2022

Low-Resource Machine Learning Techniques for the Analysis of Online Social Media Textual Data

Toktam Amanzadeh Oghaz
University of Central Florida

LOW-RESOURCE MACHINE LEARNING TECHNIQUES FOR THE ANALYSIS OF ONLINE SOCIAL MEDIA TEXTUAL DATA

by

TOKTAM AMANZADEH OGHAZ
M.Sc University of Central Florida, 2017

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2022

Major Professor: Ivan Garibay

ABSTRACT

Low-resource and label-efficient machine learning methods can be described as the family of statistical and machine learning techniques that can achieve high performance without needing a substantial amount of labeled data. These methods include both unsupervised learning techniques, such as LDA, and supervised methods, such as active learning, each providing different benefits. Thus, this dissertation is devoted to the design and analysis of unsupervised and supervised techniques to provide solutions for the following problems: 1. Unsupervised narrative summary extraction for social media content, 2. Social media text classification with Active Learning (AL), 3. Investigating restrictions and benefits of using Curriculum Learning (CL) for social media text classification. For the first problem, we present a framework that can identify the viral topics over time and provide a narrative summary for the identified topics in an unsupervised manner. Our framework can provide such information with varying time resolution. For the second problem, we present a strategy that conducts data sampling based on the local structures in the embedding space of a large pretrained language model. The data selection for annotation is conducted for the data samples that do not belong to a dominant set as these samples are less similar to the rest of the data points, and accordingly, are more challenging for the model. This criterion is a compelling technique that minimizes the need for large annotated datasets. Then for the third problem, we consider similar data difficulty notions to study the impacts of learning from such a curriculum to train models from easy samples first. This is opposite to the idea of active learning. However, instead of learning from a small number of data and disregarding a substantial amount of information, gradual training from easy samples leads to learning a trajectory to a better local minimum. Our study includes curricula based on both heuristics and model-derived.

To my family for their consistent support

ACKNOWLEDGMENTS

I wish to acknowledge the help provided by my advisor, Dr. Ivan Garibay of the Department of Industrial Engineering and Management Systems as well as the Department of Computer Science at the University of Central Florida. His consistently insightful comments and encouraging me during this process have helped me to prepare this thesis.

I wish to extend my special thanks to my committee members, Dr. Gita Sukthankar, Dr. Damla Turgut, and Dr. Niloofar Yousefi for their guidance and support. The completion of this dissertation would not be possible without your constructive criticism and insight. Particularly, Dr. Niloofar Yousefi devoted a substantial effort and played a decisive role in mentoring me in the early phase of my Ph.D.

I also would like to express my gratitude to my research colleagues and co-authors, Dr. Ece Mutlu and Jasser Jasser for their invaluable contributions. They specially impacted my research direction in the early phases of my doctoral study.

Finally, I would like to express my most profound appreciation to my family, specially, my parents, for believing in me and supporting me unconditionally. My deepest gratitude goes to my best friend and husband, Alireza, who always supported me, patiently encouraged me throughout this journey, and believed in my work. Thank you, Alireza.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Social media and microblogging platforms, such as Twitter and Facebook, provide a substantial amount of publicly available real-time information regarding significant global events, specially in the format of textual content [99]. As a result, the field of Natural Language Processing (NLP) has observed the emergence of many AI-enabled solutions and systems related to social media text analysis and mining. Machine learning development, however, is mostly focused on rich standard scenarios where a large amount of high-quality annotated data is available for the training of models, specially for data-hungry deep learning architectures [79]. Although a massive amount of user-generated social media content is publicly available from social media platforms, such as Twitter and Facebook, is cheap and many tools are designed to download such datasets, manual annotations for only a small portion of such data can be obtained [1]. Similarly, many real-life tasks and domains lack large-scale data for development and training [60].

Low resource and label-efficient machine learning methods and analysis can be described as the family of statistical and machine learning techniques that can be trained in a label-efficient manner, in which significant labeled data is not required for training [59, 58, 85, 1]. Although lack of sufficient annotated data is the most common definition of low-resource setting for machine learning, other cases are analyzing low-resource languages such as threatened languages, non-standard domains, non-standard and uncommon tasks, lack of unlabeled data, lack of auxiliary data, and computational power limitations [58, 59, 60].

Research on low-resource scenarios and machine learning techniques and models designed specifically for such settings are motivated by many factors, including the time-consuming process of data annotation, labor inefficiency of the annotation process, and the cheap collection of sufficiently large unlabeled data for most tasks [59]. Specifically, big data annotation is labor inefficient as ex-

pert annotators with domain knowledge and expertise are expensive, whereas, cheap labeling is possible through unskilled annotators with the consequence of low-quality and noisy labels [79]. As a result, using other sources has been of great interest in literature, such as using unlabeled data, lexicon and grammar [99], extracting relations, parsing [135], defining manual heuristics [17], domain knowledge [75], general knowledge transfer learning by using embeddings and pretrained models [1, 33], and transfer learning from related high-resource settings [1], and meta learning [79].

Annotated datasets to study social media are extremely limited, but an extensive literature on this topic exists. Among the most significant research areas in this field are the research on moderating online social platforms to protect the users from deliberate misinformation, toxicity, cyberbullying, and trolling, for instance, finding the indicator attributes [87, 29], preparing manually labeled datasets for disinformation detection and public stances on such content [97], classifying offensive language [123, 95], and toxicity type detection [38, 30, 152]. This is however, literature on the limitations of current traditional and deep language models for social media analysis, as well as the techniques to overcome their deficiencies with limited datasets are handful.

In the field of NLP, novel and revolutionary unsupervised techniques, such as Latent Dirichlet Allocation (LDA) [10] was developed to tackle the problem of labeled data shortage. LDA considers the co-occurrence of words in documents. This method is a statistical technique that discovers the relevant structure and co-occurrence dependencies of words within a collection of documents to capture the distribution of topic latent variables from the data. A substantial number of document modeling techniques are later derived from the LDA method, such as the ones that exploit additional sources of information to improve document modeling [155, 158]. Due to their unsupervised nature, these research efforts enable training of general-purpose systems that can be used for a variety of tasks and applications as strong classifiers [15].

Despite the fact that unsupervised learning methods seem to be ideal as they do not require any labeled data, there are lots of benefits in developing robust yet data-efficient supervised techniques, including achieving higher accuracy and reliability. Therefore, labeled data scarcity has encouraged an extensive amount of work by the machine learning research community in order to develop models that can generalize their learned knowledge from little labeled data. Examples of such research areas are few-shot learning, transfer learning, and data selection for active learning.

Meanwhile, the development of pre-trained language models, such as Google's BERT (Bidirectional Encoder Representations from Transformers) architecture [32] and OpenAI's GPT-3 (Generative Pre-trained Transformer) model [13], which are trained over massive examples of written language has revolutionized the field of natural language understanding. Through knowledge sharing, these pre-trained models allow using a small amount of labeled data for fine-tuning the model for a downstream task [33], which makes these models ideal to be used with label-efficient learning techniques.

1.1 Purpose of this Study

This work contributes to the research on machine learning methods for online social media textual content by proposing both supervised and unsupervised learning techniques. This work examines the efficiency and performance of the proposed methods via extensive analysis of the online social media textual content, including social media content containing misinformation-and offensive language. As the analysis and solutions designed for the detection of such content must tackle data imbalance, labeled data shortage, noisy and poor labels, etc., investigating content related to these domains are of our interest. Specifically, the purpose of this work is to:

I) Identify and investigate the emergence and virality of narratives around topical events in

Twitter and detect significant recurrent topics via an LDA-based generative topic model;

II) Investigate the potential of maximally cohesive structures in the embedding space of a deep language model (e.g. the BERT model) as the informative data selection criterion for fine-tuning using active learning;

III) Investigate the benefits and restrictions of incorporating curricula in low-resource imbalanced scenarios and with active learning for text classification based on data structures in the embedding space of a deep language model (e.g. the BERT model).

1.2 Research Questions

RQ.I) *How LDA-based topic modeling can be used to identify the rise and fall of topic popularity over time for social media textual data, which can be multimodal and sparse in time?* We further investigate: (a) lifetime attractiveness of topics using Shannon entropy; and (b) topic summarization to extract narratives.

RQ.II) *Does the structure of the embedding space of a deep pre-trained language model provide information for data selection in active learning? Does data selection with this technique lead to higher performance in comparison to other active learning techniques?* We further investigate: (a) the impact of increasing the number of trainable parameters, (b) maximum allowed sequence length and padding, and (c) the initial clustering method.

RQ.III) *Assuming the information on data difficulty can be derived from the structure of the embedding space of a deep pre-trained language model, does applying a curriculum learning method with such a notion of data difficulty improve the performance in fine-tuning a big language model?* We further investigate: (a) the performance of this method on both social media and other text corpora; and (b) using simple heuristics determining text difficulty.

In the following, each of the research goals discussed earlier is explained in detail and the contributions made by this doctoral thesis are outlined. Further details on the specific methodology related to each of these works are provided in Chapters 3, 4, and 5.

1.3 Probabilistic Modeling of Timestamped Social Media Data

As online social media and microblogging platforms are becoming the primary sources of real-time information on significant events, and the difficulties associated with such a fast-changing environment, designing methods that can facilitate communicating the main underlying ideas seems to be practical. A possible solution can be a narrative modeling framework that detects the topically-related content associated with varying time intervals, and specially one that allows to identify the periodic and recurrent stories associated with related events as each narrative may contain story pieces from different times.

Although abundant timestamped textual data, particularly from social media platforms and news reports are available for analysis, and that these datasets can contain multiple modalities across time, analyzing the changes in the distribution of data over time have been neglected in literature [145]. This is however modeling topics without considering the text-time relationship lead to missing the rise and fall of topics over time, the changes in terms of correlations, and the emergence of new topics and stories [40].

This work is interested in the design of a narrative modeling framework that matches the definition of narrative in literature as: i) narrative summaries can be constructed from an ordered chain of individual events with causality relationships amongst events, appeared within a specific topic; and ii) the narrative sequence may report fluctuations over time relative to the underlying events [68]; as this design allows a story-like interpretation of the text, which is a must to imply a narrative

[106].

Although continuous-time topic models such as [145] have been proposed in the literature, topical models with continuous-time distribution cannot model many modes in time, which leads to deficiency in modeling the fluctuations. Additionally, continuous-time models suffer from instability problems in the case of analyzing a multimodal dataset that is sparse in time. As studying datasets related to time-series activities on online social media platforms necessitates resolving multi-modality and sparsity, continuous-time topic models cannot accurately model the rises and falls in the distribution of topics for these platforms.

Based on the aforementioned design goals for narrative modeling, this doctoral thesis introduces an event-based narrative summary extraction framework that can identify topically-related narratives from online social media textual data. The details on the design of this method are discussed in Chapter 3. In our approach, topic discovery is influenced by both word co-occurrence and temporal information, and the model captures topic recurrence as a result of long-range dependencies in time. The significance of this framework is its capability to extract relevant sequences of text relative to the corresponding series of events associated with the same topic over time.

To achieve probabilistic modeling of narratives over topical trends, we incorporate the components of narratives including named-entities and temporal-causal coherence between events into our design. The framework containing the probabilistic topic modeling and extractive text summarization modules results in the unsupervised topic mining of narratives and to produce their summaries. Furthermore, our design allows the identification of such distributions with varying time resolutions, e.g., weekly or monthly. Via extensive analysis and comparison, we argue that this design leads to the identification of topic distributions that also approximates the user activity fluctuations over time.

Despite the existence of basic statistical methods to inspect the number of active individuals in

a topic over time, recognizing the attractiveness of a topic within a time duration based on the latent topic variable has not been investigated in literature. This doctoral work is interested in the analysis of this factor and to deliberate its applications regarding the identified topics related to a misinformation domain. The significance of such measurement is evident when speculating basic statistical information, such as the number of involved individuals in a topic across time, is not feasible as the dataset lack information regarding the number of individuals in the conversations at their production time. Yet, in the case of existence of such information, hard assignments of topics to textual data would be required to infer the attractiveness of topics across time. Whereas, our method can reveal the statistical structure related to the topics within and across document collections. Thus, Another contribution of this work is the introduction of this metric, called the significance-dispersity trade-off (SDT), which is an entropic measure to compare the identified topic distributions over time based on their lifetime attractiveness.

We evaluate our model on a large corpus of Twitter data, including more than one million tweets related to a disinformation campaign.

1.4 Active Learning Strategy based on Maximally Cohesive Structures in the Embedding Space

Manually annotating a sufficiently large dataset for the training of a machine learning model is expensive. However, for many tasks collecting a large unlabeled textual corpus is relatively cheap. Additionally, big data storage and processing are costly. When dealing with the datasets that are used for this doctoral thesis as well as other misinformation-related domains or toxic content, the annotators may undergo extra discomfort as they get exposed to offensive and abusive content. Hate speech is an example of such toxic content that can disturb the annotators. Additionally, annotation mistakes and its difficulties might get intensified as the annotators need to process a large amount of data. Furthermore, a higher probability of annotation bias for such datasets have

been observed, i.g., bias toward some attributes and races [95, 24]. Thus, NLP methods that can make the best use of significantly less labeled data points are of great interest. Active Learning (AL) techniques can mitigate the issues associated with manual labeling and improve automatic detection and classification when labeled training data is sparse [33].

The goal of active learning is to reduce the cost of labeling via using a small number of labeled samples for training, and query the class labels for the most informative subset of the data samples that are selected using an acquisition function. A practical active learning strategy must lead to the selection of certain unlabeled data samples that can lead to the maximal reduction of the classification error and variance. As a result, recently the field of natural language processing has observed the development of many active learning approaches for different machine learning applications as well as text classification [89, 165, 166] and toxicity detection [12].

The development of pre-trained language models, such as the BERT (Bidirectional Encoder Representations from Transformers) architecture [32], which is trained in an unsupervised fashion using a massive amount of textual data has enabled transfer learning in natural language processing. The transformer architecture [142] moved the boundaries for language models and obtained the highest performance across various tasks, and the BERT model built solely on transformers, shed light on fine-tuning of a big neural language model for a downstream task using a small amount of data. These NLP achievements simply can be exploited in low-recourse label-efficient techniques, such as in active learning. Thus, the aim of this work is to present a novel pool-based active learning method that uses the embedding space of pre-trained language models to minimize the annotation cost and achieve high performance. The details on the implementation and setup of our proposed model are explained in Chapter 4. This approach can be used for the fine-tuning of a large pre-trained language model using an unlabeled corpus with minimum annotation cost. We introduce a new criterion to select the most informative samples from a pool of unlabeled data points. For that, we suggest exploiting unsupervised methods, such as clustering and dominant sets [109] for

the training of deep language models using active learning. We propose finding the dominant sets [109] of local clusters in the feature space of a deep pre-trained language model. As these sets represent maximally cohesive structures in the data, they provide a notion of a cluster [14]. Accordingly, this method is also referred to as dominant set clustering. Using this technique, this doctoral work suggests that the samples that are not strongly coherent with the clusters, which are the ones that do not belong to any of the dominant sets, can be selected to be used to train the model, as these points might provide more information to the model. These data samples are selected in our method as they represent the boundaries of the local clusters and are more challenging to be classified. This approach makes data selection to be as diverse as possible via enforcing an equal number of samples to be selected from the non-dominant sets associated with each cluster. When the selection of equal number of samples associated with each cluster is not possible, we consider an adaptive cutoff value for the non-dominant sets. This method is extremely useful in the case of substantial imbalance in the dataset under study, and thus, the potential of severe imbalance in the size of identified clusters. We show that our method finds the edge cases in the feature space of a pre-trained BERT model, which are also the data samples that are the most challenging to the model. We additionally propose a hybrid strategy that allows us to incorporate the uncertainty score in the later stages of selection when the uncertainty score is more reliable.

Dominant set clustering is non-parametric and its a sequential method that only uses a predefined similarity matrix to find the cohesive structures in that space. Meanwhile, the number of classes in a classification task can be used as a known parameter for an initial clustering before finding the dominant sets in each cluster. We apply this method to allow parallelization of applying dominant set identification for clusters, which makes our method more practical for large datasets. Additionally, the preparation of pairwise similarity matrix for dominant set identification becomes substantially faster in this case as the calculation of pairwise similarity takes $O(n^2)$ in the size of the input. Thus, we divide the space by applying an initial clustering. As this approach does not

have any parameters to be tuned, it is dataset-independent. By conducting extensive experiments and analysis, we show that our proposed method can approximately achieve the same classification accuracy as using full training data, with significantly fewer data points. Additionally, this method achieves a higher performance in comparison to the state-of-the-art active learning strategies, while also being robust to outliers in the data. Furthermore, our algorithm is able to incorporate conventional active learning scores, such as uncertainty-based scores, into its selection criteria, referred to as our hybrid strategy.

We show the effectiveness of our method on different datasets and using different neural network architectures. We specifically show that our method outperforms the state-of-the-art uncertainty-based methods in the early stages of selection when the uncertainty score extracted from the model is not accurate. Via analyzing the results, we argue that an active learning task can be divided into two distinct phases. In the early stages, unsupervised techniques such as employing pre-trained models, clustering, and identifying the dominant sets outperform the supervised methods, e.g., uncertainty score extracted from the trained model. However, in the second phase, later stages of selection, taking the model uncertainty into the account can improve the selection performance. Such hybrid methods are not currently well-studied in the literature. Therefore, via the research in this work, we are hopeful to stimulate further research on similar active learning strategies.

Our proposed active learning approach has the potential to mitigate the difficulties associated with the annotation and classification of textual content, e.g. annotation cost and bias. Labeling offensive and abusive content is particularly difficult, as it can cause discomfort and emotional disturbance for the human annotators. Via extensive experiments and analysis, we show that our method is particularly practical in the classification of toxic language in online social media using a small amount of labeled data. However, our method can be used in other fields as well to facilitate and accelerate the development of AI research.

1.5 Curriculum Learning and its Application to Active Learning

In the area of machine learning, curriculum learning was first proposed by [7] and found that the exclusion of difficult samples and noisy data in the early training stage is beneficial, such as faster convergence and achieving better local minima. Due to the difference in the difficulty levels of the examples from any dataset, extensive research has investigated methods to identify the easy samples from the difficult ones and to arrange them as a curriculum for the training of a model [7]. Curriculum learning is in contrast with introducing the data to the model without any order. Some literature in this area of research suggest that the learning process can achieve remarkable improvements when using a curriculum [150]. As a result of gradually increasing the difficulty of the training samples over the training epochs, the model can take a better learning trajectory and avoid local minima. However, relatively little attention is devoted to this topic in the area of natural language processing. Examples of difficulty in language are sophisticated reasoning such as negation, lengthy text, and different types of rare words, including misspellings, abbreviations, and scientific and literary words.

Language model pre-training using large-scale unlabeled data and fine-tuning for downstream tasks have recently drawn a lot of attention in the field of natural language processing and understanding [157, 118, 119] and have led to state-of-the-art results across different tasks. This is the result of learning universal language representations using large-scale unlabeled data. However, the application of a curriculum in the finetuning stage of big language models [78, 153, 100] has not been investigated for different tasks and conditions [150]. The work of [153] is the first study that investigated the scale of impact of using curriculum learning in the NLP field. Specifically, the experiments using different datasets and testing on varying tasks in this paper show that finetuning transformer-based architectures can benefit from applying such a technique. Yet, the challenge in this area of research is defining the difficulty and noise to exclude such data in the early stages of

training, such that this difficulty evaluation remains consistent across tasks and datasets.

Curriculum learning aims to consider the level of difficulty of data samples and train the model with the easier data points first to improve the efficiency of training [51]. In recent studies, curriculum learning has been shown to be a practical technique to fine-tune deep pre-trained language models for a wide range of tasks related to natural language understanding [153] and information retrieval [110]. Although some studies have considered designing a curriculum-learning algorithm based on predefined difficulty measures, such as the length of input paragraphs for question answering [115], such methods might not be generalizable to a different task, as they can be very dataset-dependent and problem-dependent. Additionally, a defined difficulty level by human judgment might not reflect the same level of difficulty and meaning for a deep language model [153].

Active Learning (AL) [27] and Curriculum Learning (CL) [7, 51] are two closely related disciplines in machine learning. In fact, active learning is anti-curricula. Despite the existence of individual studies on the investigation of each of these techniques for a specific task or the analysis of a dataset, little information is known on the impact of each at different stages of training or fine-tuning a deep model [150] and specially, a deep language model. Additionally, a handful of studies have investigated the incorporation of the learning from a curricula into data selection for active learning [66]. While active learning goal is to minimize the number of required labeled data via smart data selection according to an informativeness score, curriculum learning can sort the training data based on some difficulty criterion.

As we are interested in the application of curriculum learning for the classification of social media textual data, we must pay attention to the problems related to this field, including short text, data imbalance, and lack of annotated data. In chapter 4, we address these issues by proposing an active learning method that outperforms state-of-the-art active learning techniques for the task of social media hate speech classification. In chapter 5, we further investigate the impact of data imbalance

on the difficulty of samples, but this time for the problem of curriculum learning.

1.6 Statement of Contributions

This doctoral work in general contributes to the fields of machine learning, natural language processing, online social media analysis, statistical modeling, and deep learning. Specifically, the expected contributions of this work are:

I) presenting a novel unsupervised learning and generative probabilistic modeling method by presenting an approach that is practical in the modeling and summarization of significant topical events in online social media;

II) introducing an innovative entropic measure for the comparison of topics based on the identified topic distributions over time and quantifying the significance of the narrative activities and recurrence of a topic via employing the Shannon entropy;

III) introducing a new data selection and active learning technique for multi-class classification, and specially, for text classification, by presenting a non-parametric and adaptive strategy using an informativeness criterion for data selection based on the dense structures in the embedding space of a large language model, e.g. the BERT model;

IV) providing evidence on the benefits and restrictions of curriculum learning strategies, which aim to improve performance and speedup convergence. This method aims to further mitigate the difficulties associated with the classification of textual content.

1.7 Statement of Originality

Parts of this doctoral thesis are included and published in conference proceedings, conference presentations, and scientific journals, and the rest are under review for conference and journal publications. At the time of writing, the rest of research material in this doctoral thesis is original and is not publicly published elsewhere:

I) Oghaz, T. A., Mutlu, E. Ç., Jasser, J., Yousefi, N., & Garibay, I. (2020, July). Probabilistic model of narratives over topical trends in social media: A discrete time model. In Proceedings of the 31st ACM Conference on Hypertext and Social Media (pp. 281-290) [105].

II) Mutlu, E. Ç., Oghaz, T. A., Tütüncüler, E., & Garibay, I. (2020, December). Do bots have moral judgement? The difference between bots and humans in moral rhetoric. In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 222-226). IEEE [99].

