WHAT DEEP LEARNING COULD BRING TO FRAME ANALYSIS

A Dissertation

by

YIKAI ZHAO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Kirby Goidel
Committee Members,	Alan Dabney
	Hart Blanton
	Timothy Coombs
Head of Department,	J. Kevin Barge

May 2020

Major Subject: Communication

ABSTRACT

This dissertation aims to solve two related questions that carry great significance for applied researchers: how do transfer learning models perform on textual classification and frame analysis under small training sizes. Transfer learning is deemed as one of the most innovative ideas in NLP (Natural Language Processing) and has broken numerous records in miscellaneous NLP tasks. It has expedited the NLP research by saving time for model training. Transfer learning may also achieve better results than prior practices on small training sizes. However, to date, there is few thorough investigation of transfer learning's performances on small training sizes.

This dissertation bridges the gap by conducting 2641 experiments of textual classification on performances of 6 different machine learning models across 5 diverse datasets and 8 different small training sizes utilizing different annotation schemes. Transfer learning models consistently outperform traditional machine learning (ML) models across different datasets and training sizes. Having said that, there are notable differences across Transfer Learning models. Two representative transfer learning models are used in this dissertation: BERT and XLNet. BERT model suffers a cold start problem with a larger variance in performances at moderately small training sizes (e.g. 400, 800) compared to other models. XLNet model should be our benchmark model in future practices because it achieves the best results across different training sizes and datasets with acceptable variances. A more compact annotation scheme, by collapsing categories into smaller number of groups, proves to increase model performances consistently across datasets and training sizes.

The second study suggests that transfer learning also benefits frame analysis greatly. With a compact annotation scheme and using a contextual Twitter dataset, which is unbalanced with 5 frames to classify, with a training size of 600, this research has achieved better than 72% accuracy with XLNet. This is optimistic for future research because even though each piece of text only contains the length of a normal tweet, which is significantly shorter than other sources of data, transfer learning could still achieve a satisfying level of result. This level of result could be used as a springboard for an iterative process that incorporates human relabeling to achieve more accurate results with less human labor.

This dissertation casts light on future research on textual classification and specifically frame analysis by offering guidance on model selection, performance evaluation, and annotation strategies. The visualization app (https://yikai-zhao.shinyapps.io/simulation_app/) made specifically for this dissertation could be used as a reference for future related research.

ACKNOWLEDGEMENTS

I would like to thank my committee chair and my advisor, Dr. Kirby Goidel, who inspired and encouraged me to pursue what I am passionate about. He is like an American father to me who showed genuine care. I would not come to this stage in study without his generous help.

I also want to thank my advisor in the Department of Statistics, Dr. Alan Dabney. His classes sparked my interests in machine learning and programming and his guidance encouraged me to finish the M.S. in Statistics.

I would also like to express my gratitude to my wonderful committee members, Dr. Hart Blanton and Dr. Timothy Coombs, for their consistent support and guidance throughout the dissertation process.

Thanks also go to the faculty and staff at the Department of Communication for making my time at Texas A&M University a valuable experience.

Finally, I want to thank God for giving me the wisdom and strength to go through the up and downs during these years. Thanks to my parents for their encouragement. And especially thanks to my girlfriend, Yuhong, for her love throughout the years that gives me hope for a better future.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Kirby Goidel, Dr. Hart Blanton and Dr. Timothy Coombs of the Department of Communication and Dr. Alan Dabney of the Department of Statistics.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This dissertation has been supported financially by the Department of Communication for the Summertime for Advancement in Research award in 2019.

NOMENCLATURE

ML	Machine Learning
DL	Deep Learning
SOTA	State-of-the-Art
BERT	Bidirectional Encoder Representations from Transformers
XLNet	General Autoregressive Pretraining for Language Understanding

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

In this chapter, I introduce the motivation for the studies in this dissertation. The research questions are presented afterwards. Finally, this chapter offers a detailed literature review regarding the key concepts including deep learning, transfer learning, and computational frame analysis.

The Motivation of the Dissertation

This dissertation aims to introduce the most updated advancements of Natural Language Processing (NLP) deep learning models to benefit social science researchers conducting framing analyses. Computer-aided textual analysis has been hindered by the demanding programming tasks of processing context-specific textual data. In other words, researchers need to write programs specifically for their particular tasks. Since 2017, breakthroughs have been achieved in NLP deep learning (DL), epitomized by the core philosophy of transfer learning (Ruder, 2018a). This dissertation investigates the applicability of transfer learning models to frame analysis. Specifically, the dissertation includes two related studies. The first study of the dissertation evaluates the performances of ML and DL models in textual classification generally under the constraint of small training sizes. In the second study, following the guidance of the first study, transfer learning models are applied to specific contextual frame analysis. The results and practical implications are discussed in terms of applying transfer learning to frame analysis.

