend & text";
- "topic & (model | modeling)";
- "topic & (model | modeling) & anomaly & detection";
- "concept & drift";
- "concept & drift & survey";
- "concept & drift & text";
- "anomaly & detection";
- "anomaly & detection & survey";
- "anomaly & detection & text".

These search queries and their combinations are used in several databases and search engines: `ScienceDirect`, `Web of Science`, `Google Scholar`, `Google` search engine, `Scopus`, `Elsevier`, `arXiv.org`, `DiVA` portal and `LIBRIS`. For `ScienceDirect` and `Web of Science`, automatic alerts were set up to keep track of new scientific articles published within the time frame of this thesis project. For those who are willing to reproduce the work conducted, it is recommended to follow these search queries in order to get an up-to-date list of relevant publications.

References

Abbas, K., Shang, M., Luo, X., and Abbasi, A. (2017). Emerging trends in evolving networks: Recent behaviour dominant and non-dominant model. *Physica A: Statistical Mechanics and its Applications*, 484:506 – 515.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

Anderson, D. (2002). Model selection and multimodel inference: a practical information-theoretic approach. *Springer-Verlag, New York, New York, USA. JACKSON ELK POPULATION DYNAMICS'Mil/ow and Smith J. Wildl. Manage*, 68(4):2004.

Ankerst, M., Breunig, M. M., peter Kriegel, H., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. pages 49–60. ACM Press.

Belford, M., Namee, B. M., and Greene, D. (2017). Stability of topic modeling via matrix factorization. *CoRR*, abs/1702.07186.

Bello-Orgaz, G., Jung, J. J., and Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45 – 59.

Berlanga-Llavori, R., Anaya-Sánchez, H., Pons-Porrata, A., and Jiménez-Ruiz, E. (2008). Conceptual subtopic identification in the medical domain. In *Ibero-American Conference on Artificial Intelligence*, pages 312–321. Springer.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Canter, M. B., Bennett, B. E., Jones, S. E., and Nagy, T. F. (1994). *Ethics for psychologists: A commentary on the APA Ethics Code.* American Psychological Association.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58.

Dasu, T., Krishnan, S., Venkatasubramanian, S., and Yi, K. (2006). An information-theoretic approach to detecting changes in multi-dimensional data streams. In *In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications*. Citeseer.

de Finetti, B. (2016). *Theory of Probability: A critical introductory treatment.*

Drechsel, T. and Tenreyro, S. (2017). Commodity booms and busts in emerging economies. *Journal of International Economics.*

Dutta, A. (2018). Implied volatility linkages between the u.s. and emerging equity markets: A note. *Global Finance Journal*, 35:138 – 146.

Elshamy, W. (2013). Continuous-time infinite dynamic topic models. *CoRR*, abs/1302.7088.

Enríquez, J., Domínguez-Mayo, F., Escalona, M., Ross, M., and Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. *Expert Systems with Applications*, 80:14 – 27.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Gama, J. a., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37.

Hinneburg, A. and Keim, D. A. (2003). A general approach to clustering in large databases with noise. *Knowledge and Information Systems*, 5(4):387–415.

Hinneburg, A., Keim, D. A., et al. (1998). An efficient approach to clustering in large multimedia databases with noise. In *KDD*, volume 98, pages 58–65.

Hobbs, N. (1948). The development of a code of ethical standards for psychology. *American Psychologist*, 3(3):80.

Hong, S., Zhou, Z., Zio, E., and Wang, W. (2014). An adaptive method for health trend prediction of rotating bearings. *Digital Signal Processing*, 35:117 – 123.

Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., and Zhang, X. (2017). A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web*, 20(2):325–350.

Hurtado, J., Huang, S., and Zhu, X. (2015). Topic discovery and future trend prediction using association analysis and ensemble forecasting. In *2015 IEEE International Conference on Information Reuse and Integration, IRI 2015, San Francisco, CA, USA, August 13-15, 2015*, pages 203–206.

Hurtado, J. L., Agarwal, A., and Zhu, X. (2016). Topic discovery and future trend forecasting for texts. *Journal of Big Data*, 3(1):7.

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., and Phelps, D. J. (2004). *A Survey of Emerging Trend Detection in Textual Data Mining*, pages 185–224. Springer New York, New York, NY.

Kuechler, B. and Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. *European Journal of Information Systems*, 17(5):489–504.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lafferty, J. D. and Blei, D. M. (2006). Correlated topic models. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 147–154. MIT Press.

Liu, X., Jiang, T., and Ma, F. (2013). Collective dynamics in knowledge networks: Emerging trends analysis. *Journal of Informetrics*, 7(2):425 – 438.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

March, S. T. and Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4):251 – 266.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*.

Mörchen, F., Dejori, M., Fradkin, D., Etienne, J., Wachmann, B., and Bundschus, M. (2008). Anticipating annotations and emerging trends in biomedical literature. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 954–962. ACM.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Nicholas Taleb, N. (2015). The black swan: The impact of the highly improbable. *Victoria*, 250:595–7955.

Oates, B. (2006). *Researching Information Systems and Computing*. SAGE Publications.

of WIPO, I. B. (2003). What is intellectual property?

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rosen-Zvi, M., Griffiths, T. L., Steyvers, M., and Smyth, P. (2012). The author-topic model for authors and documents. *CoRR*, abs/1207.4169.

Sebastião, R. and Gama, J. (2007). Change detection in learning histograms from data streams. In *Portuguese Conference on Artificial Intelligence*, pages 112–123. Springer.

Solove, D. J. (2005). A taxonomy of privacy. *U. Pa. L. Rev.*, 154:477.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

Wang, C., Blei, D. M., and Heckerman, D. (2012). Continuous time dynamic topic models. *CoRR*, abs/1206.3298.

Wang, C., Paisley, J., and Blei, D. M. (2011). Online variational inference for the hierarchical dirichlet process. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 752–760, Fort Lauderdale, FL, USA. PMLR.

Wang, P., Zhang, P., Zhou, C., Li, Z., and Yang, H. (2017). Hierarchical evolving dirichlet processes for modeling nonlinear evolutionary traces in temporal data. *Data Mining and Knowledge Discovery*, 31(1):32–64.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Xiong, L., Poczos, B., Schneider, J., Connolly, A., and VanderPlas, J. (2011). Hierarchical probabilistic models for group anomaly detection. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 789–797, Fort Lauderdale, FL, USA. PMLR.

Yu, L., Zhao, Y., Tang, L., and Yang, Z. (2018). Online big data-driven oil consumption forecasting with google trends. *International Journal of Forecasting*.

Zhang, X.-D., Li, A., and Pan, R. (2016). Stock trend prediction based on a new status box method and adaboost probabilistic support vector machine. *Applied Soft Computing*, 49:385–398.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52.

Zhuang, H., Wang, C., Tao, F., Kaplan, L., and Han, J. (2017). Identifying semantically deviating outlier documents. In *Proceeding of 2017 Conference on Empirical Methods in Natural Language Processing*, page 2738–2747. Association for Computational Linguistics.