III) Mutlu, E. C., Oghaz, T. A., Rajabi, A., & Garibay, I. (2020). Review on learning and extracting graph features for link prediction. Machine Learning and Knowledge Extraction, 2(4), 672-704 [98].

IV) Mutlu, E. C., Oghaz, T. A., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., ... & Garibay, I. (2020). A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. Data in brief, 33, 106401.

V) Oghaz, T. A. & Garibay, I. Dominant Set-based Active Learning for Text Classification and its Application to Online Social Media. Journal of Online Social Networks and Media (Under Review) [104].

CHAPTER 2: BACKGROUND AND RELATED WORK

In this chapter, we provide an overview of the areas related to this doctoral thesis. We review the basic concepts and terminologies that are used in the next chapters.[1]

2.1 Analyzing Online Social Media Textual Content

Social media and microblogging platforms, such as Twitter and Facebook, allow their users to freely and publicly express their opinions. Connecting billions of people across the world from different ethnicity, race, religion, and nationality has resulted in the propagation of massive flow of information across these environments. As a result, they are becoming the primary sources of real-time content regarding the ongoing socio-political events, including the United States Presidential Election in 2016 [35], and natural and man-made emergencies, such as the COVID-19 pandemic [26].

Due to this massive amount of information, it is extremely challenging to obtain relevant information on significant events [8, 44], distinguish between high-quality content and disinformation [56, 99, 120], understand and follow activities around different opinions within a polarized domain [43], or in general, analyze and model graphs of online social media to predict future relations [98]. Besides the threat of deliberate misinformation on online platforms, the spread of offensive language and hate speech is shown to plague society and cause violence. An example of such an incident is the spread of anti-Muslim information on Facebook that led to the 2019 massive violence in Sri Lanka. As a result, there is an unprecedented necessity of moderating these platforms to protect the users from disinformation, toxicity, cyberbullying, and trolling [87].

[1]Portions of this chapter is reprinted, with permission, from ©2020 ACM [105] `https://doi.org/10.1145/3372923.3404790`.

In response to the urgency of developing practical techniques to monitor online platforms, recently a substantial amount of research publications are focused on the analysis of online social media textual content. Examples of such efforts are finding the indicator attributes [87, 29], preparing manually labeled datasets for toxicity type classification and detection [38, 30, 152], preparing datasets for fact-checking and rumor detection [161], developing machine learning models and systems for offensive language detection [123, 95] and misinformation [131], topic modeling [80] and event detection [82], automatic summarization techniques that help to grasp the main ideas from online social media data [138], investigating whether the offensive language on social media is targeted at an individual or group [6, 160], as well as developing methods and techniques that improve the performance and efficacy of the mentioned models on different tasks by considering a priority instead of uniform training data selection, such as in Active Learning (AL) [37] and Curriculum Learning (CL) [20].

The problem of hate speech, toxic, and offensive language in general, is admittedly one of the major issues in online social media. Therefore, detecting and monitoring the propagation of such content has been a priority for social media companies. Studying the diffusion dynamics of content in online social media has revealed that hateful content spread much faster and farther than non-hateful content, reaching to significantly larger audience [87]. This has motivated many social media companies to moderate such content to prevent its potentially disastrous consequences [38].

Literature on offensive language and hate speech have shown that automated detection of hate speech is not a trivial task, as lexical detection methods cannot easily distinguish between hate speech and other instances of offensive language [29]. This signifies the need for high-quality datasets containing different instances of offensive language, as well as developing better methods of annotation and detection of cyberhate [160, 16, 148]. For example, employing syntactic features has been shown to be useful to improve the identification of the targets and intensity of hate speech [16, 48, 134, 147]. In this thesis, our focus is on the development of low-resource machine

learning methods to mitigate the difficulties around data annotation and improve the efficacy of the convergence of deep language models. Our proposed methods and results are presented in the next chapters. We show that when a supervised setting is desired, the number of annotated samples can be reduced significantly without reducing the performance.

2.2 Narrative Analysis

Narratives can be found in all day-to-day activities. The fields of research on narrative analysis include narrative representation, coherence and structure of narratives, and the strategies, aim, and functionality of storytelling [94]. From a computational perspective, narratives may relate to topic mining, text summarization, machine translation [141], and graph visualization. The latter can be achieved via using directed acyclic graphs (DAGs) to demonstrate relationships over the network of entities [49]. Narrative summaries can be constructed from an ordered chain of individual events with causality relationships amongst events, that appeared within a specific topic [68]. The narrative sequence may report fluctuations over time relative to the underlying events. Additionally, the story-like interpretation of the text is a must to imply a narrative [106].

Since social media have been admitted as a component of today's society, many studies have investigated narratives in social media content [45, 106, 144]. These narratives contain small autobiographies that have been developed in personal profiles and cover trivial everyday life events. Other types of narratives appearing in social media platforms consist of breaking news and long stories of past events [106]. Some types of narratives, such as breaking news, result in the emergence of other narratives related to the predictions or projections of events in the near future [45]. This literature views social media conversation cascades as stories that are co-constructed by the tellers and their audience and are circulating amongst the public within and across social media platforms. Moreover, the events have been considered as the causes of online user activity that can

be identified via activity fluctuations over time [4, 106]. Developing appropriate tools for social media narrative analysis can facilitate communicating the main ideas regarding the events in large data.

2.3 Statistical Topic Modeling

From the field of machine learning and data mining, topic modeling refers to the utilization of hierarchical probabilistic models that can help with the discovery and annotation of documents with thematic information [8]; e.g., to discover word patterns that reflect the underlying topics in a set of document collections [3]. In contrast to earlier efforts in the field of information retrieval to find short descriptions of document collections, e.g., the term frequency-inverse document frequency (tf-idf) and latent semantic indexing (LSI) [31], topic models reveal the statistical structure within and across document collections and lead to significant data compression [22, 8]. A detailed review of topic modeling methods and varying considered features are provided in [143, 3].

The most commonly used approach to topic modeling is Latent Dirichlet Allocation (LDA) [10, 54]. LDA is a generative probabilistic model with a hierarchical Bayesian network structure that can be used for a variety of applications with discrete data, including text corpora [69]. Using LDA for topic mining, a document is a bag of words that has a mixture of latent topics [10]. The generative probabilistic procedure of LDA, in which a multinomial variable z for each topic is selected for each word w in a given document d, can be described as follows:

I. For each topic z, draw T multinomials ϕ_z from a Dirichlet prior β;

II. For each document d, draw a multinomial θ_d from a Dirichlet prior α;

III. For each word w_{di} in d:

Variable Descriptions	Symbol
Number of topics	T
Number of documents	D
Number of word tokens in document d	N_d
Multinomial distribution of topics for document d	θ_d
Multinomial distribution of words for topic z	ϕ_z
Topic of the ith token in document d	z_{di}
ith word token in document d	w_{di}

Table 2.1: LDA symbols and definitions

(a) draw a topic z_{di} from multinomial θ_d;

(b) draw a word w_{di} from multinomial $\phi_{z_{di}}$;

The list of symbols and their descriptions can be found in table 2.1. The model parameterization is as below:

$$\begin{aligned} \theta_d|\alpha &\sim \text{Discrete}(\alpha) \\ \phi_z|\beta &\sim \text{Discrete}(\beta) \\ z_{di}|\theta_d &\sim \text{Multinomial}(\theta_d) \\ w_{di}|\phi_{z_{di}} &\sim \text{Multinomial}(\phi_{z_{di}}) \end{aligned} \tag{2.1}$$

In LDA-based methods, the calculation of the posterior probability $p(\phi, \theta, z|w)$ is intractable to compute. Among the proposed solutions to approximate the posterior distribution and maximize the log-likelihood of the data are variational inference [10] and fast collapsed Gibbs sampling [116]. The Gibbs sampling method is a selection-based approximation technique using a Markov Chain to estimate the posterior probability. Instead, variational methods change the inference problem to optimization [8].

Many advanced topic modeling approaches have been derived from LDA, including Hierarchical Topic Models [53] that learn and organize the topics into a hierarchy to address a super-sub topic relationship. Considering a tree, an L-dimensional Dirichlet is used to draw a vector of topic proportions for a root-to-leaf path with length L. Then, the words are generated along this path from a mixture of topic proportions. Thus, all topics in the same path belong to the same root topic. A hierarchical approach that can identify related events is well-suited for analyzing social media and news stories that contain rich data over a series of real-world events [136].

Modularity is a big advantage of generative methods as it can allow combining different models, each capturing varying probability distributions from the underlying data [52]. As a result, many research papers examined this via combining a mixture or product of models. For instance, the idea of word co-occurrence in LDA is extended in [36] for article categorization as a model of words in scientific paper abstracts along with their bibliography sections. Another example is the generative composite method proposed in [52], which combines a topic model with a Hidden Markov Model (HMM) to capture long-range semantic dependencies as well as short-range syntactic dependencies. In [88], LDA and author topic models are combined to allow topic as well as role discovery in social networks via modeling social network authors and their topic of interest simultaneously.

Many research articles developed ways to relax the time exchangeability assumption of LDA, which presumes all documents in the corpus are timely interchangeable and their order is not important. The assumption about the time of documents in the collection is specially not accurate when the dataset is collected over a long period of time [143]. In this regard, topic models over time with continuous-time distribution [145] and dynamic topic models [9] are developed in order to capture the rise and falls of topics within a time range. Dynamic topic models divide the data into slices based on the time of the documents, e.g., annually, and assume the topics in each time slice to be evolved from the previous slice. In this method, an evolving state model is used to chain the parameter of topics, and thus, allows modeling sequences of random variables by hav-

ing sequences of linked topic models. The continuous-time distribution method and its variations, such as topics over time [145], are designed by combining Bayesian networks with homogeneous Markov processes. Topics over time allows topic-time assignment given a word in a document, while the time of all words within a document are equal. Designing this method with a Beta distribution for timestamps of the documents leads to discovery of isolated temporal patterns, as in high peaks or U shapes. No other time patterns can be discovered with this method. This is because topical models with continuous-time distribution cannot model many modes in time, which leads to deficiency in modeling the fluctuations. Additionally, continuous-time models suffer from instability problems in the case of analyzing a multimodal dataset that is sparse in time. The assumptions in the design of such methods seem to be valid for some datasets, such as for scientific and news articles, which is what have been used frequently for the evaluation of these models. However, as studying datasets related to time-series activities on online social media platforms necessitates resolving multi-modality and sparsity, continuous-time topic models discussed above cannot accurately model the rises and falls, as well as the appearance of new topics in the distribution of topics for these platforms.

As social media textual data is usually comprised of short pieces of text, topic models that focus on word co-occurrence in short documents may fail to accurately identify the topics and the words that belong to them. As a result, many methods have been proposed specifically for short text topic modeling, such as considering word co-occurrences instead of modeling documents [156], and creating "pseudo-documents" [90]. The model proposed in [156] addresses the problem of sparsity of word co-occurrence in short text by considering a biterm design, in which any two distinct terms in a short piece of text create a biterm. In the case of social media data document aggregation to prepare pseudo-documents, hashtag co-occurrence for Twitter data and author-based aggregation seem to be practical methods [137]. However, the time information of the data will be lost in such aggregations. In this doctoral work, we discuss a novel probabilistic method for the analysis of

topics in online social media considering the timestamps of the textual content. Via analysis and explanation, we show how our design settles the constraints and limitations of the prior works. The details on the proposed techniques and analysis are available in the next chapters of this doctoral thesis.

2.4 Deep Pre-trained Language Models for Low-Resouce Settings

Deep Neural networks have revolutionized the field of machine learning, including natural language processing and language understanding. The most prominent recent development in this field is the introduction and application of Transformer-based architectures with attention mechanism [142], which identifies the global as well as local dependencies between an input sequence and the output. This architecture is based entirely on the attention mechanism. The attention mechanism is an interpretable technique introduced to sequence modeling to capture crucial dependency information required to perform a task regardless of the distance in the sequences of dependent elements. Before this achievement, the dominant methods gaining the highest performance and state-of-the-art results on varying tasks related to language modeling and understanding were Recurrent Neural Networks (RNNs) [127, 126], Long Short-Term Memory (LSTMs) [133, 61, 83, 19, 50], and Gated Recurrent Unit RNNs [21, 25, 18, 61]. Although transformer-based models with pertaining have revolutionized the field of natural language modeling and understanding in general, we discuss why these techniques are specifically significant to be applied in low-resource settings.

In resource-poor scenarios, the requirement of substantial labeled data has dampened the dominance of deep sequential models, such as LSTMs, RNNs, and GRUs, over traditional statistical methods. Having a large amount of annotated data may not be possible in many cases, such as for many low-resource languages. Even for high-resource languages such as English, the acquisition and preparation of high-quality labels for a sufficiently large amount of data can be extremely

challenging for some tasks and domains. Furthermore, resources for storing and processing a large amount of textual data and training a deep language model are costly. Additionally, contextual information to learn the dependencies directly from the input sequence might be missing or short. In the case of analyzing online social media data, despite the existence of a growing volume of user-generated content and the necessity to moderate these platforms to protect users from disturbing information, the annotation procedure is even more burdensome. For example, data that contain content with toxicity, cyberbullying, and trolling can be mentally and emotionally taxing for human annotators. Besides the annotation cost, such datasets are usually extremely imbalanced and the labeling efforts may lead to mistakes, such as enduring bias toward some attributes and races [95, 24]. These domains are sometimes also referred to as non-standard domains or tasks. Due to the importance of this area of research, a substantial amount of work has been conducted to find the indicator attributes [87, 29], preparing manually labeled datasets on sensitive content such as toxicity type classification and detection [38, 30, 152], and the development of machine learning models and systems for offensive language detection [123, 95]. Thus, NLP methods that can make the best use of significantly less labeled data points and models with little training computational cost are of great interest. Many techniques have been developed to mitigate the challenges around large annotated datasets, such as language pre-training, self and paired training, semi-supervised methods, few-shot learning, transfer learning, and active learning.

Using pre-training techniques has been common in text analysis and mining for many years. Among the proposed approaches, continuous bag of words, skip grams [92], Word2Vec [93], Global Vectors for Word Representations (GloVe) [111], FastText [72], Embeddings from Language Models (ELMO) [113], GPT-3 (Generative Pre-trained Transformer) [118, 13], and BERT (Bidirectional Encoder Representations from Transformers) [32] are the most popular. The output of such techniques is a high-dimensional vector representation with local or global contextual information by learning such relations in a unidirectional or bidirectional manner, including complex

relationships such as syntax and semantics. Language pre-training enables transfer learning as the general knowledge learned from a massive amount of unlabeled textual data can be used for a new task.

After the appearance of deep language models and pre-training and finetuning techniques, many research papers focused on employing such methods using different datasets on varying NLP tasks and releasing the pretrained models for public access. The pre-training stage uses unlabeled data to initialize the parameters of the deep base model such that in the finetuning stage for a downstream task, a few parameters need to be slightly updated. In [114] and [118] traditional unidirectional language models are used for the pre-training stage, for instance, left to right or right to left models that result in the attention of each token to the token either on its left or right. The well-known OpenAI's GPT [118] and GPT-3 [13] models have unidirectional auto-regressive architectures in which the context of each token is the result of attention to the previous token and its context. However, the context-sensitive features, in this case, might not be fully captured. Furthermore, unidirectional language modeling limits the process of pre-training as well as introduces constraints for fine-tuning. The ELMO model [113] extends unidirectional language modeling to bidirectional via concatenation of the representations. However, it does not provide a representation as in deep bidirectional models and it is rather known as a feature-based method.

Instead of using traditional unidirectional language models and concatenating the representations, BERT [32] is pre-trained jointly on two tasks in an unsupervised manner: 1. Masked Language Modeling (MLM) [139], and 2. Next Sentence Prediction (NSP). This bidirectional pre-training method relaxes the restrictions of unidirectional language models. Bidirectional pre-training is specially reported to be crucial in tasks such as question answering, where the contextual information from both directions is of great concern [32]. In MLM, via randomly masking a percentage of the token sequences the model learns to predict the masked tokens by exploiting contextual information in both left and right directions for all layers. BERT is pre-trained with 15% token masking.

The idea behind NSP is to capture the relationships between 2 sequences by training the model to predict the next sentence. Employing this technique specially benefits tasks such as question answering and natural language inference. Finetuning the transformer-based language models such as BERT can be end-to-end, training the task-specific layers only, or training selective parameters such as cross-attention parameters [46]. Finetuning is not expensive as training, however, each of these techniques might have benefits for specific tasks and data domains. Specially, when the dataset for a downstream task has a small size, finetuning can lead to overfitting and catastrophic forgetting [32, 46]. Thus, it is suggested to perform finetuning for only a small number of epochs (2-4 epochs in most literature) to avoid such effects. BERT uses the exact transformer architecture with multi-head self-attention introduced in [142].

In sequential models such as LSTM and RNN, sequential training procedure for long documents and big data is challenging as for sequential computation, such as in RNNs, the hidden state of each part of the input sequence (token) needs to be calculated based on the hidden state of the previous token. The transformer-based architecture [142] overcame this challenge by enabling parallel computation of all hidden representations of tokens. This is inspired by the high performance of using attention mechanism [76, 108] such that stacking many self-attention and fully connected layers, leads to computing a representation of the input sequence according to the position of each part of the input to the rest. As a result of overcoming the problem of sequential computation, the transformer architecture is widely used for pre-training by applying the techniques discussed above and using a massive amount of unlabeled data. It has been reported that this technique is effective across different tasks related to natural language processing when little labeled data is used for fine-tuning the deep pre-trained language models. As the base layers are initialized via pre-training, the small annotated data is observed to be sufficient to train a task-specific layer on top of the BERT model [32].

Figure 2.1 depicts the transformer model architecture using BERT model as the base for the task of

text classification. BERT uses [CLS] and [SEP] tokens as special symbols for the model showing the beginning of the sequence and the separators, for example separating questions from answers. The final hidden state of the [CLS] token can be used for the task of classification as it aggregates the representations for all tokens in the sequence. However, it is also common to use all final hidden states of tokens with a pooler layer and feed that to dense layers for classification instead of only the final hidden state of the [CLS] token as is depicted in Figure 2.1.

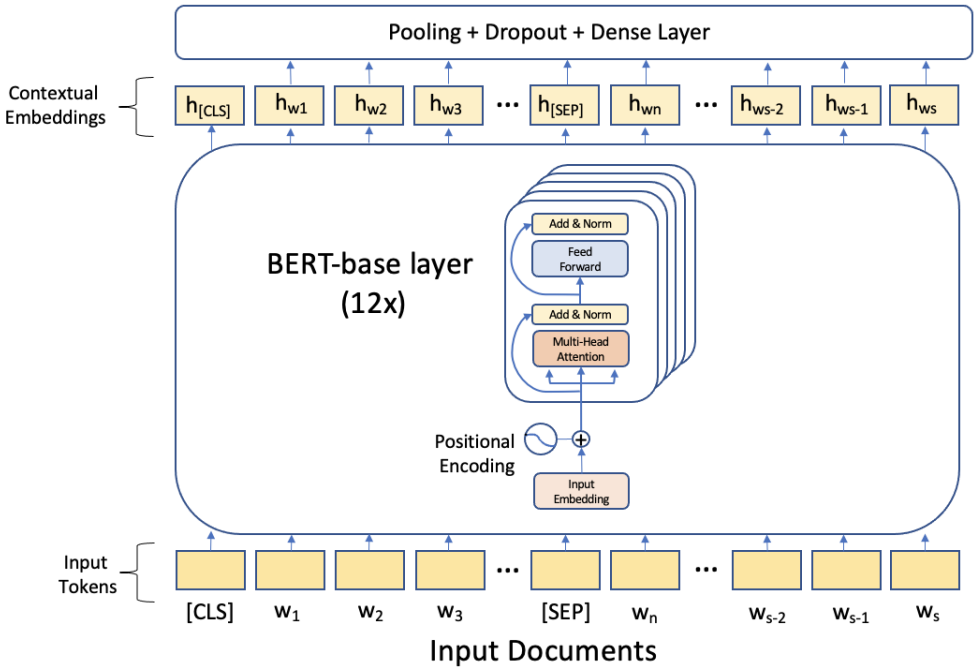


Figure 2.1: The overall architecture of a text classifier with a BERT-base model (number of layers L=12, hidden size H=768, number of self-attention layers A=12, number of parameters=110M) to extract the contextual representations of the tokens for the task of text classification. The [CLS] token shows the beginning of input sequences. This figure shows when all final hidden states of tokens are used for classification instead of only the final representation of the [CLS] token.

Data parallelization and improving computational efficacy of the BERT model led to analyzing long sequences of text and to recognize global dependencies between the input and the output by reducing this required process to a constant number of operations [142]. As discussed earlier, transformers are encoder-decoder architectures that are built entirely from stacked self-attention

layers and fully connected layers. The self-attention mechanism results in obtaining a representation of a sequence by making relations between an input and an output sequence. The process by the attention mechanism can be simply defined as mapping a query (Q), and a key-value $(K-V)$ pair vector to an output vector that itself can be calculated via a weighted sum over the values as:

$$Attention(Q,K,V) = softmax(\frac{QK^{\intercal}}{\sqrt{d_k}})V, \tag{2.2}$$

where d_k is the dimension of the K vector. A multi-head attention mechanism is then achievable by a linear projection of h concatenated vectors, which are the output of h attention heads. Each of these attention heads apply a scaled dot-product attention function and the process is performed in parallel. The attention functions receive h different learned projections of queries, keys, and values as input. As a result of learning from different representations, the multi-head self-attention enables attending to varying information at different positions.

As a result of knowledge sharing available from the explained pre-training techniques introduced in BERT, a small amount of labeled data for fine-tuning the BERT model for a downstream task can be sufficient to achieve a high-performance [33]. This can be accomplished by adding only a single output layer to achieve state-of-the-art results across varying tasks. The discussed features of pre-trained language models, specially BERT, make them ideal to be used with low-recourse label-efficient techniques, such as with active learning.

In this doctoral thesis, we aim to show how pre-trained language models can be used to minimize the annotation cost, while achieving high performance for the task of text classification of social media data. We investigate this by studying pre-trained language models with active learning and curriculum learning. The details on the implementation and analysis of our approaches are provided in the next chapters.

2.4.1 Active Learning

Active Learning (AL) refers to the process of efficient selection of the most informative data when the data is plentiful, but the labels are scarce [129]. Active learning techniques can mitigate the issues associated with manual labeling and improve automatic detection and classification when labeled training data is sparse [33]. The goal of active learning is to

i) reduce the cost of labeling via using a small number of labeled samples for training, and

ii) query the class labels for the most informative subset of the data samples that are selected using an acquisition function.