The two studies in this dissertation cast light on communication research. Communication research often deals with small samples (because of the limits of human coding) or small training sizes. It is hard for communication research to scale up to large volume of data or longitudinally across a long period. A tool that could learn from a small training dataset would be invaluable for analysis of textual data. This tool could either be used in automating the annotation process of a large training size, or in helping build the annotation dataset iteratively. This dissertation conducted two studies to showcase the effects of transfer learning as the tool for social science researchers to scale up the research. The first study reveals the best practices in model selection, performance evaluation, and annotation strategies for textual analysis under small training sizes, which is the generic version of frame analysis from the angle of algorithms. The second study demonstrates the effeteness of transfer learning on frame analysis, with only a training size of around 600, on one of the few shortest forms of textual data, tweets. Transfer learning should be considered a powerful tool for communication researchers.

This dissertation intends to facilitate frame analysis in two ways. Firstly, with the help of DL algorithms, we could scale up our analyses to larger datasets across time efficiently without demanding human labor to annotate the data. Secondly, such effective DL tools could improve the validity of the computational approach adopted in communication research. Even the most recent frame analysis taking the computational approach (e.g. Boon, 2019; F. Lind, Eberl, Heidenreich, & Boomgaarden, 2019; Lucas et al., 2015; Terman, 2017; Walter & Ophir, 2019), claiming using statistical learning or machine learning, are unsupervised in nature or built on top of a bag-of-words approach, which is a simple mapping between the frequency of co-occurrences of certain word(s) and the target categories, with no capabilities to understand syntactical meanings and interdependence of words. The unsupervised or bag-of-words approach lacks validity in that using them to extract frames or topics is like asking a foreigner to extract meanings while the foreigner only recognizes some individual keywords. In terms of meaning-making and interpretation, human brain is the most well-crafted algorithm still when compared to all the rest of ML/DL models. ML/DL algorithms, if trained well, could learn the human way of classification effectively and apply it automatically to a large scale. Thus, to use the best of both human brain and ML/DL algorithms, an intuitive approach is like this: firstly, apply most of the human effort to inductively making sense of the data to come up with an annotation scheme; after annotating a small portion of the data according to the annotation scheme, we could train a ML/DL model to further scale up the analysis automatically to larger volume of data. This dissertation serves to experiment the applicability and effectiveness of this approach and promising results have been found by virtue of transfer learning in this dissertation.

Transfer learning is based on the idea of applying a well-trained prototypical deep learning model to any downstream textual dataset without much context-specific programming. The rationale is that transfer learning models, trained with a gigantic corpus of processed text, already possessed a high level of understanding of the language and even knowledge about the world (Ruder, 2018b). Transfer learning models could greatly lower the time and computing power required for further training a DL model for other downstream tasks. As the name “deep learning” suggests, the machine needs to consume enormous “teaching” material (in the context of deep learning, these “teaching” materials are called training data), as well as a large amount of computing power. In order to replicate the state-of-the-art (SOTA) result that Google achieved with their NLP model BERT (Bidirectional Encoder Representations from Transformers), which even surpasses human performance on the task of question answering (Devlin & Chang, 2018), we need to train such a model on a standard (NVIDIA RTX 2080 Ti) 4 GPU (Graphics Processing Unit) desktop for over 32 days (Dettmers, 2018). With the idea of transfer learning, we could directly build our customized model on top of well-trained transfer learning models to solve context-specific tasks in any downstream datasets, spending less time and effort in fine-tuning the customized model. Such context-specific tasks include textual classification, named entity recognition, textual summarization and machine translation (Ruder, 2018b). As a consequence of the lowered cost of model training, transfer learning helps automate and expedite the whole analytical process while achieving a high degree of accuracy without demanding labor to annotate an enormous set of training data. Such benefits could potentially scale up the utility of frame analysis from traditionally closed-reading of a small sample of texts to large scale textual analysis across time. For example, transfer learning has been found to achieve satisfying levels of accuracy on sample sizes as small as 100 movie reviews on a task of textual classification (Howard & Ruder, 2018).

My dissertation proposes that transfer learning could be applied to frame analysis and will achieve better results than the bag-of-words style of traditional machine learning methods, which is the current convention for conducting textual analyses. This dissertation aims to make a methodological contribution to the social science community by evaluating the efficiency and accuracy of the most advanced transfer learning models, especially when using small training sizes.

There are two studies in this dissertation. The first study evaluates the performances of ML/DL models in textual classification and compares the performances of DL models against traditional ML models. The ML models are featured with the bag-of-words style and such models include Linear Support Vector Machine, Random Forest, Logistic Regression, and Multinomial Naïve Bayes. They are chosen because they are the most common ML models for textual data (Pedregosa et al., 2011). The DL models include the most representative transfer learning models: BERT (from Google 2018) and XLNet (Generalized Autoregressive Pretraining for Language Understanding, from Google and Carnegie Mellon University, 2019). These DL models are chosen because of their contributions to the understanding of transfer learning and their proved performance on miscellaneous NLP tasks (Ruder, 2018b; "Text Classification," 2020). Although both DL models use the framework of neural networks, they are constructed through different mechanisms (or architectures) to discover patterns in textual data. We would expect to see differences in their performances across different datasets. This analysis also showcases the differences in performances between the DL models. A comparison of NLP DL models and ML methods yields insights for best practices in model selection in computer-aided textual analysis research.

The first study is designed as a set of comprehensive experiments investigating and contrasting the performances of DL and ML models on textual classification under small training sizes. Specifically, this dissertation limits the scope of the experiments to small training sizes (e.g. within the range of 100 to 5000) for two reasons. First, this study is designed to better guide the practice of applied researchers, such as social science researchers who often do not start their research with large training data sets. Given the great progress (SOTA results) achieved on multiple datasets and miscellaneous NLP tasks by transfer learning models, we expect to see the transfer learning models greatly outperform traditional ML models and achieve satisfying level of accuracy, even on small training sizes. This challenges the general impression that DL models need to consume very large training data. Second, it is surprisingly rare to see NLP literature discussing model performances under the constraint of small training sizes (Hanoz Bhathena, 2019). The relatively sparse research examining model performances on small training sizes (e.g. Howard & Ruder, 2018) only show a single accuracy point at each training size without mentioning the stability of the performances. Such a practice does not provide guidance to applied researchers whose goal is to apply the DL tools to downstream contextual tasks, rather than designing a DL model to break the SOTA record with little consideration of the cost of annotation of training data or computation resources. Thus, this research intends to bridge this gap by experimenting different ML and DL models across small training sizes (from 100 to 5000) in multiple datasets to reveal general guidance for model selection and evaluation of model performance.

In this first study, 2541 experiments have been run to offer a comprehensive overview of model performances under small training sizes across 5 different datasets. Among the 5 datasets, 4 of them are NLP benchmark datasets thus allowing us to compare the performances of our models against the SOTA results. The other one is a common dataset that NLP practitioners use for prototyping. The experiments can be categorized into three stages. The first stage investigates the performance of different ML and DL models on 5 different NLP datasets at relatively large training sizes. In the second stage, each model is run on different training sizes (from 100 to 5000) across datasets. We can directly evaluate the accuracies and generalizability of the model performances across datasets on small training sizes from this stage. The third stage aims to contrast the model performances, under the same training size, but with different number of classes contained within the training data. This stage suggests that the scheme of annotation, or called coding scheme in social science literature, matters for the model performances. From a research practitioner's perspective, collapsing the categories of the coding scheme into less groups might help improve the model performance.