The best active learning strategy successfully selects certain unlabeled data samples from the distribution of available data, such that using this data portion for training leads to the maximal reduction of the classification error and variance. Thus, a wealth of work has been made on the development of various active learning approaches for different machine learning applications [42, 63], and particularly for text classification [89, 165, 166] and specially, for toxicity detection [12].

Main approaches of active learning can be categorized into methods based on uncertainty scores such as minimum margin (least confidence) [81, 125] and Bayesian active learning (Monte Carlo dropout) [41, 42], the entropy of class predictions [62], prediction disagreement in ensemble-based settings such as Active-Decorate [91], expected gradient length [70], expected loss value [84], and representation-based methods such as Core-Set [128] which aims to select samples that cover the learned representation space. The work of [129] provides a thorough summary of active learning techniques before the advancement of deep neural networks. The development of deep neural networks resulted in deep active learning techniques that are trained in batches instead of single data queries in classic AL [41]. A survey of advancements in deep AL is presented in [122].

Uncertainty-based AL techniques use the classification probabilities extracted from the model as the informativeness criterion with the hope that selecting samples in this way leads to a lower model uncertainty. The main intuition behind uncertainty-based active learning methods is that if the model is uncertain about a sample, it likely lies near the decision boundary of the classes. Thus, knowing its label can help the classifier to better estimate the decision boundary. However, when the classifier is trained on only a few samples, or even no samples at all, the uncertainty score obtained from it is not a reliable metric. In other words, the classifier does not yet know what it does not know. Also, the diversity-based techniques, which select samples that cover the feature space or select from the dense regions of the space can lead to suboptimal performance, as many of the selected data points can be redundant and less informative. On the contrary, literature report the highest performance scores for the uncertainty-based AL methods, such as Bayesian AL [41], for a relatively large query size or in later cycles [47].

Because of the high chance of bias in sampling with such methods, many AL techniques have been developed that query samples that best represent the unlabeled data, usually via finding the cluster structure of the data [28]. These methods rely extensively on the quality of the initial unsupervised method. Despite the existence of many clustering techniques, graph-theoretic clustering methods, such as spectral clustering, dominant sets [109], and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [74] achieve higher performance than k-means in the discovery of the clusters of arbitrary shapes, and are more robust to noise and outliers. Among these methods, dominant sets clustering [109] has the lowest computational complexity.

In this doctoral thesis, we exploit unsupervised methods such as clustering and dominant set clustering[109] for the training of deep language models using active learning. The work of [64] also uses unsupervised techniques, such as dominant set and spectral clustering for active learning. However, it is not designed to be applied to the training of a deep model, where the data needs to be processed in batches. The authors find that spectral clustering and dominant sets complement each

other and using them hierarchically and exclusionary can boost active learning. However, their method uses heuristics to filter the data in multiple steps and build high-quality clusters, which necessitates a substantial amount of labeled data to build "pure" initial clusters. These clusters are later used for the clustering of unlabeled data. Finally, the nature of their algorithm is sequential, which is not practical for deep AL methods due to slow convergence and local optimization algorithms. One-by-one data selection can be viewed as having the batch size of size one, which delays convergence and is expensive.

Querying based on a hybrid strategy, e.g., using density information as well as the classification margins is investigated in many research, including [103]. However, an ad hoc combination of such measures can slow the convergence, and lead to suboptimal performance as a result of biased data selection. Diversity-based sampling [65] was proposed to systematically combine informativeness and representativeness measures such that the selected samples have the highest uncertainty scores for both sets of labeled and unlabeled data. This method was designed as a binary classification task and the selection mechanism is sequential instead of batch mode AL. Core-set strategy [128] is a diversity-based technique that aims selecting samples that cover the feature space using a greedy algorithm. Despite high diversity in selected samples, [33] reports low representatives scores and slow convergence for this algorithm comparing to other techniques.

2.4.1.1 Active Learning for Natural Language Processing

Although a high volume of research on active learning exists, this topic has received little attention in the field of natural language processing. The works of [132] and [33] present empirical analysis of employing deep active learning strategies for natural language processing tasks. In [132], the best performance across different tasks was achieved for Bayesian active learning based on disagreement using the uncertainty scores from Dropout [41] or Bayes-by-Backprob [11] methods.

An active learning approach for multi-label text emotion classification task is proposed in [73] that uses a probabilistic distance between the expected label distribution and uniform distribution. The main goal is to collect balanced data, when the pool of data is imbalanced. The task of named entity recognition from clinical text using active learning has been studied in [149], which models the informativeness as well as the annotation cost. This approach was specifically designed for scenarios where the labeling cost, e.g. time, for different samples is different. Therefore, an estimation of the labeling cost was also taken into account for selection.

Zhang et al. [165] employed the embedding space of neural networks for word and sentence classification. The words or sentences that would potentially change the embedding function the most are selected to be labeled. This is estimated by calculating the expected gradient length. However, recent advancements in pre-training the embedding function with extremely large unlabeled datasets have eliminated the need for such consideration. For example, the application of active learning for binary text classification with the BERT pre-trained model has been investigated in [33], which provides prominent evidence that such pre-trained models are powerful tools for text classification when combined with active learning strategies. Even for tasks other than text classification, the employment of pre-trained deep language models with a practical active learning strategy seems to provide a promising direction via reducing the burden of manual labeling and resulting in a strong performance across many diverse tasks.

In this thesis, we aim to introduce a scalable technique based on unsupervised learning to achieve a general method for data selection and training of a model. The proposed methods and evaluation results are discussed in detail in the next chapters.

2.4.2 Curriculum Learning for Natural Language Processing

Similar to active learning, curriculum learning tries to consider a priority into using data samples for training instead of uniform data selection [7]. However, dissimilar to active learning it usually all the dataset instead of partial data, and in the order of increasing difficulty instead of selecting the most challenging samples first [100]. The idea behind curriculum learning is inspired by the learning process of humans, which is based on ordering the information to an easy to difficult arrangement. Literature on this topic also refer to curriculum learning as gradual learning. For non-convex functions such as deep learning methods, curriculum learning is reported to be a powerful strategy for global optimization of the function [150].

The focus of literature on this topic is on understanding and defining difficulty across different tasks. Applying curriculum learning with deep learning has provided evidence of faster training convergence and obtaining a high-quality local optima [153, 150]. In the field of natural language processing, examples of the most commonly used difficulty criterion are the frequency of rare words [7] and length of input sequences (block size) [100]. Although this area of research is not new, the number of research papers that have investigated the application of curriculum learning in machine learning are very limited [150].

Well-known NLP models, such as GPT-3 [13] and T5 [119] are trained according to a curriculum. However, this does not mean that such a technique can enhance the performance of deep language models for any task and under any condition. As a result, further investigation of the conditions and specific notions of difficulty for tasks related to natural language processing is necessary. The work of [153] is the first study that investigated the scale of impact of using curriculum learning in the NLP field. Specifically, their experiments using different datasets and evaluation on varying tasks show that finetuning transformer-based architectures can benefit from applying such a technique. Yet, the challenge in this area of research is defining the difficulty and noise to exclude such data

in the early stages of training. Also, the application of a curriculum in the finetuning stage of big pre-trained language models [78, 153, 100] has not been investigated for different tasks and conditions [150], e.g., imbalanced short text data.

The available literature that is focused on the training and fine-tuning of deep pre-trained language models based on curriculum learning is a handful. In recent studies, curriculum learning has been shown to be a practical technique to fine-tune deep pre-trained language models for tasks related to natural language understanding [153] and information retrieval [110]. Some studies have considered designing a curriculum-learning algorithm based on predefined difficulty measures, such as the length of input paragraphs for question answering [115]. However, such methods might not be generalizable to a different task, as they can be very dataset-dependent and problem-dependent. Additionally, a defined difficulty level by human judgment might not reflect the same level of difficulty and meaning for a deep language model [153].

To identify the most difficult data sample, [7] associates the loss value of pre-trained models with data difficulty. The work of [71] defines a difficulty score based on investigating the consistency of a model in predicting the class label of an example in i.i.d. draws from the data. The work of [140] relates difficult samples to the phenomenon of catastrophic forgetting and suggests the easy samples to be identified as those that do not get forgotten by the model over the training procedure. The impact of data imbalance on sample difficulty is investigated in [146] and this work suggests a dynamic technique for the adjustment of strategy and weight of loss per batch.

As in this doctoral thesis, we are interested in the analysis of social media data for tasks such as text classification, we must pay attention to the challenges related to such data, such as data imbalance, lack of context in short text, and annotated data limitation. The problem of data imbalance for curriculum learning can be tackled by oversampling from the minority class, downsampling from the majority class, and adjusting weights for the loss function [34, 55, 86]. It must be noted

that in this context, downsampling of the training data refers to selecting a small subset of the majority class examples. Techniques such as oversampling and undersampling are not practical when lacking prior knowledge on the classes. Additionally, oversampling can lead to overfitting of the model as repetitive information is being fed to the model. Also, downsampling can lead to losing a substantial amount of information that is required for the task under study. Dealing with data imbalance in curriculum learning has been previously investigated in [146] and a dynamic solution with batch-level strategy and loss weight adjustment was suggested.

CHAPTER 3: PROBABILISTIC MODEL OF NARRATIVE OVER TOPICAL TRENDS IN SOCIAL MEDIA: A DISCRETE-TIME MODEL

In this chapter we describe our framework for unsupervised narrative summary extraction for online social media timestamped content.[1] Recent advances in natural language processing (NLP) in online social media are evidently owed to large-scale datasets. However, labeling, storing, and processing a large number of textual data points, e.g., tweets, has remained challenging. On top of that, in applications such as hate speech detection, labeling a sufficiently large dataset containing offensive content can be mentally and emotionally taxing for human annotators. Thus, NLP methods that can make the best use of significantly less labeled data points and unsupervised practices are of great interest as they require little to no labeled training data. As social media activities generate abundant timestamped multimodal data, many studies such as [23] have presented algorithms to discover the topics and develop descriptive summaries over social media events. From the field of machine learning and data mining, topic modeling refers to the utilization of hierarchical probabilistic models to discover word patterns that reflect the underlying topics in a set of document collections [3].

3.1 Our Framework

Here, we explain our proposed narrative framework. The framework comprises of 2 steps: I. Narrative modeling based on topic identification over time; and II. extractive summarization from the identified narratives.

To discover the narratives over topical events, first, we use our discrete-time generative narrative

[1] Portions of this chapter is reprinted, with permission, from ©2020 ACM [105] https://doi.org/10.1145/3372923.3404790.

model as an unsupervised learning algorithm to learn the distribution of textual contents from daily conversation cascades. Then, we extract narrative summaries over topical events from sentences in the time categories. This is achieved by sampling from the identified distribution of narratives and performing sentence ranking. Narrative modeling and summarization steps are explained below in separate subsections.

3.1.1 Topic Modeling to Identify Narratives

To model narratives, we design our topic model such that the discovered topics present a series of timely ordered topical events. Accordingly, the topical events deliver a narrative covering distinct events over the same topic. In this regard, we present Narratives Over Categorical time (NOC)[2], a novel probabilistic topic model that discovers topics based on both word co-occurrence and temporal information to present a narrative of events. According to the topic-time relationship explained above, we refer to the topics over time as narratives, topical events as events, and the extracted timely ordered sentences of documents with a high probability of belonging to each event as the extracted narrative summary. To fully comply with the definition of narrative, we assume a causality relation between the conversation cascades in social media. However, we do not investigate the causality relation across the conversation cascades or named-entities.

The differences between our narrative model with dynamic topic models [9], topic models with continuous-time distribution [145], and hierarchical topic models [53, 117] include: not filtering the data for a specific event, imposing sharp transition for topic-time changes with time slicing, discovering topical events without scalability and sparsity issues, allowing multimodal topic distribution in time as a result of categorical time distribution, and allowing to select a time slicing

[2]The code for NOC is available at our GitHub repository on `https://github.com/toktammm/Twitter_Topics_over_Time`

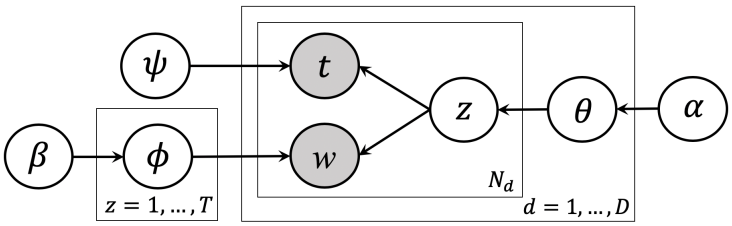


Figure 3.1: The graphical model for NOC with Gibbs sampling.

size such that distinct topical events be recognizable. Additionally, categorical time distribution enables discovering topical events with varying time resolution, for instance, weekly, biweekly, and monthly. In contrary to dynamic topic modeling, this method does not requiring filtering the documents to have an equal number of documents per time slice.

Time discretization brings the question of selecting the appropriate slicing size or the number of categories that depend on the characteristics of the dataset under study. On the contrary, topical models with continuous-time distribution cannot model many modes in time. Additionally, continuous-time models such as [145] suffer from instability problems if the dataset is multimodal and sparse in time. Furthermore, categorical time enables discovering topic recurrence which results in identifying topical events related to distinct narrative activities, which is of our interest in this work. Narrative activities in social media refer to the amount of textual content that is circulating in online platforms over time, corresponding to a specific topic.

The generative process in NOC, models timestamps, and words per documents using Gibbs sampling which is a Markov Chain Monte Carlo (MCMC) algorithm. The graphical model of NOC is illustrated in Figure 3.1. As can be seen from the graphical model, the posterior distribution of topics is dependent on both text and time modalities. This generative procedure can be described as follows:

I. For each topic z, draw T multinomials ϕ_z from a Dirichlet prior β;

II. For each document d, draw a multinomial θ_d from a Dirichlet prior α;

III. For each word w_{di} in d:

(a) draw a topic z_{di} from multinomial θ_d;

(b) draw a word w_{di} from multinomial $\phi_{z_{di}}$;

(c) draw a timestamp t_{di} from categorical $\psi_{z_{di}}$;

where the timestamps t_{di} for words w_{di} in each document d are identical. The list of symbols and their descriptions can be found in table 3.1. The model parameterization is as below:

$$\begin{aligned} \theta_d|\alpha &\sim \text{Discrete}(\alpha) \\ \phi_z|\beta &\sim \text{Discrete}(\beta) \\ z_{di}|\theta_d &\sim \text{Multinomial}(\theta_d) \\ w_{di}|\phi_{z_{di}} &\sim \text{Multinomial}(\phi_{z_{di}}) \\ t_{di}|\psi_{z_{di}} &\sim \text{Categorical}(\psi_{z_{di}}) \end{aligned} \tag{3.1}$$

In this model, Gibbs sampling provides an approximate inference instead of exact inference. To calculate the probability of topic assignment to word w_{di}, we first need to calculate the joint probability of the dataset as $\mathbb{P}(z_{di}, w_{di}, t_{di}|w_{-di}, t_{-di}, z_{-di}, \alpha, \beta, \psi)$ and use chain rule to derive the probability of $\mathbb{P}(z_{di}|w, t, z_{-di}, \alpha, \beta, \psi)$ as below, where $-di$ subscript refers to all tokens except w_{di}:

$$\begin{aligned} \mathbb{P}(z_{di}|w, t, z_{-di}, \alpha, \beta, \psi) &\propto (m_{dz_{di}} + \alpha_{z_{di}} - 1) \\ &\times \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v} + \beta_v) - 1}\, p(t_{z_{di}} \in b_k) \end{aligned} \tag{3.2}$$

Variable Descriptions	Symbol
Number of topics	T
Number of documents	D
Number of unique words	V
Number of word tokens in document d	N_d
Multinomial distribution of topics for document d	θ_d
Multinomial distribution of words for topic z	ϕ_z
Categorical distribution of time for topic z	ψ_z
Topic of the ith token in document d	z_{di}
ith word token in document d	w_{di}
Timestamp for ith word token in document d	t_{di}
Time category for timestamp associated with a token	b_k
jth sentence of document d	s_{dj}

Table 3.1: NOC symbols and definitions

where n_{zv} refers to the number of words v assigned to topic z, m_{dz} refers to the number of word tokens in document d that are assigned to topic z, and b_k represents the kth time slice. The details on the Gibbs sampling derivation can be found in subsection 3.1.3. After each iteration of Gibbs sampling, we update the probability of $p(t_{z_{di}} \in b_k)$ as follows:

$$p(t_{z_{di}} \in b_k) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(t_{z_{di}} \in b_k) \tag{3.3}$$

where $\mathbb{I}(.)$ is equal to 1 when $t_{z_{di}} \in b_k$, and 0 otherwise.

3.1.2 Narrative Summary Extraction

We employ the discovered probabilities of topics over documents, θ, probabilities of words per topic, ϕ, and probabilities of topics per time category, ψ to perform sentence ranking. This ranking allows extracting the sentences with the higher scores of belonging to each topic. This is

achieved via performing weighted sampling on the collection of documents based on the probabilities of topics per time category ψ and draw D documents from θ. The weighted sampling leads to drawing more documents from the time categories b_k with a higher ψ as this time slices contain more documents related to the topic z. Each document contains a sequence of sentences $(s_1, s_2, \ldots, s_J) \in d$ from the aggregated conversation cascades per day. Information on the aggregation of conversation cascades and document preparation can be found in section 3.2.1.

Since the social media narrative activity over a topic evolves from the circulation of identical or similar textual content in the platform, the content involves significant similarity. For instance, the Twitter conversation cascades include replies, quotes, and comments, where replies and quotes duplicate the textual content. Therefore, we applied Jaro-Winkler distance over the timely ordered sentences and dismissed the sentences with similarity above 70%, while keeping the longest sentence. After removing redundant text as described earlier, we calculate the probability of each sentence s_j by measuring the sum of the probabilities of topics for words $w_{di} \in s_j$. Then, we select the sentences with the highest accumulative probability of words w per topic z. Summary coherence was induced as suggested in [5] by ordering the extracted sentences according to their timestamps such that the oldest sentences appear first. The results of the extracted summaries are provided in Chapter 3.5.

3.1.3 Gibbs Sampling Derivation for the Discrete-Time Narrative Model

Starting with the joint distribution $\mathbb{P}(w, t, z|\alpha, \beta, \psi)$, we can use conjugate priors to simplify the equations as below:

$$
\begin{aligned}
&\mathbb{P}(w,t,z|\alpha,\beta,\psi) = \mathbb{P}(w|z,\beta)\ p(t|\psi,z)\ \mathbb{P}(z|\alpha) \\
&= \int \prod_{d=1}^{D}\prod_{i=1}^{N_d} \mathbb{P}(w_{di}|\phi_{z_{di}}) \prod_{z=1}^{T} p(\phi_z|\beta) d\Phi \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
&\times \int \prod_{d=1}^{D} \Big(\prod_{i=1}^{N_d} \mathbb{P}(z_{di}|\theta_d)\ p(\theta_d|\alpha) \Big) d\Theta \\
&= \int \prod_{z=1}^{T}\prod_{v=1}^{V} \phi_{zv}^{n_{zv}} \prod_{z=1}^{T} \Big(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{zv}^{\beta_v - 1} \Big) d\Phi \\
&\times \int \prod_{d=1}^{D}\prod_{z=1}^{T} \theta_{dz}^{m_{dz}} \prod_{d=1}^{D} \Big(\frac{\Gamma(\sum_{z=1}^{T}\alpha_z)}{\prod_{z=1}^{T}\Gamma(\alpha_z)} \prod_{z=1}^{T} \theta_{dz}^{\alpha_z - 1} \Big) d\Theta \\
&\times \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
&= \Big(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \Big)^{T} \Big(\frac{\Gamma(\sum_{z=1}^{T}\alpha_z)}{\prod_{z=1}^{T}\Gamma(\alpha_z)} \Big)^{D} \prod_{d=1}^{D}\prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
&\times \prod_{z=1}^{T} \frac{\prod_{v=1}^{V}\Gamma(n_{zv}+\beta_v)}{\gamma(\sum_{v=1}^{V}(n_{zv}+\beta_v))} \prod_{d=1}^{D} \frac{\prod_{z=1}^{T}\Gamma(m_{dz}+\alpha_z)}{\gamma(\sum_{z=1}^{T}(m_{dz}+\alpha_z))},
\end{aligned}
\tag{3.4}
$$

where $\mathbb{P}$ and p refer to the probability mass function (PMF) and probability density function (PDF), respectively. The conditional probability $\mathbb{P}(z_{di}|w,t,z_{-di},\alpha,\beta,\psi)$ can be found using the chain rule as:

$$
\begin{aligned}
\mathbb{P}(z_{di}|w,t,z_{-di},\alpha,\beta,\psi) &= \frac{\mathbb{P}(z_{di},w_{di},t_{di}|w_{-di},t_{-di},z_{-di},\alpha,\beta,\psi)}{\mathbb{P}(w_{di},t_{di}|w_{-di},t_{-di},z_{-di},\alpha,\beta,\psi)} \\
&\propto \frac{\mathbb{P}(w,t,z|\alpha,\beta,\psi)}{\mathbb{P}(w_{-di},t_{-di},z_{-di}|\alpha,\beta,\psi)} \\
&\propto \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v}+\beta_v) - 1}\ (m_{dz_{di}} + \alpha_{z_{di}} - 1)\ p(t_{di}|\psi_{z_{di}}) \\
&\propto (m_{dz_{di}} + \alpha_{z_{di}} - 1)\ \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v}+\beta_v) - 1}\ p(t_{z_{di}} \in b_k)
\end{aligned}
\tag{3.5}
$$

The probability of $p(t_{di} \in b_k)$ can be measured as follows:

$$p(t_{z_{di}} \in b_k) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(t_{z_{di}} \in b_k), \tag{3.6}$$

where $\mathbb{I}(.)$ is equal to 1 when $t_{z_{di}} \in b_k$, and 0 otherwise.

3.1.4 Proposed Metric: Analyzing Lifetime Attractiveness of Topics with Shannon Entropy

The topic attractiveness to social media users can be investigated by basic statistical methods as a measure of the length of conversation cascades, the number of initiated textual content, and the number of unique users performing an activity relative to the underlying topic. The user activity fluctuations for timestamped data may contain activity bursts that are illustrative of significant events. Similarly, the generation and propagation of textual content within an online platform can illustrate the narrative activity relative to the events over time, where a burst represents a significant narrative activity. Additionally, the recurrence of a topic can be considered as an attractiveness measure for the associated topic.

In this regard, we propose the significance-dispersity trade-off (SDT) metric to compare the identified narratives against each other. SDT measures the lifetime attractiveness of the identified narratives based on the distribution of narratives over topical events. The significance of such measurement is evident when speculating the number of involved individuals in a topic is not feasible as the dataset lack information regarding the number of individuals in the conversations. The proposed metric quantifies the significance of the narrative activities and recurrence of a topic via employing the Shannon entropy for the discovered narrative distributions. The intuition behind the SDT score is that the value of the entropy is maximum when the probability distribution is uniform. On the contrary, this value is minimum if the distribution is a delta function. This is visualized in Figure 3.2.