In the second study of the dissertation, we apply transfer learning models and ML models to test their applicabilities to frame analysis. From an algorithmic and engineering perspective, frame analysis is a subcategory of text classification and it is no different from other textual classification tasks such as classifying if a sentence is grammatically correct (Warstadt, 2018) or the overall topic of news articles (Auer et al., 2007). Thus, we expect that the effective deep learning models can be greatly beneficial to understanding and classifying frames. However, such expectations require further investigation because the foundation of frame analysis depends heavily on human interpretation, setting it apart from other textual classification tasks. Frame analysis is difficult because it requires human interpretation to make sense of a corpus of textual data to develop a coding scheme from an abstract theoretical framework. As a consequence, the categories from such coding schemes become more context dependent. They are not a simple mapping of association with some keywords but are contingent upon the whole corpus of text, or even the contextual background that is not manifested explicitly within the corpus. For example, in news topics classifications (e.g., the DBPedia dataset used in this dissertation), each category is associated with clear and distinct indicators of certain words or phrases, thus making it easier for ML/DL models to capture the patterns to make accurate classification. In the case of frame analysis, the ML/DL algorithms need to 'comprehend' the whole piece of information (e.g. sentence, paragraph or even the whole article) to make the judgement. Furthermore, the frames are usually multifaceted in that they are not derived from a single dimension of meaning such as sentiment (positive vs negative) or mutually exclusive and collectively exhaustive (MECE) topics (e.g. 'sports', 'finance', 'politics' or 'entertainment' in news topic classification). Rather, frames are flexible in their scope of meanings and they are usually defined deductively from nuanced contextual theories or inductively out of the raw corpus of texts. One example of deriving frames is to follow the guidance of the theory of framing effects, which states that frames are to "promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described." (Entman, 1993, p.52) The frames constructed from such theories focus on different hierarchical or topical aspects of research topics or events. Frames might focus on causal interpretations, moral evaluations, or treatment recommendations. To compound the complexities, even within certain focus, there are further subcategories that might become frames. For instance, within the focus on moral evaluations, researchers might extract frames from the reasoning process behind a moral/immoral debate, or from another angle involving the motivation of the decision making. Therefore, frames could be operationalized very flexibly from the texts. Even though we have witnessed the power of transfer learning models on textual classification tasks, it is conceptually harder to classify frames simply based on a set of keywords or phrases without a comprehensive understanding of the texts, or even some meta-information such as the research contexts. Therefore, this dissertation will examine the effectiveness of transfer learning models in identifying frames in study 2.

Research Questions and Research Hypothesis

All of the abovementioned specific nuances about frame analysis make the application of transfer learning more challenging than other textual classification tasks. Thus, this dissertation raises the following research questions that are of particularly interest to social science researchers:

RQ 1: How do ML/DL models perform on textual classification tasks under small training sizes (e.g. 100 to 5000)?

RQ2: Would transfer learning models outperform machine learning models in textual classification tasks?

As we established above, the annotation scheme of the training data also impacts the model performance. Thus, we would like to investigate:

RQ 3: In general, under the same training size, will more compact annotation styles yield better model performance than more extensive annotation styles?

RQ 4: How do transfer learning models perform on a frame analysis task?

In the next section of this chapter, literature review on deep learning, transfer learning, and annotated frame analysis will be provided to offer more contextual understanding of this dissertation.

Deep Learning and Transfer Learning

Before introducing deep learning and transfer learning in details. I would like to clarify the terms used in this dissertation. Except for the following section of literature review about deep learning and transfer learning where we make distinctions between them, the rest of the dissertation interchangeably use the two terms because the DL models in this dissertation, BERT and XLNet are also transfer learning models. The majority of the SOTA DL models since 2017 are transfer learning models and it is likely to be a trend following the steps of transfer learning's dominant presence in computer vision. Thus, this dissertation would equate the transfer learning and DL in the rest of chapters.

Deep Learning

DL and ML are subfields of artificial intelligence. Both DL and ML use algorithms to discover patterns within data and using these patterns to predict real world outcomes without resorting to hand-coding software routines. Common machine learning algorithms include Logistic Regression, Random Forest, Support Vector Machine, clustering, Nearest Neighbors, and Neural Networks.

DL is a subcategory of ML that features miscellaneous algorithmic designs on the basis of deep neural networks and can be compared to other ML design such as tree-based models (e.g. Random Forest or XGBoost), linear models (e.g. linear regression), or kernel models (Kernel Support Vector Machine). Neural networks are developed as a crude approximation of the nervous system found in biological organisms (Perez, 2016). Within the neural network, information is transported between paths of neurons. Neurons have the capability to collect and aggregate information (numerical input) from previous neurons and then apply a non-linear transformation to accommodate more complex interactions. Although there is no definite consensus about the definition of deep learning, An AI leading researcher, Jeremy Howard, offered a set of features that distinguish DL from general machine learning models. First, DL models feature an infinitely flexible function, which theoretically could map the feature space (in social science terms, variable space) into any arbitrary target space, and the performance of such a function is theoretically guaranteed by the universal approximation theorem (e.g. Copeland, 2016; Mcneela, 2017; Nielsen, 2019). However, to achieve the capacity to fit such flexible functions, the DL model is usually tremendously large (e.g. 345 million parameters in BERT) and it takes a long time to train/fit the model, especially compared to traditional machine learning methods.