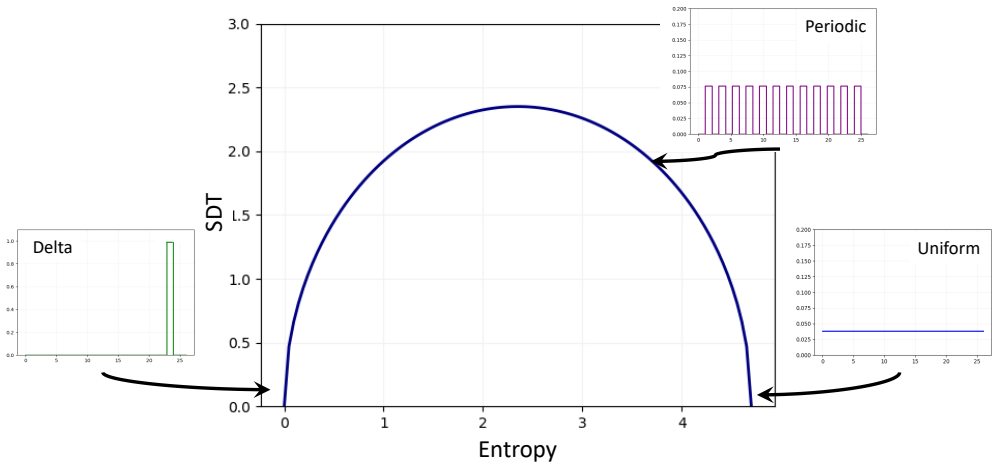


Figure 3.2: A visualization of SDT for different entropy values using $\gamma = 0.5$. The SDT values for delta, uniform, and periodic distributions are marked on the graph.

We define dispersity of a categorical time topic distribution as a measure of the dispersion of the time categories. Based on this definition, SDT score of topic z can be obtained as:

$$SDT_z = H^{\gamma}(H_{max} - H)^{1-\gamma}, \tag{3.7}$$

where H is the Shannon entropy for the categorical distribution of time for topic z:

$$H_z = -\sum_{k}^{K} p_z log_2 p_z, \tag{3.8}$$

$$H_{max} = log_2(K),$$

and K refers to the number of time slices in the distribution.

We rely on this intuition that social media topics with high lifetime attractiveness are significant and recurrent. However, the probability distribution imposes a trade-off on the two. The parameter γ

Variable Descriptions	Symbol
Entropy of topic z	H_z
Number of time slices	K
Weighted geometric mean	γ

Table 3.2: SDT symbols and definitions

provides a weighted geometric mean of H and $H_{max} - H$ that enables promoting either significance or recurrence, dependent on the application under study. A larger value of parameter γ promotes dispersity for SDT score, and a smaller amount of this parameter promotes mode significance. The bounds for the SDT score are:

$$SDT_i = \begin{cases} 0 & \text{if } H = 0 \ \& \ \gamma! = 0 \\ 0 & \text{if } H = H_{max} \ \& \ \gamma! = 1 \\ \gamma^{\gamma}(1-\gamma)^{1-\gamma} H_{max} & \text{if } H = \gamma H_{max} \ \& \ 0 < \gamma < 1 \end{cases} \tag{3.9}$$

where $H = 0$ occurs when the distribution under study is uniform, and $H = H_{max}$ relates to delta distribution.

Since the time categorical distribution of our narrative model allows many modes in time, recurrent narratives can be identified. Additionally, the narrative activity fluctuations can be modeled using categorical time distribution in topic analysis. The results presented in the next chapter suggest that SDT score can be used to identify the narrative with higher lifetime attractiveness in a timestamped dataset.

The evaluation of the proposed technique, choice of parameters, and details on the dataset are provided next.

3.2 Experiments and Results

We conduct extensive analysis to investigate the performance of the proposed narrative summary extraction framework. The investigated dataset, experiments, and results are presented next.

3.2.1 Dataset Description

To analyze topical events and provide narratives, we investigate the Twitter dataset on the domain of White Helmets of Syria over a period of 13 months from April 2018 to April 2019. This dataset was provided to us by Leidos Inc[1] as part of the Computational Simulation of Online Social Behavior (SocialSim)[2] program initiated by the Defense Advanced Research Projects Agency (DARPA). We analyze more than 1,052,000 tweets from April 2018 to April 2019.

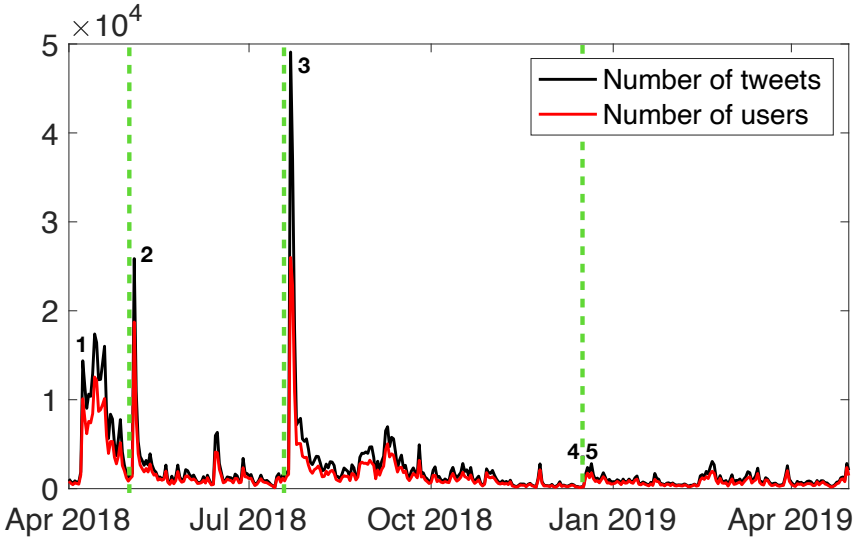
For the investigated Twitter dataset, Figure 3.3 demonstrates the daily number of user activities and the number of unique users who are involved in these activities, which can help to obtain a better understanding of the data, as well as determining the number of topics. It can be observed that some essential events triggered the burstiness of the specific Twitter cascades at specific times that can be considered as the topics in this dataset. We manually investigated these specific times and mark these events on this figure.

To prepare the model inputs, we filter the tweets from the non-English text. Then, we clean up the data by removing usernames, short URLs, as well as emoticons. Additionally, we remove the stopwords, perform Part of Speech (POS) tagging, and Named Entity Recognition (NER) on each tweet using Stanford Named Entity Recognizer[3] model. Using the NER tool, we extract persons,

[1] `https://www.leidos.com/`

[2] `https://www.darpa.mil/program/computational-simulation-of-online-social-behavior`

[3] `https://nlp.stanford.edu/software/CRF-NER.html`



1 - Douma chemical attack (source: *New York Times 04/08/18*)
2 - Trump's Syria aid freeze hits 'White Helmets' rescue workers (source: *ABC News 05/04/18*)
3 - Syrian White Helmets evacuated to Jordan through Israel (source: *CBC News 07/22/18*)
4 - The Russia-backed campaign to link the volunteer rescuers with al-Qaida exposes how conspiracy theories take root: 'It's like a factory' (source: *The Guardian 12/18/18*)
5 - Air strikes kill at least 9 civilians in Syria's Idlib - White Helmets (source: *TRT 12/23/18*)

Figure 3.3: Daily number of user activities (black line) and unique user involved (red line). Via manual investigation, titles of the news possibly related to the bursts in Twitter cascades are detected. These news titles are marked with green dashed line.[3]

locations, and organizations and removed all pseudo-documents that do not contain named entities similar to [90]. Furthermore, We remove the tweets shorter than 3 words.

Anywhere on the same page where the float appears

As Twitter maintains a maximum allowed character limit of 280 characters, collected tweets lack context information and have very low word co-occurrence. We tackle the challenge of topic

[3]This Figure is reprinted, with permission, from ©2020 IEEE/ACM [99] `https://doi.org/10.1109/ASONAM49781.2020.9381386`.

modeling on short-text tweets and to include plentiful context information by preparing pseudo-documents for our model inputs via aggregating daily root, parent, and reply/quote/retweet comments in each conversation cascade. We maintain the order of the conversation according to the timestamps associated with each tweet. This text aggregation method results in preparing pseudo-documents rich in context and related words with a daily time resolution. We use the pre-processing phase output as the model input pseudo-documents, referred to as documents in this paper. The results for this model is provided in section 3.2.4.

3.2.2 Performance Measurements

3.2.2.1 Coherence Metric

The identified narratives can be evaluated using effective evaluation metrics for topic models. Accordingly, we calculate pointwise mutual information [101] to measure the coherence of a topic z as follows:

$$Coh_z = \frac{2}{K(K-1)} \sum_{j<k\leqslant K} log \frac{p(w_j, w_k)}{p(w_j)p(w_k)}, \tag{3.10}$$

where K is the number of most probable words for each narrative, $p(w_j)$ and $p(w_k)$ refer to the probabilities of occurrence for words w_j and w_k, and $p(w_j, w_k)$ represents the probability of co-occurrence for the two words in the collection of documents.

3.2.2.2 Significance-Dispersity Trade-off

The Significance-Dispersity Trade-off (SDT) score is a measurement that we proposed. We defined dispersity of a categorical time topic distribution as a measure of the dispersion of the time

categories for the identified topics. Based on this definition, SDT score of can be calculate as:

$$SDT_z = H^{\gamma}(H_{max} - H)^{1-\gamma}, \tag{3.11}$$

where H is the Shannon entropy for the categorical distribution of time for topic z. The parameter γ provides a weighted geometric mean of H and $H_{max} - H$ that enables promoting either significance or recurrence. A larger value of parameter γ promotes dispersity for SDT score, and a smaller amount of this parameter promotes mode significance.

3.2.3 Experiment Setup

In this work, we report results with bi-weekly categorical time resolution. To determine the values for hyper-parameters α and β and to investigate the sensitivity of the model to these values, we repeated our experiment with symmetric Dirichlet distributions using values $\alpha \in [0.1, 0.5, 1]$, $\beta \in [0.01, 0.1, 0.5, 0.8, 1]$. We observed that the model did not show significant sensitivity to the values of these hyper-parameters. Thus, we fix $\alpha = 1$ and $\beta = 0.5$, both as symmetric Dirichlet distributions. We initialize the hyperparameter ψ in 2 ways for comparison: I. random initialization (model referred as NOC_R); and II. based on the probability of user activity per time category, illustrated in Figure 3.6, (model referred as NOC_A).

To estimate the number of topics for our experiments, we first visualize the tweets' hashtag co-occurrence graph. We measure the graph modularity to examine the structure of the communities in this graph. We observe the highest modularity score of 0.41 using modularity resolution equal to 0.85. Figure 3.4 illustrates a downsampled version of this graph, where each color represents a modularity class. The edges of the graph are weighted according to the number of hashtags' co-occurrence in the document collection. Our modularity analysis suggests that few distinct hashtag

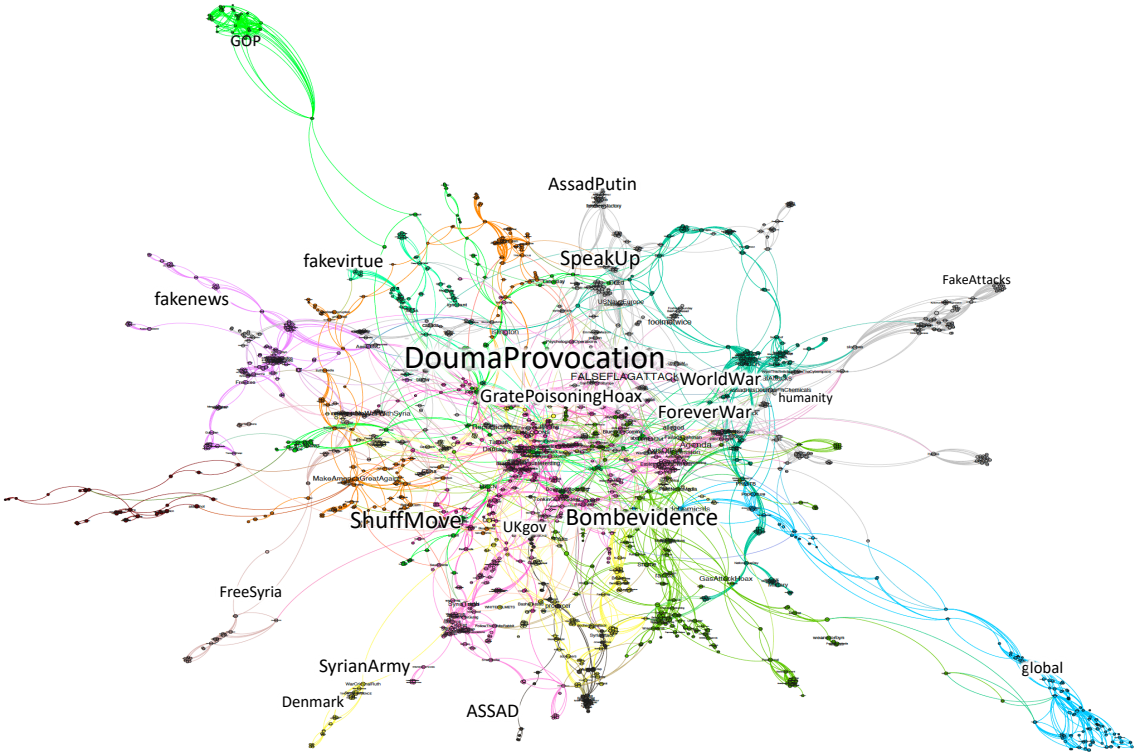


Figure 3.4: The hashtag co-occurrence graph for twitter dataset on the domain of White Helmets of Syria for a period of 13 months, from 2018 to 2019. This graph represents a down-sampled version of the hashtag co-occurrence for this data for the sake of visualization.

communities exist. Additionally, the dataset under study contains tweets associated with a single domain. As a result, we assume the number of topics to be relatively low. To choose an appropriate number of topics, we repeated LDA with the number of topics as $T \in [4, \ldots, 20]$ with increments of size 1. We evaluated the c_v coherence of topics identified by LDA and observed the highest coherence score for $T = 5$ and $T = 6$, respectively. Thus, we report our experimental results using these values.

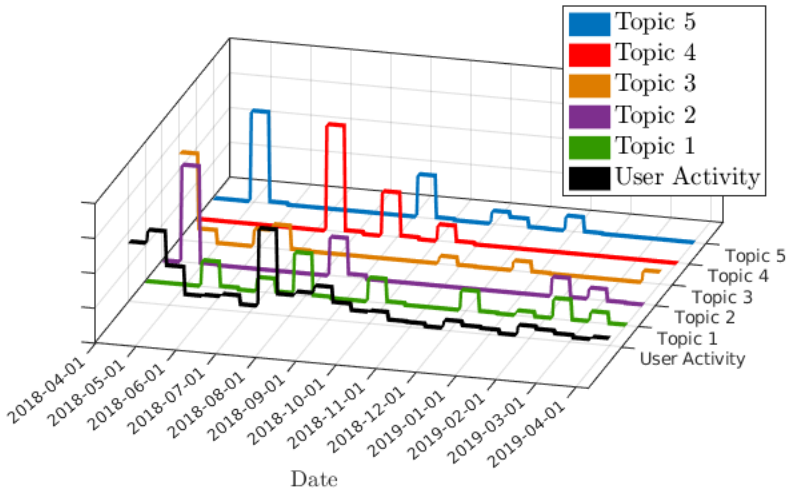
As discussed in 3.1.2, we employ the discovered probabilities of topics over documents, θ, probabilities of words per topic, ϕ, and probabilities of topics per time category, ψ to perform sentence ranking to extract a summary for topics, such that sentences with the higher scores of belonging

to each topic appear in the topic summaries. We apply weighted sampling on the collection of documents based on the probabilities of topics per time category ψ and draw D documents from θ. Summary coherence was induced as suggested in [5] by ordering the extracted sentences according to their timestamps such that the oldest sentences appear first. Table 3.5 in the 3.2.4 section contains the extracted narrative summaries for 5 topics for a sample run.

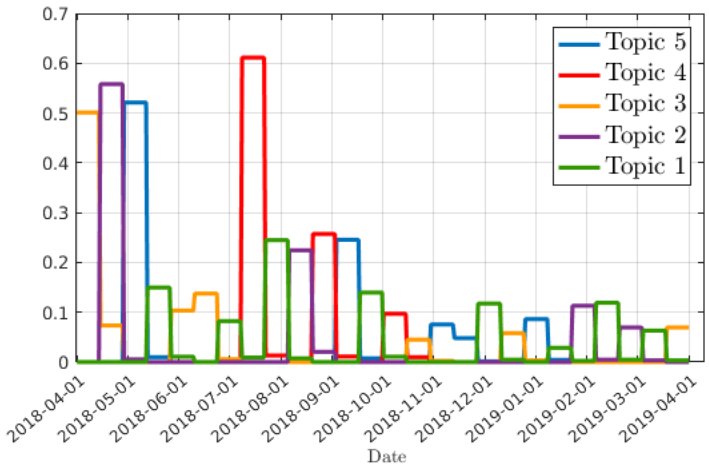
3.2.4 Results

As mentioned earlier, the discovered topics by NOC present a series of timely ordered topical events. Thus, the topical events deliver a narrative covering distinct social media events over the same topic. Figure 3.5 demonstrates the generated narrative distributions with NOC, where the hyperparameter ψ was randomly initialized (referred to as NOC_R). In this figure, the vertical axis represents the empirical probabilities of the topics. This figure represents that the identified narratives by our model are distinct from each other and the collapsed distribution of all narratives approximates the distribution of social media user activity over time.

Below we compare our model with LDA and TOT[145] based on the coherence score achieved for both methods. TOT is a probabilistic topic model over time with Beta distribution for time. Table 3.3 displays the average coherence score measured across the discovered topics by LDA, TOT, and NOC. For NOC, we investigate initializing the parameter ψ with random and user activity-based initialization, referred to as NOC_R and NOC_A, respectively. We consider $K = 500$ most probable words from each topic. This comparison suggests that the narratives identified by NOC are more coherent than the identified topics by LDA, with an improvement in coherence of about 35%. The observed improvement compared with TOT was about 27%. Additionally, initializing the hyperparameter ψ in NOC using the distribution of user activity improves the narrative coherence by about 3%.



(a)



(b)

Figure 3.5: The distribution of extracted topics and user activity over time: a) The distribution of user activity over time is depicted by black, followed by the 5 distribution of the extracted topics, using NOC_R; b) The collapsed distribution of the 5 extracted topics. The vertical axis represents the empirical probabilities of the topics.

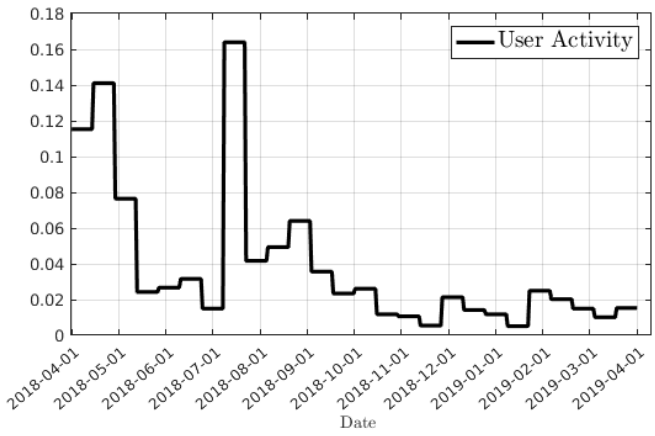


Figure 3.6: The distribution of user activity over time. Comparing this figure with Fig. 3.5 suggest that the distribution of extracted topics approximates the distribution of user activity over the same time period.

Model	***LDA***	***TOT***	***NOC_R***	***NOC_A***
T = 5	5.980	6.36	7.95	8.23
T = 6	5.546	5.99	7.75	7.98

Table 3.3: The comparison of coherence scores for 4 models:

Table 3.4 provides a comparison for the SDT scores measured for the 5 identified narratives, using varying values of γ. The illustration of the distribution of the extracted narratives can be seen in Figure 3.5a. We can clearly see in this figure that narratives 1 and 3 have the highest dispersity. On the contrary, narratives 4 and 2 have the highest significance. We compare SDT_i for narrative i with the number of user activity associated with narrative z. The results suggest that SDT score can be used to identify the narrative with higher lifetime attractiveness in a timestamped dataset. In our experiments, this is achieved for topic 1 when the value of γ is greater than or equal to 0.7. As it can be seen, this topic is associated with the highest user activity count, reported in the same table.

Topic / Narrative	*z = 1*	*z = 2*	*z = 3*	*z = 4*	*z = 5*
$\gamma = 0$	1.59	2.90	2.39	**3.21**	2.75
$\gamma = 0.4$	2.08	**2.40**	2.36	2.36	**2.40**
$\gamma = 0.7$	**2.54**	2.08	2.33	1.87	2.16
$\gamma = 1$	**3.11**	1.80	2.31	1.49	1.95
User Activity	**353,280**	317,686	244,674	247,895	175,343

Table 3.4: The comparison of SDT scores for 5 topic narratives:

3.3 Conclusion

In this chapter, we addressed the problem of narrative modeling and narrative summary extraction for social media content. For this, we investigated the Twitter dataset on the domain of White Helmets of Syria over a period of 13 months from April 2018 to April 2019. This dataset was provided to us as part of the Computational Simulation of Online Social Behavior (SocialSim) program initiated by the Defense Advanced Research Projects Agency (DARPA), under grant number FA8650-18-C-7823.

We presented a narrative framework consisting of I. Narratives over topic Categories (NOC), a probabilistic topic model with categorical time distribution; and II. extractive text summarization. The proposed narrative framework identifies narrative activities associated with social media events. Identifying topics' recurrence and significance over time categories with our model allowed us to propose significance-dispersity trade-off (SDT) metric. SDT can be employed as a comparison measure to identify the topic with the highest lifetime attractiveness in a timestamped corpus. Results on real-world timestamped data suggest that the narrative framework is effective in identifying distinct and coherent topics from the data. Additionally, the results illustrate that the identified narrative distributions approximate the user activity fluctuations over time. moreover, informative, and concise narrative summaries for timestamped data are produced.

Topics	Keywords	Summary
Topic 1	Terrorist, Idlib, Civilian, Child, City, Attack, Aleppo, Rescue, Weapon, Killed	WhiteHelmets Syria News: One child was injured in the north of Aleppo. Their aim is to save lives in war zones inside Syria. Has credibly substantiated 336 uses of ChemWeapons in Syria 98% of attacks by Assadallies. These are the WHITE HELMETS or Syria Civil Defense as our US Dept of State calls them!! Russian airstrikes killed two men and one baby in DMZ areas RussianWarCrimes.
Topic 2	Chemical, Attack, Douma, Video, Idlib, Staged, Boy, War, Child, Witness	Remember first they said the video including the pics of the chlorine cylinder was fake. Whitehelmets One America News Pearson Sharp Visits Hospital in Douma Where White Helmets Filmed Chemical Attack Hoax Multiple Eyewitness Doctors Say No Chemical Attack Took Place Syria. This is the video evidence of the airstrike on Zardana an Idlib town controlled by Very expensive camera on the helmet of the WhiteHelmets rescuer. White Helmets making films of chemical attacks with children in Idlib.
Topic 3	Chemical, Attack, Douma, Terrorist, Fake, Child, Propaganda, Video, Russian, Russia	From the fabrication of the plays of the chemist and coverage of the crimes of terrorism to the public cooperation with the Israeli army the white helmets. They are holding children! Another chemical attack is imminent its all they've got left! 4 dead including two children and more than 50 wounded mostly women and children. Love the White Helmets propaganda almost as untruthful as the BBC.
Topic 4	Israel, Terrorist, Idlib, Chemical, Attack, Life, Rescue, Russian, People, Al Qaeda	WHITE HELMETS ARE PREPARING CHEMICAL ATTACK ON CITIZENS AGAIN! Those are basically just members of Al Qaeda Al Nusra right? The Al Qaeda smear is deliberate propaganda. Its war crime only If US intervenes in Kashmir Kashmir will be liberated like Raqqan with a dozen US bases having Thaad missiles aimed at China and with AlQaeda WhiteHelmets taking out children's organs of Kashmiris.
Topic 5	Funding, Freeze, Trump, Terrorist, Group, Chemical, Attack, Idlib, Civilian, News	Trumps USA has built a rationale for its public that it will need to support rebels in holding on to a large chunk of Syria. I wonder how it is possible that criminal associations such as WhiteHelmets and the Syrian Human Rights Observatory can make the world go round as they want by influencing the policies of world leaders. U.S. freezes funding for Syrias White Helmets. White helmets are terrorists. Former Head of Royal Navy Lord West on BBC White Helmets Aren't Neutral They're On The Side Of The Terrorists.

Table 3.5: Identified representative keywords and topic summaries (narratives).

The summaries provided here are the results for a sample run of the proposed narrative framework and do not reflect authors' personal opinions.

Further improvement of the narrative framework can be achieved via incorporating the causality relation cross the social media conversation cascades and social media events into account. Other future directions are identifying topical hierarchies and extract summaries associated with each hierarchy.

CHAPTER 4: DOMINANT SET-BASED ACTIVE LEARNING FOR TEXT CLASSIFICATION AND ITS APPLICATION TO ONLINE SOCIAL MEDIA

In this chapter, we describe our active learning method for social media text classification. The growing volume of online publicly available user-generated content has encouraged an enormous amount of research on the design and application of natural language processing techniques using online social media datasets [6, 96, 57]. Recent advances in natural language processing (NLP), including research on online social media analysis and mining are evidently owed to large-scale datasets, deep language models with attention mechanisms, and transferring of general knowledge via per-training. However, labeling, storing, and processing a large amount of textual data, e.g., tweets, which is required for training deep language models has remained challenging. On top of that, manually analyzing textual data that contains rumors, misinformation, hate speech and bullying, etc. to label a sufficiently large dataset can be mentally and emotionally taxing for human annotators. As a result, a substantial amount of recent NLP research is focused on techniques that can make the best use of significantly less amount of labeled data.

Active learning refers to the process of efficient selection of the most informative data when the data is plentiful, but the labels are scarce [129]. Active Learning (AL) techniques can mitigate the issues associated with manual labeling and improve automatic detection and classification when labeled training data is sparse [33]. The best active learning strategy successfully selects certain unlabeled data samples from the distribution of available data, such that using this data portion for training leads to the maximal reduction of the classification error and variance.

Although this area of research is not new, very few research papers are focused on applying active learning for natural language, defining data informativeness for contextual data, and investigating

the potential of popular data selection techniques on low-quality data, such as imbalanced data that misses context. This doctoral thesis is specifically focused on extremely imbalanced data, which is a common issue in research on many social media datasets, e.g., spam detection, social bot detection, and hate speech classification [95, 24]. Thus, we evaluate our proposed technique using two common social media datasets related to hate speech classification and show that our method consistently achieves a higher performance in comparison to the state-of-the-art active learning strategies. We show the effectiveness of our method on different datasets and using different neural network architectures.

4.1 Our Framework

Here, we explain our proposed active learning method. The framework comprises of 2 main steps that occur in every cycle of active learning: I. extracting feature vectors that are rich with contextual information and general language knowledge using deep pre-trained language models as the embedding function; and II. identifying the non-dominant samples of clusters and use them for fine-tuning the language model.

We are interested in pool-based active learning [89], in which an acquisition function is used to query the label of a small set of selected samples. In this method, the model is initially trained using a small set of labeled data. Then, according to an informativeness criterion, the acquisition function selects a few data points to query their labels from the oracle. For a model $\mathscr{M}(x;\theta)$, pool data $\mathscr{D}_{pool}$, and input $x \in \mathscr{D}_{pool}$, the acquisition function $a(x,\mathscr{M})$ is defined as:

$$x^* = \arg\max_{x \in \mathscr{D}_{pool}} a(x,\mathscr{M}).$$

The most informative samples are drawn from the pool and added to the training set by repeating

the above step. For instance, investigating the uncertainty of the model as the informativeness criterion is a common approach for many active learning acquisition functions, with the hope that selecting samples based on such criterion leads to a lower model uncertainty. In this work, we use various acquisition functions for comparison with the proposed strategy, which are provided in section 4.2.3.

Recent advancements in the development of pre-trained language models, such as Google's BERT (Bidirectional Encoder Representations from Transformers) architecture [32] and OpenAI's GPT-3 (Generative Pre-trained Transformer) model [13] have revolutionized the field of natural language processing and deep learning. As these models are trained over massive examples of written language, they can be easily fine-tuned for a downstream task using a small amount of labeled data [33]. Accordingly, employing these models in active learning is ideal and leads to the reduction of cost and burden of manual labeling, without loss of generalizability and performance across many diverse tasks.

We discuss and investigate our proposed active learning strategy for the task of social media text classification. Our goal is to find the smallest set of training samples that leads to the maximal reduction of the classification error and variance. As we are interested in the application of this strategy to text classification, we exploit the feature vectors from the BERT model to calculate the informativeness scores using our proposed acquisition functions described below. Since the BERT model is pre-trained over a huge amount of data points and can provide us with high-quality feature space, even before supervised fine-tuning, it is an ideal candidate for tasks such as active learning.

4.1.1 Dominant Set Clustering and its Applications

Despite the existence of many clustering techniques, graph-theoretic clustering methods [130, 124, 112], such as spectral clustering, dominant sets [109], and Density-Based Spatial Clustering of Ap-

plications with Noise (DBSCAN) [74] achieve higher performance than k-means in the discovery of the clusters of arbitrary shapes, and are more robust to noise and outliers. Among these methods, dominant sets clustering [109], which refers to strongly coherent subsets of local clusters [109], has the lowest computational complexity.

We propose to exploit dominant sets into informative data selection for active learning. However, our work is not the first study that employs dominant or non-dominant sets for machine learning and classification, nor the first study that investigates these sets for active learning. In fact, dominant sets are widely used for different applications and tasks. The work of [14] provides a review of methods that have adopted dominant sets, as well as the extensions of this algorithm. A related work on the exploitation of dominant sets for active learning is [64], which also uses both spectral clustering and dominant sets to find the outliers for label investigation. However, this work is not designed to be used for the training of deep models. The dissimilarities to this doctoral work are discussed in the next part.

4.1.2 Active Learning Using Non-Dominant Set (NDS)

In this doctoral thesis, we first, identify the local clusters of data in feature space and for each cluster detect the most similar set of points, i.e., the dominant set. The samples that do not belong to the dominant set of any of the clusters are the least similar samples to their corresponding clusters, and therefore the most challenging to classify. Furthermore, these samples are more likely to lie on the decision boundary. Thus, we can collect these informative samples without needing a classifier. Figure 4.1 illustrates an overview of the steps of the proposed method for a toy 2D feature space and 3 classes. Dominant set clustering is non-parametric and is a sequential method that only uses a predefined similarity matrix to find the cohesive structures in that space. We use the number of classes in our classification task as a known parameter for an initial clustering before

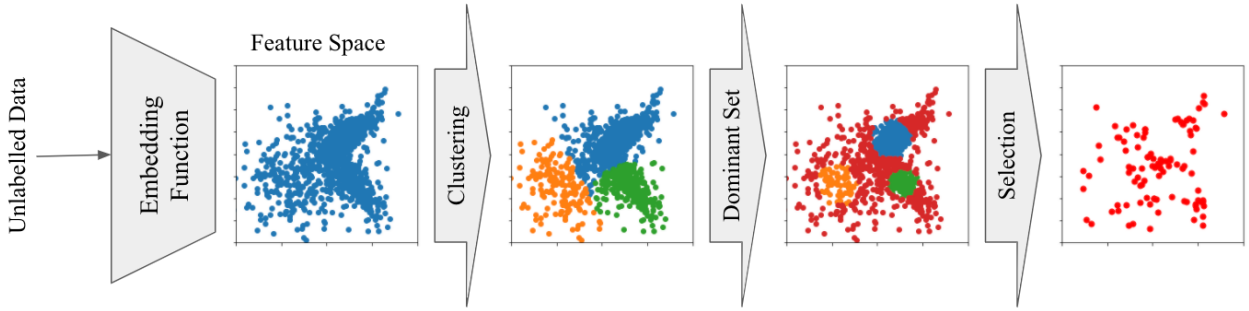


Figure 4.1: Overview of the proposed NDS active learning framework for a toy 2D feature space with 3 classes. At each active learning cycle, an embedding function (e.g., pre-trained BERT) maps the input onto the feature space. Then, spectral clustering is used to cluster the extracted feature vectors. For each cluster, the dominant samples are identified by measuring the sample-cluster similarity. Finally, samples are selected randomly from the non-dominant set and labeled by an annotator. The selected samples represent the structure of the data well and most likely lie near the decision boundaries.

finding the dominant sets in each cluster. We assume that with little domain knowledge the number of classes must be known. Also, applying this method allows parallel dominant set identification across clusters, which makes it more practical for large datasets.

Here, we explain NDS in detail, which is a feature similarity-based active learning approach. Our selection strategy starts with spectral clustering of the feature vectors for the pool of unlabeled data to obtain a set of K clusters $C_1 \dots C_K$ based on a pairwise local feature similarity score (e.g. Euclidean distance), where K is equal to the number of classes in our classification problem. Although this is not an entirely realistic assumption, we suppose in an "ideal case" the model can provide a feature space that contains the same number of clusters as the classification problem (K). We extract a dominant set from each cluster C_k by constructing an undirected edge-weighted graph $G_k(V_k, E_k, w_k)$ with no self-loops. V_k and E_k represent the set of vertices and edges per cluster C_k, and neighborhood relationships define the existence of edges. $a_k^{ij} = w_k(i,j)$ refers to the edge weight between feature vectors i and j if $(i,j) \in E_k$, and $a_k^{ij} = 0$ otherwise. In other words, the vertices in the graph G_k represent the samples in the k^{th} cluster and the edges represent the distance

Variable Descriptions	Symbol
Model	$\mathscr{M}$
Pool of training data for selection	$\mathscr{D}$
k_{th} cluster	C_k
Set of vertices for cluster k	V_k
Set of edges for cluster k	E_k

Table 4.1: NDS-AL symbols and definitions

between the samples. The proper distance function depends on the embedding function. In our experiments, where we employ BERT, Euclidean distance is used.

To find the similarity of a sample i to its corresponding cluster C_k, we can assign each vertex i in the cluster a non-negative value z_i, representing the participation of the vertex in the cluster. The larger the z_i the more corresponding node is associated with the cluster. If $z_i = 0$, then the sample i is not associated with the cluster. This is a common way to represent the nodes in a cluster in classical graph-theoretical approaches [130, 124, 112] and a central quantity in dominant set identification [109]. The participation value $z_i, \forall i \in V_k$ can be found by solving the following optimization problem:

$$\begin{aligned} \min \quad & z^{\intercal} W_K z \\ \text{subject to} \quad & z_i \geq 0 \\ & \sum_i z_i = 1, \end{aligned} \tag{4.1}$$

where z is a vector containing $z_1, z_2, \ldots$ and W_k is a matrix containing the pairwise distances $w_k(i, j)$. Similar to [109], we can use the replicator dynamics optimization technique, which is an evolutionary game theory approach, to extract z_i associated with each node in the cluster. Using this optimization algorithm, the similarity of vertex i with respect to the cluster k is obtained. We follow the procedure in [109] that suggests to divide the samples into dominant and non-dominant

samples using the median of the obtained positive similarity scores of each cluster as a cutoff for the sample-cluster similarity (parameter z_i). This means that half of the points with non-zero participation are considered as the dominant set.

Using this pairwise clustering and dominant set extraction method, it is expected that the identified dominant sets represent the highly compact structures within the embedding space, and thus, belong to the same class. We believe that using the embedding space of a pre-trained language model such as BERT is crucial to obtain clusters with a low amount of noise. With these assumptions, the non-dominant sets within each cluster C_k contain more interesting samples for active learning as they hold less similarity to these maximally cohesive clusters. Accordingly, NDS concentrates on these sets from the pool data. To maximize the diversity over the embedding space and therefore the classes, we uniformly sample an equal number of data points from the non-dominant set of each cluster. Algorithm 1 shows the steps of the proposed NDS algorithm.

Algorithm 1 Selection Using NDS

Require: A pool of unlabeled data points $\mathscr{D}_{pool}$, number of samples to be selected m, number of classes K, an embedding function, and a distance function.

Output: Selected samples for annotation.

Embedding: Extract the feature vectors corresponding to all the points $x \in \mathscr{D}_{pool}$ using the embedding function.

Clustering: Extract the clusters $C_1, \ldots, C_K$ using spectral clustering

for $k = 1, \cdots, K$

    Calculate the pairwise distance matrix W_k

    Calculate the sample-cluster similarity z using (4.1)

    Calculate the threshold as $\tau = median(z[z > 0])$

    Randomly select $\frac{m}{K}$ samples with $z_i \leq \tau$

end

4.1.2.1 Incorporating Uncertainty into NDS

Although uncertainty-based methods such as variation ratio [39] and Bayesian AL [42] are widely used for active learning, these sampling strategies are known to have a higher tendency to the

selection of outlier samples in the early cycles [33]. As the size of training data increases in the later active learning cycles, the uncertainty-based methods are able to provide more reliable uncertainty scores.

On the contrary, we show that NDS is an effective method in the early sampling cycles as it selects the most critical samples from the pool. After a few active learning cycles, and specially, in the case of extremely imbalanced datasets, NDS can run out of the non-dominant set pool to select from. This is because the previously sampled training data is removed from the pool in every iteration, which results in the shrinkage of clusters. In such a scenario, we can increase the size of the non-dominant set by increasing the threshold of the dominant-set detection. As the model has learned more challenging examples at this point, the drawback of modifying this threshold can be the selection of redundant examples.

We can also extend our approach via proposing NDS+, which is a compound sampling strategy that benefits from both NDS and uncertainty-based methods. We can think of NDS as a random selection with equal weights over the non-dominant sets while setting the weight of dominant samples to 0. Considering a uniform distribution over each of the identified per cluster non-dominant sets, the linear combination of NDS and uncertainty-based approaches becomes possible. To investigate the impact of such compound strategy on the classification performance, we simply define NDS+ with a smooth transition such that only NDS is used in the very early cycle, and the uncertainty scores influence the drawing procedure of samples in the later stages. Let $\Phi_U \in [0,1]$ be an uncertainty measure, e.g. minimum margin, the hybrid sampling weight for NDS+ can be defined as:

$$\alpha\Phi_{NDS} + (1-\alpha)\Phi_U,$$

where Φ_{NDS} is the NDS sampling weights as defined above and α is a parameter that regulates the relative effect of NDS selection versus the uncertainty-based strategy. The value of parameter α

initially starts from 1, the impact of uncertainty score can be gradually increased by reducing the parameter α over the cycles of active learning.

The evaluation of the proposed technique, choice of parameters, and details on the dataset are provided next.

4.2 Experiments and Results

4.2.1 Dataset Description

To evaluate the effectiveness of the proposed approach, we conduct our analysis using two datasets that we will briefly describe in this section. We refer the reader to the provided source citations for further details.

I) The abusive language Twitter dataset [38] contains the *abusive, hateful, spam*, and *normal* classes. The class sizes are 22,766, 4,496, 13,996, and 53,560, respectively. This dataset has been referred to as the Twitter-abusive dataset in our study.

II) The second dataset that we use in this work is the *Wikipedia Talk Labels: Personal Attacks*[152], which belongs to the Wikipedia Detox Research Project. In this study, we refer to this dataset as Wiki-attack. This dataset contains comments from Wikipedia talk pages, and is annotated for binary classification of *attack* versus *normal* classes. The class sizes are 10,792 and 61,174, respectively.

To prepare the model inputs for both datasets, first, we follow the text normalization procedure in [102]. We remove the emojis from the text corpus. Then, we replace the URLs with the special token “HTTPURL". For the Twitter dataset, we also replace the usernames with the special token

“@USER". For the Wiki-attack text data, we remove the special tokens “TAB_TOKEN" and “NEWLINE_TOKEN". The data preparation procedure for the Wiki-attack dataset is the same as in [151]. Next, we use the BERT tokenizer to convert the text inputs to sequences of tokens. The train and test sets for the Twitter-abusive dataset are determined via randomly splitting this dataset with the ratio of 8:2. We used the predefined train and test splits for the Wiki-attack dataset.

4.2.2 Performance Measurements

4.2.2.1 Precision, Recall, F1 Score

We evaluate our model based on the achieved F1 score, which itself is a function of precision and recall scores. The precision score is defined as:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4.2}$$

The recall value is obtainable by:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.3}$$

And finally, the F1 score can be calculated as:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{4.4}$$

4.2.3 Acquisition Functions

Here, we briefly review some of the most common acquisition functions in active learning literature. This work is interested in the pool-based active learning [89], in which an acquisition function is used to query the label of a small set of selected samples. As discussed earlier, the model is initially trained using a small set of labeled data. Then, according to an informativeness criterion, the acquisition function selects a few data points to query their labels from the oracle. The most informative samples are drawn from the pool and added to the training set. We use various acquisition functions for comparison with our proposed strategy, which are reviewed below:

I. Random Acquisition (baseline): This function selects data points uniformly at random from the pool of the unlabeled data.

II. Bayesian AL (Monte-Carlo Dropout): Proposed in [41], this method is an uncertainty-based AL strategy in which the class probability for each sample is approximated via calculating the average over N inference cycles using Monte Carlo Dropout.

$$\textit{Bayesian-AL} = \frac{1}{N}\sum_{n} p(y = c|x, \theta_n)$$

The authors in [41] use Variation Ratios [39] as the acquisition function for their Bayesian AL strategy, defined as:

$$\textit{VarRatio}(x) = 1 - \max_{y} \; p(y = c|x, \theta),$$

in which larger values indicate a higher uncertainty score. Intuitively, the less probable our most probable class is, the more uncertain we are about the class of the sample.

III. Minimum Margin: Proposed in [125], this approach is also an uncertainty-based acquisition

function, which is more appropriate for a multiclass scenario. This function evaluates the uncertainty as the difference between the two most probable predictions as:

$$MinMargin(x) = p(y = c_1|x, \theta) - p(y = c_2|x, \theta),$$

where classes c_1 and c_2 have the first and the second highest prediction scores.

IV. Core-Set: Proposed in [128], this method selects central samples using a greedy algorithm by minimizing the distance between data points and their nearest centers to lead to covering the entire learned feature space.

4.2.4 Model Configuration and Training Details

To map the inputs to a feature space, we use the BERT-base architecture, which is known to capture the global as well as local context of text data. For a fair comparison with the state-of-the-art models, we use the IBM's low resource text classification toolkit[1] and add our approach to it as a new acquisition function. We use two different BERT-base architectures to conduct the experiments for text classification: i) the pre-trained BERT-base model 4.2, and ii) the pre-trained BERT-base model with 3 additional self-attention layers followed by a GRU layer 4.3. In the rest of this chapter, we refer to these architectures as BERT and BERT-GRU, respectively. The implementation details of BERT-GRU follow the setting used in [2]. The additional self-attention layers for this architecture contain 8 heads, and the GRU layer outputs a 512-dimensional feature vector. We investigate this architecture to examine the effect of extra attention layers and a larger number of parameters for different AL methods. A single dense layer as the classification head has been used for both architectures.

[1] https://github.com/IBM/low-resource-text-classification-framework

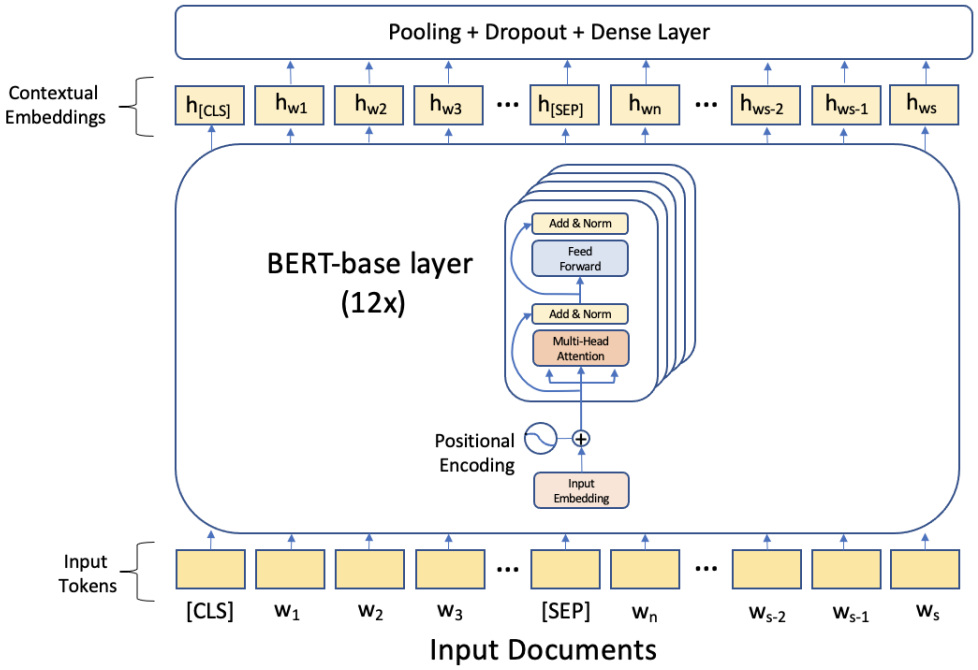


Figure 4.2: The overall architecture of a text classifier with a BERT-base model (number of layers L=12, hidden size H=768, number of self-attention layers A=12, number of parameters=110M) to extract the contextual representations of the tokens for the task of text classification. This figure shows when all final hidden states of tokens are used for classification instead of only the final representation of the [CLS] token.

Using the feature vectors from the last hidden layer of the BERT-base model, each text input x is converted to a word embeddings $F \in \mathbb{R}^{S \times d}$, where $d = 768$ refers to the dimension of the embedding space, and S is the maximal sequence length of the text inputs. Similar to [33], we use $S = 100$, and learning rate of 2×10^{-5}. We use draw size of 30, and batch size of 64 across all the experiments. However, due to memory constraints for the larger model we use $S = 50$. The same setting is used to train the two architectures using both datasets, exempting the number of training epochs, and initial training size. We choose to run all the experiments using the draw size of 30 as we noticed that a smaller draw size than this value can lead to high standard deviations and challenge the comparisons of the AL methods. The models were trained for $e \in \{5, 10\}$ number of epochs. However, we observed a reduced performance in terms of F1 score and recall of all active

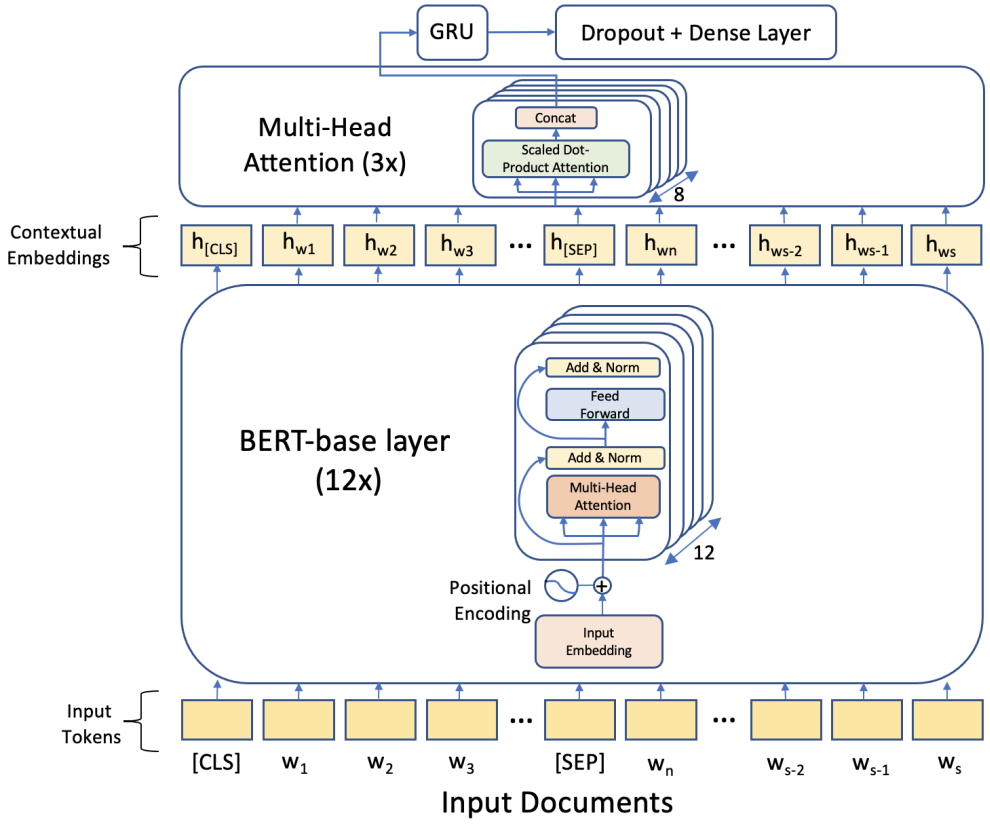


Figure 4.3: The overall architecture of a text classifier with a BERT-base model and 3 additional self-attention layers (each layer having 8 heads) followed by a GRU layer, and the classification head.

learning methods for the the BERT architecture when using 10 training epochs. This was specially noticed when the Wiki-attack dataset was used for training. On the contrary, higher performance scores were achieved for BERT-GRU using 10 training epochs. Thus, we report the experimental results using different epoch sizes for the two architectures.

We follow the imbalanced-practical setting in [33] for the initial training data preparation, which is suggested to avoid unstable BERT runs for extremely imbalanced scenarios, i.e., when the size of one class is less than %15 of the others. In this setting, the annotation budget is 100, performed on 100 random samples. Additionally, a simple (imperfect) heuristic, i.e., a query based on a frequent

pattern observed in the data, is used to retrieve examples from the class with the smallest size. Also, domain knowledge can be used to prepare such queries. A random sample of 100 retrieved examples, which are expected to belong to the class with the smallest size are then added to the training data without the need for annotation. Similar to [33] we use the query [[A-Z]!]] that refers to containing an upper case word that follows by an exclamation mark (e.g., IDIOT!). The training sets of each dataset are used as the initial pool for data selection and label querying.

4.2.5 Active Learning Details

After the initial training iteration, we use the model from the last training epoch for the evaluation with the test set, as well as the calculation of informativeness metrics by the acquisition functions. To evaluate the performance of the proposed approach, we perform comparative analysis using different active learning strategies, details of which are provided in the 4.2.3 section. The minimum margin uncertainty scores are calculated via conducting one forward pass. However, the Bayesian AL uncertainty scores are obtained by averaging over 10 forward passes with an additional dropout unit with parameter 0.2, added before the fully-connected layer, as proposed in [41]. Lastly, NDS and NDS+ use the learnt embedding space of the BERT model to discover the non-dominant sets associated with each cluster. We use the published code by [33] with the default values for the greedy Core-Set method.

As mentioned before, NDS and NDS+ methods use a cutoff parameter, which determines the dominant versus non-dominant sets of each cluster. This can become problematic when the size of identified non-dominant sets per associated class goes below $\frac{m}{k}$. We experienced this issue specially with extremely imbalanced data. Thus, we consider an adaptive cutoff value by increasing this parameter whenever the size of non-dominant set pools is less than $\frac{m}{k}$. In our experiments, we multiply the cutoff value by 10 until the size of the non-dominant set per cluster is sufficient for

sampling. An observation for NDS+ was that incorporation of the uncertainty scores postponed increasing the cutoff value to later AL cycles, and in some runs even prevented it.

After the calculation of the informativeness scores for the pool of training samples, m number of training inputs are drawn randomly from the pool of interest for each strategy. We use the minimum margin uncertainty scores for NDS+, and initially use $\alpha = 1$. This parameter is gradually decreased by 2% at each active learning cycles. We did not tune this parameter. Instead, this value is selected based on the number of AL cycles in our experiments, such that the draw of examples in the last cycle get influenced almost equally by NDS and uncertainty-based strategies. Next, the drawn samples are added to the training set, the models are reset, and the new set of training inputs are used to train a new model. The explained procedure continues until a total of 500 training samples are selected and used for the training of the models. We conduct 10 runs for each experiment setting, and report the average of the performance scores per active learning strategy (Figure 4.4).

4.2.6 Results

In this section, we report the results of our proposed AL methods, NDS and NDS+, in comparison to random selection, minimum margin [125], Bayesian AL with variation ratio [42], and Core-Set [128] acquisition functions using BERT and BERT-GRU architectures.

4.2.6.1 Does random selection from the non-coherent structures of the embeddings space of a deep language model for active learning improve performance over other methods? What is the impact of increasing the number of parameters by choosing a larger model?

To answer this question, we investigate the classification F1 scores of the proposed approaches, NDS and NDS+, versus state-of-the-art active learning methods. The reported scores in Figure 4.4

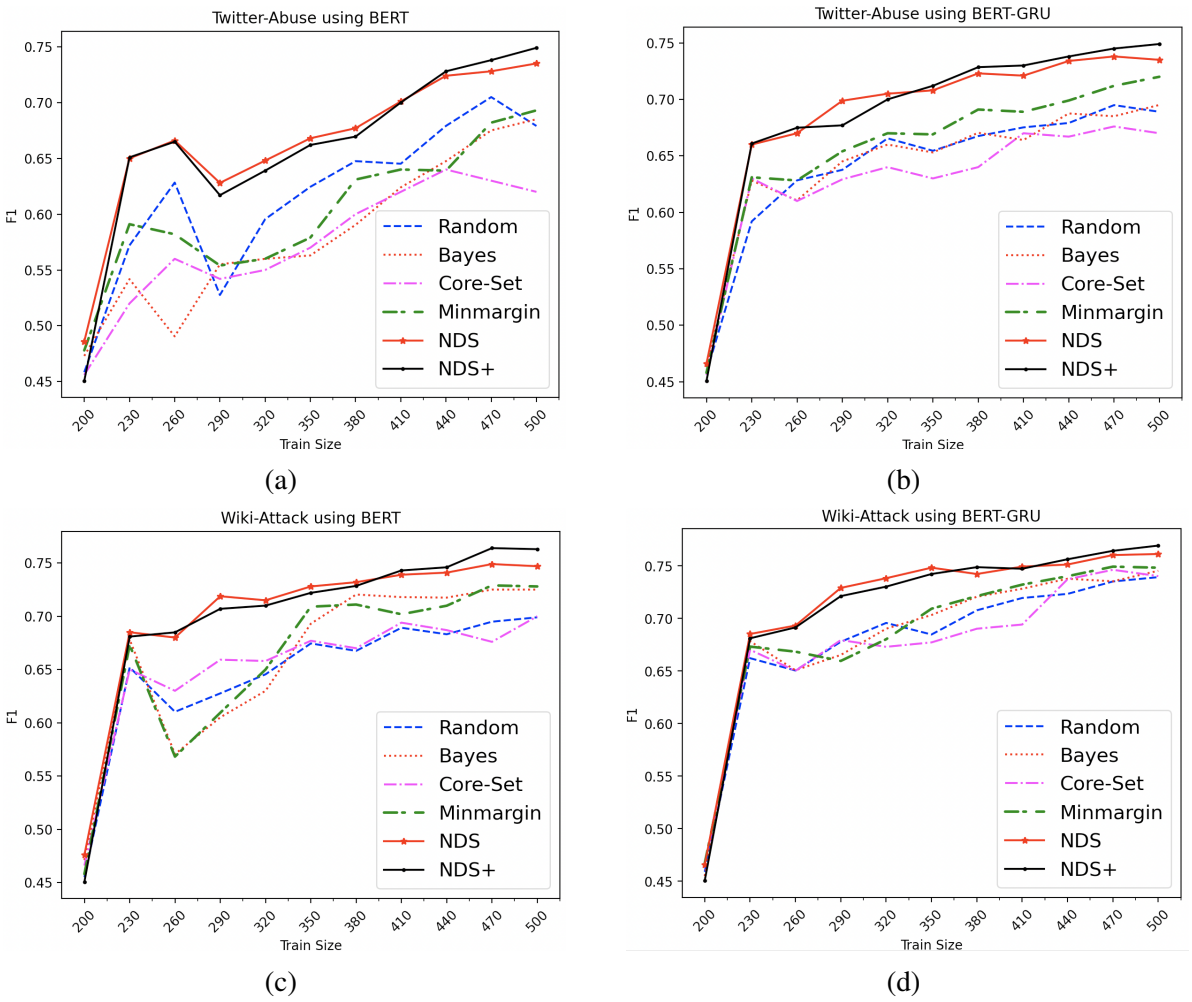


Figure 4.4: (a) and (b) are the average classification F1 scores for the Twitter-Abusive dataset using BERT and BERT-GRU, respectively. Similarly, (c) and (d) are the average F1 scores for the Wiki-attack dataset using BERT and BERT-GRU, respectively.

are the average of the classification F1 scores over 10 training runs using 2 different architectures and 2 different datasets. (a) and (b) report the results using *abusive language Twitter dataset* [38], and (c) and (d) are for the Wikipedia talk labels: personal attacks dataset [152]. From the figures, it is evident that the two proposed sampling strategies that concentrate on the dominant sets consistently outperform other AL methods.

It can be observed in Figure 4.4 that the uncertainty-based AL methods, Bayesian AL and minimum margin, fall behind the random and dominant set-based strategies in the first few AL cycles, observed for both architectures. This is in contrast with the results and analysis of uncertainty-based methods reported in [33], which can be due to selecting a larger number of samples per AL cycle in that paper. However, the results related to the Core-Set strategy is similar to the reported imbalanced-practical scenario in [33]. An observation that is consistent with literature is that as the size of training data grows, all methods manage to produce more reliable scores, and thus, discover more critical samples for training. Thus, a superior technique is the one that performs better in the early cycles. The performance of the core-set approach, which is a diversity-based method, does show the same drop as for the uncertainty-based methods in the early AL cycles, which is also reported in [33]. This technique has reported to perform well for the convolutional neural network architectures [128].

On the contrary, the two non-dominant set-based methods, NDS and NDS+, seem to select the critical samples for training, even in the early AL cycles. This difference is more significant when comparing different AL methods to fine-tune the model with smaller number of parameters, as shown for the BERT architecture (Figure 4.4. (a) and (c)). Rather, the three additional self-attention layers, as well as the GRU layer in BERT-GRU lead to compensation for the shorter input size and the deficiency of the selection strategies, with the cost of higher computational complexity ((Figure 4.4. (b) and (d))). Further comparison of the obtained results for BERT versus BERT-GRU also suggests that the additional layers in BERT-GRU result in more stability and less fluctuations that can be interpreted as having a higher robustness to noise. specially, the performance of all strategies in the early AL cycles is improved using this architecture. This makes the comparison of different methods more challenging.

Another important factor to discuss about the non-dominant set-based strategies is that the sizes of non-dominant sets shrink over AL cycles with this technique, resulting in a smaller pool of

samples for selection in the next iterations. This is specially a concern when using an extremely imbalanced dataset. Accordingly, the compound strategy that exploits an independent score, such as model uncertainty scores, can achieve superior results over the long run. This is noticeable from the obtained results reported for NDS+ in Figure 4.4. Specially, our results represent the superiority of this method for later AL cycles as expected. However, the results in Figures 4.4. (a)-(d) also suggest competing performance for the two dominant set-based methods in early and middle AL cycles. In later cycles, NDS+ seems to take advantage from both NDS and uncertainty-based selection methods, observed consistently over 10 runs and across the experiments. However, we did not investigate the samples that are being selected using each of these strategies in this hybrid method.

Acknowledging the sharp decline in the F1 scores associated with the uncertainty-based and Core-Set (diversity-based) strategies in the early AL cycles, the significance of using non-dominant sets for the selection pool becomes evident. Our extensive analysis and comparison of the proposed methods with other AL approaches using two different text classification corpus, and two deep architectures suggest that randomly selecting from the non-dominant set associated with each cluster is a powerful strategy to reduce the labeling cost and effort. That being said, further improvements can be achieved via considering a joint strategy, as reported for NDS+.

4.2.6.2 How does the length of data inputs and padding affect the performance of different active learning methods?

To answer this question, we need to look at this problem from different aspects. First, the datasets that are studied in this research have short text. Even in many related research on using active learning for language processing such as [33] and [159] that experiment on long texts, the sequence length is set as $s = 100$ and $s = 128$, respectively. This is partially because of memory constraints.

This constraints results in choosing a smaller batch size if the sequence length is high, and thus, a slower convergence. Second, when the input sequences have short length and we set a high maximum sequence length s, we are increasing the padding size. Although we are masking the padded sequences and the attention mechanism ignores the padded tokens, many research have found that excessive padding deteriorates the performance of deep models and slows convergence [154], including for transformer architectures [162]. To overcome this problem, many research papers apply smart batching, which groups inputs according to their length and pads them to the maximum length in the mini-batch [121]. This method has achieved a substantial convergence speed-up on both CPUs and GPUs. The work of [121] reports this speed-up as 89% and 48% for CPUs and GPUs, respectively.

In this doctoral thesis, we approach this problem by using the BERT architecture discussed earlier, and evaluating the same active learning strategies considering the same parameters as in the previous experiment, except for the length of the input sequences (s). We repeat the experiments for new sequence lengths of $s = \{64, 128\}$ and compare the performance scores with the $s = 100$ reported earlier. Again, we report the classification F1 scores of different active learning strategies. The results can be found in Figure 4.5, which are the average of the classification F1 scores over 10 training runs using the BERT architecture evaluated on 2 different datasets. (a) and (b) report the results using *abusive language Twitter dataset* [38], and (c) and (d) refers to the results for the Wikipedia talk labels: personal attacks dataset [152] dataset. From the figures, it is evident that using $s = 64$ significantly worsens the results for all active learning strategies and challenges comparing different methods for both datasets. Using such a short sequence length leads to losing a substantial amount of information that the deep model requires to learn, and thus, the classification score decreases. Not learning the class features leads to model instability and high uncertainty [33], which affects the performance of all data selection strategies. We do observe noticeable improvement for different active learning methods in the early cycles when setting the sequence max

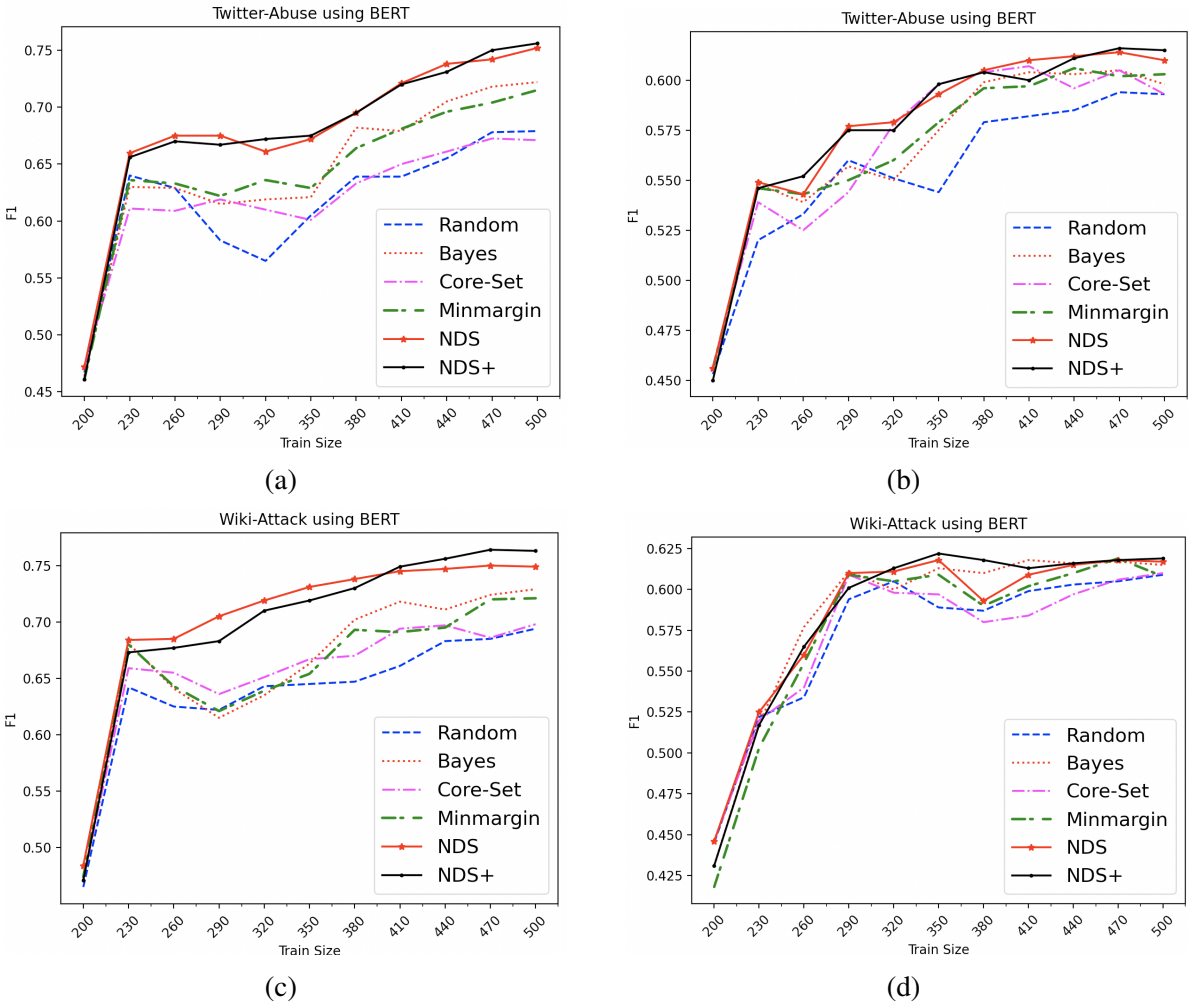


Figure 4.5: (a) and (b) are the average classification F1 scores for the Twitter-Abusive dataset using BERT architecture with maximum sequence length of $s = 128$ and $s = 64$, respectively. Similarly, (c) and (d) are the average F1 scores for the Wiki-attack dataset using BERT architecture with the maximum sequence lengths of $s = 128$ and $s = 64$, respectively.

length as $s = 128$ vs 100. This observation is specially evident for the Wiki-Attacj dataset that contains lengthier data versus Twitter-Abuse. But, the results in the later active learning cycles for both detests seem to be similar to those when the sequence length is set as 100. Investigating higher values of sequence length is not possible in this research due to memory constraints.

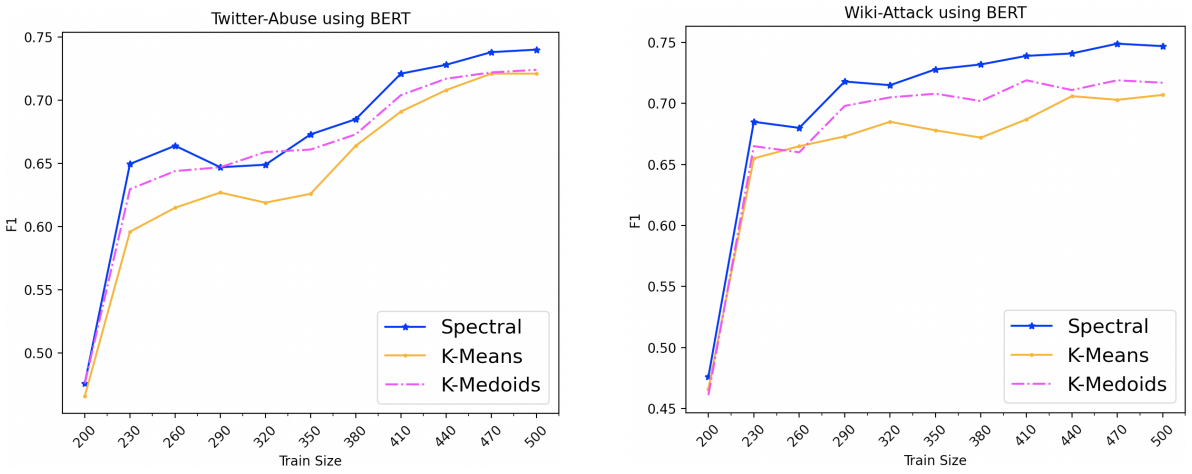
4.2.6.3 *What is the impact of conducting initial spectral clustering versus other clustering techniques for non-dominant set-based active learning?*

To answer this question, we investigate the performance of NDS using 3 different clustering methods: spectral clustering, K-means, and K-medoids. After these 3 clustering methods, the non-dominant set of each cluster are extracted in the same way as discussed earlier. We conduct experiments using all the parameters the same as the first experiment and repeat the evaluation of NDS using spectral clustering vs K-means and K-medoids.

The results of this experiment can be found in Figure 4.6, which are the average of the classification F1 scores over 10 training runs using the BERT architecture evaluated on 2 different datasets. (a) reports the results using *abusive language Twitter dataset* [38], and (b) refers to the results for the Wikipedia talk labels: personal attacks dataset [152] dataset. From theses figures, it is evident that using spectral clustering before non-dominant set extraction for active learning outperforms the other clustering techniques. This can be because spectral clustering has been frequently reported as more powerful than K-means in the discovery of the clusters of arbitrary shapes, and is are more robust to noise and outliers [67, 74].

4.3 Conclusion

The task of labeling offensive and abusive content is difficult, as it can cause discomfort and emotional disturbance in the human annotators. Thus, we focused on a new criterion for active learning to select the most informative samples from a pool of unlabeled data points. Our proposed approach has the potential to mitigate the difficulties associated with the annotation, and classification of textual content, e.g., annotation cost and bias, even in imbalanced scenarios. We showed the effectiveness of our approach via conducting extensive analysis on text classification of toxic



(a) (b)

Figure 4.6: (a) and (b) are the average classification F1 scores using the BERT architecture and show the impact of initial clustering algorithm, evaluated on the Twitter-Abusive and Wiki-Attack datasets, respectively.

language and hate speech in online social media data, using only a small amount of labeled data. These datasets are extremely imbalanced by nature and have short text, and thus, many common techniques fail to perform well in such scenarios. We leverage pretrained language models and unsupervised techniques to detect the local clusters in the embedding space and select the samples that are not strongly coherent with the cluster, i.e., the non-dominant set. Specially, our method significantly outperforms the state-of-the-art AL techniques in the early AL cycles in which the number of annotated samples are limited.

We also propose a hybrid algorithm that is able to incorporate the uncertainty score into its decision criteria. The results and analysis using this technique suggest that an active learning task can be divided into two distinct phases. In the early stages, unsupervised techniques such as employing pre-trained models, clustering, and identifying the dominant sets outperform the supervised techniques, e.g., uncertainty score extracted from the trained model. However, in the second phase,

later stages of selection, taking the model uncertainty into the account can improve the selection performance. Such hybrid methods are not currently well-studied. The proposed method can be considered parameter-free, as we did not fine-tune for any values, used the default or suggested settings in the literature, and all of the parameters for the architectures were shared across all methods.

We also showed that increasing the size of the model and training a larger number of parameters on a downstream task with active learning significantly improves the performance of different active learning strategies. We further showed that for short text such as social media textual data, choosing a larger sequence length and increasing the padding size for the training of a deep language model, such as BERT, does not have a significant impact on the performance. This observation was consistent for different active learning strategies. However, a very small sequence length can result in losing critical information that is required for the training of a model. As a result, it is vital to set this parameter wisely.

This doctoral thesis also reported the results of using different clustering techniques before the extraction of dominant and non-dominant sets for active learning data selection. We showed that spectral clustering followed by non-dominant set extraction achieves the best results for text classification. Still, additional experiments are needed on the effect of data imbalance, the number of clusters, size of transition parameter (α) in NDS+, and using different feature similarity scores instead of the Euclidean distance, such as the Mahalanobis distance.

CHAPTER 5: APPLICATION OF CURRICULUM LEARNING INTO TEXT CLASSIFICATION OF ONLINE SOCIAL MEDIA DATA: BENEFITS AND RESTRICTIONS

The introduction of Curriculum Learning (CL) to the field of machine learning was first suggested in [7], which shows performance improvements. The results showed that the exclusion of difficult samples and noisy data in the early training stage is beneficial, such as faster convergence and achieving better local minima. Due to the difference in the difficulty levels of the examples from any dataset, extensive research has investigated ways to identify the easy samples from the difficult ones and to arrange them as a curriculum for the training of a model [7]. Some literature in this area of research suggest that the learning process can achieve remarkable improvements when using a curriculum [150]. However, relatively little attention is devoted to this topic in the area of natural language processing. Examples of difficulty in language are lengthy text, short text that lacks context, rare words, and sophisticated reasoning such as negation.

As we are interested in the application of curriculum learning for the classification of social media textual data, we must pay attention to the problems related to this field, including annotated data limitation and data imbalance. In the previous chapter, we addressed these issues by proposing an active learning method that outperforms state-of-the-art active learning techniques for the task of social media hate speech classification. In this chapter, we further investigate applying similar unsupervised techniques, but this time as part of designing a curricula to fine-tune a pre-trained language model. We investigate the impact including benefits and drawbacks of different curriculum learning methods for the classification of online social media data.

The problem of data imbalance for curriculum learning can be tackled by oversampling from the minority class, downsampling from the majority class, and adjusting weights for the loss function.

It must be noted that in this context, downsampling of the training data refers to selecting a small subset of the majority class examples.

Based on our findings from Chapter 4 we assume that deep pre-trained language models, such as the BERT model [32], can be used as a powerful mapping function for the short-text social media data to achieve high-dimensional contextual vector representations of such data. Our last assumption is regarding the dominant and non-dominant sets [109], which is also from the findings of the previous chapter. We assume that the non-dominant set of each cluster contains the more challenging samples, and the dominant set of each cluster contains the easy samples for the language model as they are most likely located within the decision boundaries of the classifier. We exploit this idea to design our curricula and provide comparisons with other methods.

5.1 Our Framework

Here, we explain our proposed curriculum learning method. The framework comprises of 2 main steps: I) extracting feature vectors that are rich with contextual information and general language knowledge using deep pre-trained language models as the embedding function; and II) identifying the dominant and non-dominant samples of clusters, and sorting based on their difficulty to fine-tune the language model. Briefly, this is performed via randomly selecting from the dominant samples and shift the difficulty to the selection of non-dominant samples.

Our CL approach uses BERT [32], which is the most influential pre-trained language model that achieved state-of-the-art results on a wide range of tasks related to natural language processing. The output of the model, which is being used for sequence classification, is the pooled representation of all contextual vector embeddings for each token position.

For a target task, let $\mathscr{D}_{pool}$ be the pool of examples for training, and $\mathscr{M}$ be the language model

that is being trained to fit $\mathscr{D}_{pool}$. Here, we explain how we assign a difficulty score d_i to every example $i \in \mathscr{D}_{pool}$. We denote d as the difficulty level corresponding to training data. In the second stage, based on d the pool data $\mathscr{D}_{pool}$ is organized into a sequence of N ordered learning stages $\{S_{d_1}, ..., S_{d_N}\}$ with an easy-to-difficult order, resulting in the final curriculum where the model will be trained on.

5.1.1 Difficulty Evaluation

Although applying heuristics such as the length of textual data, frequency of rare words, and depth of dependency tree seem reasonable for data difficulty analysis, these techniques might not be generalizable to different tasks. Also, what human judgment might find difficult might not necessarily be challenging for a deep language model. As a result, we believe that the difficulty score of a sample as its intrinsic property should be decided by the model itself. Examples of such measurements are accuracy, F1 score, and different types of model uncertainty.

In chapter 4 we compared different acquisition functions for data selection in active learning. As the objective in active learning is minimizing the annotation effort, we decided to ignore the dominant sets of each cluster and assumed that they share the same class labels as their surrounding data samples that lie farther from the dense parts of each cluster, i.e., non-dominant sets. This assumption may not be necessarily valid and may hurt the performance of a model, specially, when we have the annotation budget or have access to the labels, which is the case in curriculum learning. Even if this assumption stands true, disregarding dominant sets from the training data may lead to losing a substantial amount of information that might be required for the model to learn the class features. Thus, we design our curriculum strategy with the same assumptions that dominant sets of each cluster are easier samples to learn by the model and non-dominant sets are more challenging. We compare this curriculum with the methods discussed earlier.

To evaluate the difficulty of the samples, we first, apply a clustering method. Then, we need to identify the local clusters of data in feature space and for each cluster detect the most similar set of points, i.e., the dominant set. To find the similarity of a sample i to its corresponding cluster C_k, we can assign each vertex i in the cluster a non-negative value z_i, representing the participation of the vertex in the cluster. The larger the z_i the more corresponding node is associated with the cluster. If $z_i = 0$, then the sample i is not associated with the cluster. This is a common way to represent the nodes in a cluster in classical graph-theoretical approaches [130, 124, 112] and a central quantity in dominant set identification [109]. The participation value $z_i, \forall i \in V_k$ can be found by solving the following optimization problem:

$$\begin{aligned} \min \quad & z^{\intercal} W_K z \\ \text{subject to} \quad & z_i \geq 0 \\ & \sum_i z_i = 1, \end{aligned} \tag{5.1}$$

where z is a vector containing $z_1, z_2, \ldots$ and W_k is a matrix containing the pairwise distances $w_k(i, j)$. Similar to [109], we can use the replicator dynamics optimization technique, which is an evolutionary game theory approach, to extract z_i associated with each node in the cluster. Using this optimization algorithm, the similarity of vertex i with respect to the cluster k is obtained. Although in the previous chapter we suggested to divide the samples into dominant and non-dominant samples using the median of the obtained positive similarity scores of each cluster as a cutoff for the sample-cluster similarity (parameter z_i), here we choose a smaller size for this parameter and gradually increase it over the epochs. Although this parameter needs to be tuned, we do not investigate this in our study and will leave it for future work. Instead, we heuristically choose the 25th percentile for the first curriculum learning step and increase this parameter with the size of 25 percent in each next steps. The 100th percentile means that all cluster data is now considered as the cluster dominant set. This leads to having four difficulty levels for curriculum learning.

The clustering is performed once. However, the identification of dominant and non-dominant

sets are being repeated each time we modify the cutoff parameter z_i. According to the size of parameter z_i in each curriculum learning step, the dominant samples are identified. The samples that do not belong to the dominant set of any of the clusters are the least similar samples to their corresponding clusters, and therefore are more challenging to classify. We use the number of classes in our classification task as a known parameter for an initial clustering before finding the dominant sets in each cluster. We assume that with little domain knowledge the number of classes must be known. The differences of our curriculum learning with the NDS active learning method discussed in the previous chapter are as follows:

I. Instead of spectral clustering, here we use K-medoids clustering;

II. Instead of setting the dominant/non-dominant set cutoff parameter z_i as the median of the obtained positive similarity scores of each cluster, we choose 25th percentile at first and gradually increase it;

III. Every time we change the cutoff parameter for the dominant and non-dominant set identification, we repeat this step. Then, the training data is selected from the new identified dominant set and the model is retrained;

IV. The data selection strategy is opposite to what we discussed for NDS (section 4.1.2), as we must select the least challenging data first and increase the difficulty via adding training samples from the non-dominant sets.

We choose to use K-medoids despite the fact that we previously observed lower performance for K-means and K-medoids for our active learning method when compared to spectral clustering. K-medoids and K-means are less powerful in the discovery of the clusters of arbitrary shapes and when outliers and noise exist. Yet, in chapter 4, compared these methods with spectral clustering and observed that the performance of these methods can improve under specific conditions. In

NDS, we are dealing with the most dense part of the clusters and K-means and K-medoids have found to effectively identify dense cluster shapes, where all members of each cluster are in close proximity to each other (in the Euclidean sense). Also, in 4 we showed that the clustering technique matters the most in the early cycles of active learning when the training data size is small. In curriculum learning all the dataset is being used for training. Thus, the performance of different clustering techniques was observed to be much closer as the size of training data increases, we substitute spectral clustering with K-medoids to speedup our approach. In this chapter. All the other steps of identifying the dominant and non-dominant sets remain the same as section 4.1.2. Next, we sort the data based on the discussed difficulty evaluation method. The result is arranging the training data $\mathscr{D}_{pool}$ to a sequence of N different ordered learning difficulties d as $\{S_{d_i} : d = 1, ..., N\}$ having an easy-to-difficult order.

5.1.2 Gradual Training

We use the same architecture discussed in chapter 4, referred to as the BERT architecture for the experiments in this chapter. We train the model step-by-step based on different difficulty levels. In our experiments, we choose the 25th percentile for the first curriculum learning step and increase this parameter with the size of 25 percent in each next step. In the last step, all cluster data is considered to be added to the training data. This leads to having $N = 4$ difficulty levels for curriculum learning. In this study, we have not investigated different values for this parameter. This is decided based on similar values in related research. We first train the model for one epoch with the difficulty level of one. Then, in the next epoch we retrain the model using data with difficulty levels one and two. We continue this process until the training data include the first to the last difficulty levels.

5.2 Experiments and Results

5.2.1 Dataset Description

To evaluate the effectiveness of the proposed approach, we conduct our analysis using five datasets that we will briefly describe in this section. We refer the reader to the provided source citations for further details.

I) The abusive language Twitter dataset [38] contains the *abusive, hateful, spam*, and *normal* classes. The class sizes are 22,766, 4,496, 13,996, and 53,560, respectively. This dataset has been referred to as the Twitter-abusive dataset in our study.

II) The second dataset that we use in this work is the *Wikipedia Talk Labels: Personal Attacks*[152], which belongs to the Wikipedia Detox Research Project. In this study, we refer to this dataset as Wiki-attack. This dataset contains comments from Wikipedia talk pages, and is annotated for binary classification of *attack* versus *normal* classes. The class sizes are 10,792 and 61,174, respectively.

III) The third dataset is the Covid-CQ stance Twitter dataset [97] that contains the *favor, against*, and *neutral* labels as the stances toward the use of "chloroquine" and "hydroxychloroquine" for the treatment or prevention of Covid-19. The class sizes are 6,841, 4,685, and 2,848, respectively. In this study, we refer to this dataset as Covid-CQ.

IV) The fourth dataset is the IMDb subjectivity data [107] which contains 5000 *subjective* and 5000 *objective* review snippets from the IMDB website. This dataset does not include reviews shorter than 10 words.

V) The last dataset is the AG news corpus [163], which contains news articles on the web. This dataset has 127,600 samples and 4 classes of World, Sports, Business, and Sci/Tech

regarding categorized news articles from more than 2000 news sources.

The procedure of data preprocessing is the same for all datasets. To prepare the data, first, we follow the text normalization procedure in [102]. We remove the emojis from the text corpus. Then, we replace the URLs with the special token “HTTPURL". For both of the Twitter datasets, we also replace the usernames with the special token “@USER". For the Wiki-attack text data, we remove the special tokens “TAB_TOKEN" and “NEWLINE_TOKEN".

We use the BERT tokenizer to convert the text inputs to sequences of tokens. The train and test sets for the Twitter-abusive, Covid-CQ, and IMDb subjectivity datasets are determined via randomly splitting these datasets with the ratio of 8:2. We used the predefined train and test splits for the Wiki-attack and AG News datasets.

5.2.2 Commonly Used Curricula in NLP

We compare our method with the following commonly used curricula in the field of natural language processing:

1. Sequence Length: This method is simply a criterion based on the number of words in a sequence. We refer to this method in the experiments as SL. The training data for this criterion should be arranged as lengths from short to long;

2. Average Word Frequency Rank: This method ranks the words based on word frequencies across the dataset, and calculates the average rank for each input sequence. The order of training data difficulty, in this case, is sorting the sequences containing the most to the least frequent words [164, 77]. in this work, we refer to this method as WFR.

5.2.3 Model Configuration and Training Details

To map the inputs to a feature space, we use the BERT-base architecture, which is known to capture the global as well as local context of text data. A single dense layer as the classification head has been used. Using the feature vectors from the last hidden layer of the BERT-base model, each text input x is converted to a word embeddings $F \in \mathbb{R}^{S \times d}$, where $d = 768$ refers to the dimension of the embedding space, and S is the maximal sequence length of the text inputs. We use $S = 100$, the learning rate of 2×10^{-5}, and the batch size of 64.

5.2.4 Results

In this section, we report the results of our proposed curriculum learning method, in comparison to different strategies. The results are provided and discussed for each research question.

5.2.4.1 *Does applying a curriculum learning-based algorithm based on the local structures of the embedding space of a language model improve finetuning a deep language model for text classification? Does this technique outperform using simple heuristics for text difficulty?*

To answer this question, we conduct experiments using five different datasets introduced earlier and provide the results for difficulty criteria of sequence length and rare words in inputs in comparison to our suggested curriculum arrangement method. Below we discuss the investigated strategies:

1. Random using all data (Random_All): sequence of randomly ordered samples. This is the same as applying no curricula and instead, using all the training datasets for some number of epochs to fine-tune the language model. Here, we train the model for 4 epochs.

2. Random: sequence of randomly selected samples without a curriculum, but using the same size of training data at each step as in the curriculum learning method. Here, for $N = 4$ levels of difficulties, we randomly select 25% of the data for training and add an additional 25% in each next step to have $\{S_{d_1}, ..., S_{d_N}\}$.

3. Sequence Length (SL): ordering the training data from the shortest to the lengthiest input sequences. After sorting the data in this way, to have $N = 4$ levels of difficulty we select 25% of the data for training and add an additional 25% in each next step to have $\{S_{d_1}, ..., S_{d_N}\}$.

4. Word Frequency (WFR): ordering the training data from sequences with the highest to lowest average word frequency rank [164, 77]. After sorting the data in this way, to have $N = 4$ levels of difficulty we select 25% of the data for training and add an additional 25% in each next step to have $\{S_{d_1}, ..., S_{d_N}\}$.

5. Dominant Set (DS): according to the value of cutoff parameter z_i in equation 5.1, find the dominant set of each cluster and aggregate them. This is S_{d_i} with difficulty level d_i. Add the randomly ordered aggregated data to the sequence of training data $\{S\}$. Repeat this step for every value of cutoff parameter z_i and add the new data to the end of ordered training data.

We provide the classification F1 scores of the proposed technique for the mentioned strategies above. The reported scores are available in Figure 5.1, which shows the evaluation of three curricula versus two baselines. The figures are the average performance of different methods over 10 runs of curriculum learning.

The flat line shows the performance of the BERT-base model for 4 epochs using all the training data available to the model without a curriculum. The training phases are from 1 to 4 for all methods, in each, 25 percent of the sorted data samples are added to the training set. As a result, it is not fair to compare the Random-All case that refers to the performance of the model trained over 4 epochs

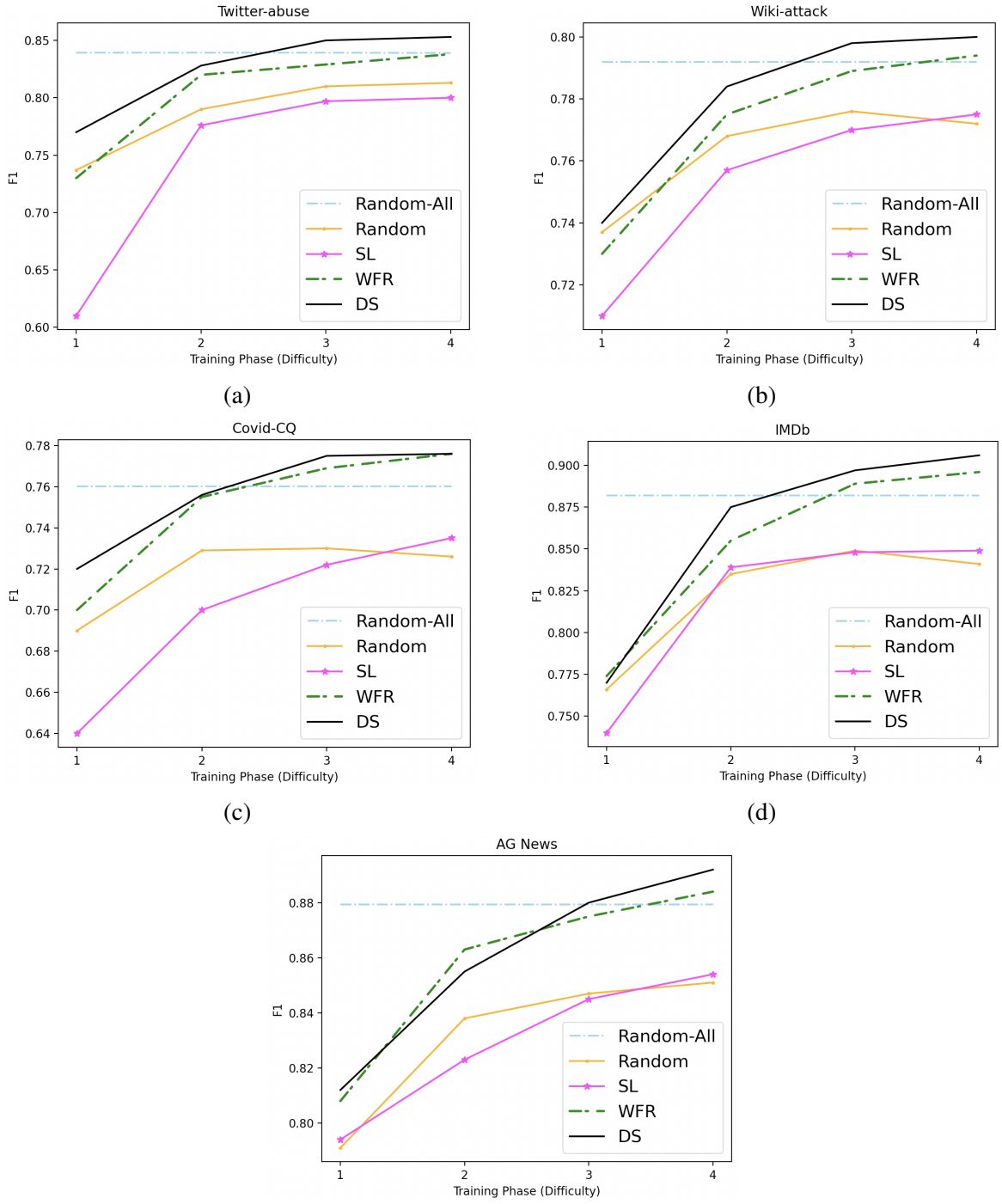


Figure 5.1: Average classification F1 scores for 10 runs of curriculum learning and increasing the difficulties within 4 phases to Fine-tuning the BERT-base model.

with the other techniques in the first 3 epochs of curriculum learning. Instead, the performance of Random-All can be compared with the scores achieved by other methods for the final difficulty level.

From the results, it is evident that not all curriculum learning methods outperform the Random case (Yellow line), in which 25 percent of data is randomly selected and added to the training set. Specifically, using the sequence length as the curriculum learning criterion performs worth than the Random case as well as the rest of the methods, and for all of the investigated datasets. The low performance of the Sequence Length (SL) strategy can be because all these investigated datasets contain short sequences of data. Also, the maximum sequence length for the BERT model is set as 100 tokens, meaning that the rest of the tokens in lengthier sequences are ignored. Additionally, using SL method to sort the difficulty of short text might not be practical as the shortest sequences may lack critical information and even be more challenging to the model. Instead, such a technique can be useful for other tasks such as question answering, in which a short answer can be easier to the model than a lengthy one.

Another important observation from all five figures is that the Random case consistently performs worth than the Random-All base case, which uses all the training data for all 4 training epochs. This was expected as the initial randomly selected training data (25 percent) may be biased toward a class or contain noise. Such data is being used for a single training epoch and thus, the model would not be able to learn critical features from such limited data. The result is a low performance that is probably associated with poor local minima. As the choice of data affects the learning trajectory of the model, the impact from the initially randomly selected data seems to remain even when the training data size increases and the model is retained on the new data. In contrast, it can be seen from all figures that having all the training data available to the model as the standard case (Random-All) can cancel the influence of poor data selection.

The WFR method and DS outperform the other techniques for all datasets. Still, WFR, which orders and selects input sequences according to the frequency of the words that appear in the sequence does not achieve a higher performance than the Random-All base case. Again, it must be noted that the BERT-base model is trained for 4 epochs on all available training data, which can lead to a more stable model. The WFR method shows having more significant improvement over other criteria for the Covid-CQ, IMDb, and AG News. This is not surprising as these three datasets contain vocabularies that are less frequent in ordinary language. On the contrary, less number of rare words were observed in the Twitter-abuse and Wiki-attack datasets. As a result, WFR, which is a heuristic difficulty evaluation method designed by human judgment does not seem to generalize across different datasets. But, WFR might be a more powerful strategy for curriculum learning when the dataset contains uncommon terminologies, such as in literary and scientific language. This technique specifically shows improvements over Random-All case for Covid-CQ, IMDb, and AG News datasets.

Our proposed curriculum method (DS) consistently beats the other methods for all investigated datasets. This is the case of selecting the data samples that lie closer to the center of clusters and gradually increasing the cutoff parameter that splits the data into dominant versus non-dominant samples of each cluster. As the difficulty of samples in this method is being assessed directly by the model instead of human judgment and assumptions, this method is more generalizable across different datasets. A notable observation is that our method substantially outperforms other criteria in the early phases of curriculum learning. This can be justified as the importance of data selection criterion when little data is being used for the training of a deep language model, as a better set of training data can lead the model to the direction of the optimal solution and speedup the convergence, which are the goal of curriculum learning.

5.3 Conclusion

In this chapter, we focused on the task of text classification using curriculum learning, which considers a difficulty criterion for training samples and trains the model using easy data first. We proposed a difficulty criterion that directly comes from the model instead of simple heuristics. Via conducting analysis on five different general and social media datasets, we showed that such a technique is generalizable across datasets and improves the classification performance of the model.

In our analysis, we observed that applying curriculum learning techniques in comparison to the standard case when all training data is available to the model does not necessarily improve the text classification performance when fine-tuning language models. Before the final phase of curriculum learning, the model is trained on partial data in each step. As a result, the choice of difficulty criterion is of great importance. The consequence of poor difficulty evaluation and using initial training data that includes uninformative, short, and noisy samples, is a performance decline in comparison to a standard training case. The cause of performance deterioration can be getting stuck in a local minimum, overfitting, etc. This can be specially observed in figures 5.1b, 5.1c, and 5.1d for the Random selection method. In these figures, the performance score drops in the later phases of curriculum learning despite the increase in the size of training data.

Although considering simple heuristics such as the length of sequence inputs, seems to be a reasonable difficulty method, these criteria may not be generalizable to different datasets and tasks. In our analysis, we showed that as social media textual data is intrinsically short, the application of such a technique does not provide benefits to text classification, but drawbacks. Despite testing our proposed method with imbalanced datasets such as Wiki-attack, Twitter-abuse, and Covid-CQ, we observed that our method is consistently superior to other techniques. This is also the case in experiments with balanced datasets.

Still, additional experiments are required to further investigate the impact of curriculum learning on text classification of short imbalanced data. Among the important research questions, investigating the correlations of selected data using different curricula can provide substantial information to understand the reasons behind the similarity and dissimilarity of performance scores for varying difficulty evaluation criteria. Also, studying the correlations of selected data for curriculum learning versus active learning in the early and late phases is important. Another future work is comparing the proposed dominant set-based difficulty criterion with other curriculum learning techniques, such as data difficulty arrangement based on the depth of sequence dependency tree and the sentence score from the GPT2 language model.

CHAPTER 6: CONCLUSION

This doctoral work is focused on the importance of designing low-recourse machine learning techniques for natural language processing, and specifically, for online social media textual content. Low-recourse approaches can be defined as techniques that do not require a substantial amount of labeled data or resources for the training of machine learning methods. Unsupervised learning methods do not require the label information of any data points, and accordingly, are one of the focuses of this doctoral work. However, as supervised techniques can provide lots of additional benefits, such as a higher reliability and performance, this work also proposes techniques that can be used with minimal annotation cost, as well as techniques to facilitate training of deep models.

An important factor that interests us for the research presented in this doctoral thesis is the development of deep pre-trained language models, such as the BERT (Bidirectional Encoder Representations from Transformers) architecture which is solely designed based on transformers in a stacked concatenated fashion referred to as multi-head self attention. Such method is a powerful technique that can be fine-tuned for a downstream task using small amount of labeled data and achieve the state-of-art results via knowledge sharing and attention mechanism. As a result, this doctoral work investigates ways to exploit the structure of BERT vector embeddings to design a data selection criterion for active learning and a data difficulty criterion for curriculum learning.

We presented the design of a label-efficient low-recourse technique for online social media by proposing a topic modeling framework that can extract topics around significant events, as well as providing topic summaries. We showed that our framework identifies narrative activities associated with social media events. Although conversation cascades created by commenting on an original post, re-sharing, or quoting, these activities are not necessarily related to the original post. Also, it is frequently observed that the information on the context and story are missing. We believe that if

the information on the causal relations of the social media events is known, further improvement of the narrative framework can be achieved via incorporating such relation across the social media conversation cascades and social media events into account.

We investigated the problem of social media data annotation and provided an active learning solution to select and annotate informative data. The objective is to achieve high performance using a small amount of labeled data. As social media corpus usually have short text and lack context, we used the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model as a mapping function to convert the text to high dimensional vector representations with general language knowledge and contextual information. In this study, we present a new criterion to select the most informative samples from a pool of unlabeled data points. Our proposed algorithm detects the local clusters in the embedding space and selects the samples that are not strongly coherent with the cluster, i.e., the non-dominant set. Additionally, we studied the impact of increasing the number of trainable parameters by adding additional attention layers and showed that different active learning methods achieve better performance when training a larger model comparing to BERT-base. We also propose a hybrid algorithm that is able to incorporate the uncertainty score into its decision criteria in the later stages of selection. Our results and analysis suggest that an active learning task can be divided into two distinct phases. In the early stages, unsupervised techniques such as employing pre-trained models, clustering, and identifying the dominant sets outperform the supervised methods, i.e., uncertainty score extracted from the trained model. However, in the second phase, later stages of selection, taking the model uncertainty into the account can improve the selection performance. Such hybrid methods are not currently well-studied.

Finally, in this thesis we investigated the benefits and drawbacks of finetuning pre-trained language models according to a curricula. Curriculum learning approaches training machine learning models with easy-to-difficult training data. This can be viewed as the opposite of active learning as the most informative data selection for active learning can be considered as selecting the most chal-

lenging data samples and ignoring the easy data. For this reason, active learning is also referred to as anti-curricula. Although these methods are the opposite of each other, still both can provide benefits due to having different objectives. In curriculum learning, the model is being retrained over and over on the easy samples, and in each step, more and more difficult samples are added to the training data. In our curriculum learning study, we employed the BERT model to enrich short-text social media data to high dimensional vector representations with general language knowledge and contextual information and to define a curricula based on the structures in the embedding space of the BERT model. Our proposed difficulty analysis criteria is a distance-based method and we consider the data samples that lie closest to the center of clusters as the easiest samples for the language model. Again, we used dominant sets and adjust a cutoff parameter that controls the participation of every data sample in such dense region of every cluster. Different difficulty levels in our method are available by gradually increasing this cutoff parameter. The results is having a criterion that comes directly from the model rather than human judgment and heuristics such as text length.

We believe that this doctoral thesis provides promising research and results for low-resource machine learning scenarios such as social media text analysis and classification.

LIST OF REFERENCES

[1] Gustavo Aguilar and Thamar Solorio. From english to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, 2020.

[2] Ramya Akula and Ivan Garibay. Interpretable multi-head self-attention architecture for sarcasm detection in social media. *Entropy*, 23(4):394, 2021.

[3] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

[4] Jeffery Ansah, Lin Liu, Wei Kang, Jixue Liu, and Jiuyong Li. Leveraging burst in twitter network communities for event detection. *World Wide Web*, pages 1–26, 2020.

[5] Regina Barzilay, Noemie Elhadad, and Kathleen R McKeown. Sentence ordering in multidocument summarization. In *Proceedings of the first international conference on Human language technology research*, pages 1–7. Association for Computational Linguistics, 2001.

[6] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[8] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[9] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, 2006.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[11] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

[12] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, 2019.

[13] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[14] Samuel Rota Bulò and Marcello Pelillo. Dominant-set clustering: A review. *European Journal of Operational Research*, 262(1):1–13, 2017.

[15] Sophie Burkhardt and Stefan Kramer. A survey of multi-label topic models. *ACM SIGKDD Explorations Newsletter*, 21(2):61–79, 2019.

[16] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242, 2015.

[17] Gerardo Canfora, Andrea Di Sorbo, Enrico Emanuele, Sara Forootani, and Corrado A Visaggio. A nlp-based solution to prevent from privacy leaks in social network posts. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pages 1–6, 2018.

[18] Abhisek Chakrabarty, Onkar Arun Pandit, and Utpal Garain. Context sensitive lemmatization using two successive bidirectional gated recurrent networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491, 2017.

[19] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, 2017.

[20] Weilong Chen, Feng Hong, Chenghao Huang, Shaoliang Zhang, Rui Wang, Ruobing Xie, Feng Xia, Leyu Lin, Yanru Zhang, and Yan Wang. Curriculum learning for wide multimedia-based transformer with graph target detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4575–4579, 2020.

[21] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[22] Gobinda G Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[23] Freddy Chong Tat Chua and Sitaram Asur. Automatic summarization of events from social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[24] Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. Mitigating biases in toxic language detection through invariant rationalization. *arXiv preprint arXiv:2106.07240*, 2021.

[25] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[26] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020.

[27] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[28] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008.

[29] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[30] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, 2018.

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

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[33] Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for bert: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, 2020.

[34] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[35] Gunn Enli. Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. *European journal of communication*, 32(1):50–61, 2017.

[36] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5220–5227, 2004.

[37] Parsa Farinneya, Mohammad Mahdi Abdollah Pour, Sardar Hamidian, and Mona Diab. Active learning for rumor identification on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4556–4565, 2021.

[38] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[39] Linton C Freeman and Linton C Freeman. *Elementary applied statistics: for students in behavioral science*. New York: Wiley, 1965.

[40] Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Tom Ewing, Keith N Hampton, and Naren Ramakrishnan. Themedelta: Dynamic segmentations over temporal topic models. *IEEE transactions on visualization and computer graphics*, 21(5):672–685, 2015.

[41] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[42] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[43] Ivan Garibay, Alexander V Mantzaris, Amirarsalan Rajabi, and Cameron E Taylor. Polarization in social media assists influencers to become more influential: analysis and two inoculation strategies. *Scientific Reports*, 9(1):1–9, 2019.

[44] Ivan Garibay, Toktam A Oghaz, Niloofar Yousefi, Ece C Mutlu, Madeline Schiappa, Steven Scheinert, Georgios C Anagnostopoulos, Christina Bouwens, Stephen M Fiore, Alexander Mantzaris, et al. Deep agent: Studying the dynamics of information spread and evolution in social networks. *arXiv preprint arXiv:2003.11611*, 2020.

[45] Alexandra Georgakopoulou. 17 small stories research: A narrative paradigm for the analysis of social media. *The Sage Handbook of social media research methods*, 2017.

[46] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, 2021.

[47] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

[48] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[49] Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. Hieve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, pages 3678–3683. ELRA, 2014.

[50] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[51] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. PMLR, 2017.

[52] Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum. Integrating topics and syntax. *Advances in neural information processing systems*, 17, 2004.

[53] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24, 2004.

[54] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[55] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[56] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.

[57] Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. Asad: Arabic social media analytics and understanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118, 2021.

[58] Michael A Hedderich and Dietrich Klakow. Training a neural network in a low-resource setting on automatically annotated noisy data. *ACL 2018*, page 12, 2018.

[59] Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, 2021.

[60] Michael A Hedderich, Lukas Lange, and Dietrich Klakow. Anea: distant supervision for low-resource named entity recognition. 2021.

[61] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[62] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[63] Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Twenty-Ninth AAAI conference on artificial intelligence*, 2015.

[64] Weiming Hu, Wei Hu, Nianhua Xie, and Steve Maybank. Unsupervised active learning based on hierarchical graph-theoretic clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(5):1147–1161, 2009.

[65] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010.

[66] Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. Active curriculum learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 40–45, 2021.

[67] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[68] Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344. Association for Computational Linguistics, 2012.

[69] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[70] H Jiaji, C Rewon, R Vinay, L Hairong, S Sanjeev, and C Adam. Active learning for speech recognition: The power of gradients. In *The 30th Conference on Neural Information Processing Systems, NIPS. Barcelona, Spain*, pages 1–5, 2016.

[71] Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. In *International Conference on Machine Learning*, pages 5034–5044. PMLR, 2021.

[72] Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. Bag of tricks for efficient text classification. *EACL 2017*, page 427, 2017.

[73] Xin Kang, Xuefeng Shi, Yunong Wu, and Fuji Ren. Active learning with complementary sampling for instructing class-biased multi-label text emotion classification. *IEEE Transactions on Affective Computing*, 2020.

[74] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE, 2014.

[75] Soon Jye Kho, Swati Padhee, Goonmeet Bajaj, Krishnaprasad Thirunarayan, and Amit Sheth. Domain-specific use cases for knowledge-enabled social media analysis. In *Emerging research challenges and opportunities in computational social network analysis and mining*, pages 233–246. Springer, 2019.

[76] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *International Conference on Learning Representations*, 2017.

[77] Tom Kocmi and Ondřej Bojar. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, 2017.

[78] Kai A Krueger and Peter Dayan. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394, 2009.

[79] Hung-yi Lee, Ngoc Thang Vu, and Shang-Wen Li. Meta learning and its applications to natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 15–20, 2021.

[80] João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, 2020.

[81] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[82] Quanzhi Li, Armineh Nourbakhsh, Sameena Shah, and Xiaomo Liu. Real-time novel event detection from social media. In *2017 IEEE 33Rd international conference on data engineering (ICDE)*, pages 1129–1139. IEEE, 2017.

[83] Pengfei Liu, Xipeng Qiu, Xinchi Chen, Shiyu Wu, and Xuan-Jing Huang. Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2326–2335, 2015.

[84] Bo Long, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng. Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274, 2010.

[85] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. *Advances in neural information processing systems*, 30, 2017.

[86] Tomasz Maciejewski and Jerzy Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *2011 IEEE symposium on computational intelligence and data mining (CIDM)*, pages 104–111. IEEE, 2011.

[87] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182, 2019.

[88] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. 2005.

[89] Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, pages 359–367. Citeseer, 1998.

[90] Andrew J McMinn and Joemon M Jose. Real-time entity-based event detection for twitter. In *International conference of the cross-language evaluation forum for european languages*, pages 65–77. Springer, 2015.

[91] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.

[92] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[93] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[94] Elliot G Mishler. Models of narrative analysis: A typology. *Journal of narrative and life history*, 5(2):87–123, 1995.

[95] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[96] Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118, 2019.

[97] Ece C Mutlu, Toktam Oghaz, Jasser Jasser, Ege Tutunculer, Amirarsalan Rajabi, Aida Tayebi, Ozlem Ozmen, and Ivan Garibay. A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, 33:106401, 2020.

[98] Ece C Mutlu, Toktam Oghaz, Amirarsalan Rajabi, and Ivan Garibay. Review on learning and extracting graph features for link prediction. *Machine Learning and Knowledge Extraction*, 2(4):672–704, 2020.

[99] Ece Çiğdem Mutlu, Toktam Oghaz, Ege Tütüncüler, and Ivan Garibay. Do bots have moral judgement? the difference between bots and humans in moral rhetoric. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 222–226. IEEE, 2020.

[100] Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. Pre-training a bert with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996, 2021.

[101] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 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, 2010.

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

[103] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.

[104] Toktam A Oghaz and Ivan Garibay. Dominant set-based active learning for text classification and its application to online social media. *arXiv preprint arXiv:2202.00540*, 2022.

[105] Toktam A. Oghaz, Ece Çiğdem Mutlu, Jasser Jasser, Niloofar Yousefi, and Ivan Garibay. *Probabilistic Model of Narratives Over Topical Trends in Social Media: A Discrete Time Model*, page 281–290. Association for Computing Machinery, 2020.

[106] Ruth Page. Seriality and storytelling in social media. *StoryWorlds: A Journal of Narrative Studies*, 5:31–54, 2013.

[107] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, 2004.

[108] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, 2016.

[109] Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):167–172, 2006.

[110] Gustavo Penha and Claudia Hauff. Curriculum learning strategies for ir. In *European Conference on Information Retrieval*, pages 699–713. Springer, 2020.

[111] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[112] Pietro Perona and William Freeman. A factorization approach to grouping. In *European Conference on Computer Vision*, pages 655–670. Springer, 1998.

[113] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, 2017.

[114] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. *NAACL*, 2018.

[115] Martin Popel and Ondřej Bojar. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, (110):43–70, 2018.

[116] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577, 2008.

[117] Jay Pujara and Peter Skomoroch. Large-scale hierarchical topic models. In *NIPS Workshop on Big Learning*, volume 128, 2012.

[118] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[119] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[120] Amirarsalan Rajabi, Chathika Gunaratne, Alexander V Mantzaris, and Ivan Garibay. Modeling disinformation and the effort to counter it: A cautionary tale of when the treatment can be worse than the disease. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1975–1977, 2020.

[121] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[122] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

[123] Magnus Sahlgren, Tim Isbister, and Fredrik Olsson. Learning representations for detecting abusive language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 115–123, 2018.

[124] Sudeep Sarkar and Kim L Boyer. Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors. *Computer vision and image understanding*, 71(1):110–136, 1998.

[125] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[126] Ingo Schellhammer, Joachim Diederich, Michael Towsey, and Claudia Brugman. Knowledge extraction and recurrent neural networks: An analysis of an elman network trained on a natural language learning task. In *New Methods in Language Processing and Computational Natural Language Learning*, 1998.

[127] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[128] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *stat*, 1050:21, 2018.

[129] Burr Settles. Active learning literature survey. 2009.

[130] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[131] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[132] Aditya Siddhant and Zachary C Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, 2018.

[133] Hava T Siegelmann. Recurrent neural networks. *Computer Science Today*, pages 29–45, 1995.

[134] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *Tenth international AAAI conference on web and social media*, 2016.

[135] Sattaya Singkul, Borirat Khampingyot, Nattasit Maharattamalai, Supawat Taerungruang, and Tawunrat Chalothorn. Parsing thai social data: A new challenge for thai nlp. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–7. IEEE.

[136] PK Srijith, Mark Hepple, Kalina Bontcheva, and Daniel Preotiuc-Pietro. Sub-story detection in twitter with hierarchical dirichlet processes. *Information Processing & Management*, 53(4):989–1003, 2017.

[137] Asbjørn Steinskog, Jonas Therkelsen, and Björn Gambäck. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st nordic conference on computational linguistics*, pages 77–86, 2017.

[138] Sansiri Tarnpradab, Fereshteh Jafariakinabad, and Kien A Hua. Improving online forums summarization via unifying hierarchical attention networks with convolutional neural networks. *arXiv e-prints*, pages arXiv–2103, 2021.

[139] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[140] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018.

[141] Josep Valls Vargas. *Narrative information extraction with non-linear natural language processing pipelines*. Drexel University, 2017.

[142] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[143] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[144] Kristin Veel. Make data sing: The automation of storytelling. *Big Data & Society*, 5(1):2053951718756686, 2018.

[145] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[146] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2019.

[147] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, 2012.

[148] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.

[149] Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, et al. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322, 2019.

[150] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In *International Conference on Learning Representations*, 2020.

[151] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[152] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Wikipedia talk labels: Personal attacks, 2017.

[153] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, 2020.

[154] Jingyun Xu and Yi Cai. Incorporating context-relevant knowledge into convolutional neural networks for short text classification. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 10067–10068, 2019.

[155] Kang Xu, Guilin Qi, Junheng Huang, and Tianxing Wu. Incorporating wikipedia concepts and categories as prior knowledge into topic models. *Intelligent Data Analysis*, 21(2):443–461, 2017.

[156] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456, 2013.

[157] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[158] Liang Yao, Yin Zhang, Baogang Wei, Lei Li, Fei Wu, Peng Zhang, and Yali Bian. Concept over time: the combination of probabilistic topic model with wikipedia knowledge. *Expert Systems with Applications*, 60:27–38, 2016.

[159] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, 2020.

[160] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, 2019.

[161] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1942–1951, 2019.

[162] Wancong Zhang and Ieshan Vaidya. Mixup training leads to reduced overfitting and improved calibration for the transformer architecture. *arXiv preprint arXiv:2102.11402*, 2021.

[163] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[164] Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*, 2018.

[165] Ye Zhang, Matthew Lease, and Byron Wallace. Active discriminative text representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[166] Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, 2008.