The second feature of DL is an all purposeful parameter fitting process such as Gradient Descent. Gradient Descent is a common optimization method that could iteratively adjust the parameters to minimize the loss function, such as mean squared error or cross entropy loss. The third feature of DL is its utilization of GPUs that makes training of DL models fast and scalable. GPUs were designed to facilitate gaming graphics which deals with large amount of matrix calculations simultaneously. GPU is not as fast and powerful in its processing power as that of a CPU, however, it is good at multitasking. Fitting a deep learning model is, in essence, a large number of iterative matrix calculations, which could be dramatically expedited by the use of GPU's multitasking capability.

Even though deep learning has been proven to achieve the best results in almost all NLP tasks (Ruder, 2018b), it requires great programming and field experience to build up the neural networks and operationalize them to achieve the best result. Thus, deep learning is still considered cutting edge technology that could only be manipulated by a small group of computer scientists. With the development of transfer learning, DL could became more accessible to applied researchers.

Transfer Learning

Transfer learning made it possible to popularize deep learning by making it more accessible to a wider range of researchers. Transfer learning denotes a practical philosophy that the “knowledge” a well-trained general model learned could be re-applied to other context-specific tasks. The general model is trained to learn tasks that do not pertain to any context, thus making the “knowledge” learned more generalizable. Transfer learning has been proposed and empirically tested in computer vision in 2012. A competition-winning deep learning model trained from ImageNet, a tremendously large pool of miscellaneous annotated images, has proven useful in initializing weights for completely different datasets and improving performance (Ruder, 2018a). Since then, transfer learning models have triggered an explosion in research focusing on computer vision, achieving SOTA results in multiple tasks such as image classification and object detection. The intuition behind transfer learning in computer vision is that the DL layers trained on large amount of training materials ‘learned’ to distinguish some fundamental vision cues: such as dots, lines, shapes and repetitive patterns (Zeiler & Fergus, 2014). These vision cues aggregated together and made up the complicated image objects that the DL models could recognize and classify. Thus, we could recycle those DL layers that could recognize those visual cues and apply them to a new task on a new dataset, leading to a tremendous reduction in requirements of time and computing power. This process of recycling and reusing some well-trained DL layers is called transfer learning.

In NLP, transfer learning was mostly limited to the use of pre-trained word embeddings, using a vector of weights to represent each word, which improved baselines significantly in early stages of the development prior to 2017. After 2017, researchers moved toward transferring entire models from one task to another (Ruder, 2018a). In the current literature of NLP deep learning, transfer learning particularly pertains to DL models. However, DL models are not transfer learning models if those DL models do not generate reusable pieces for other models or tasks. The most representative transfer learning models include ULMFiT (Universal Language Model Fine-tuning), BERT, and XLNet. This research will specifically choose BERT and XLNet as representative of transfer learning models and evaluate performances under small training sizes. However, before discussing BERT and XLNet in detail, it is helpful to first look at the inception of modern NLP transfer learning, ULMFiT. Doing so helps us to better understand the concept of transfer learning in NLP.

One of the earliest representative methods of NLP transfer learning is ULMFiT. ULMFiT achieved the best results at the early stages of transfer learning in NLP in multiple tasks and it is still one of the best models in text classification (https://nlpprogress.com/). It has also been utilized in tasks involving other languages such as German and Dutch (Rother & Rettberg, 2018; van der Burgh & Verberne, 2019). ULMFiT was built based on 3-layer LSTM (long short term memory) architecture called AWD- LSTM, which is a multi-layer bi-LSTM network (Merity, Keskar, Bradbury, & Socher, 2018). The AWD-LSTM is a specific and effective way of constructing neural networks to capture the meanings of a sequence of words rather than any individual word or set of words or phrases.

In ULMFiT, the transfer learning layers, in combination, are called a universal/general-domain language model (LM). It was pre-trained on a gigantically large corpus of texts. A commonly used LM in ULMFiT is called Wiki103, which was trained from 28,595 preprocessed Wikipedia articles and 103 million words (Merity et al., 2018). After the training of the LM, the LM itself learned how to predict the next word from the past sequence of words. This capacity of LM is so intuitive that it is considered to understand the language itself. Subsequently, such a universal LM could be applied to any target dataset and further trained in order to gain a contextual understanding of the language within downstream data.

As the transfer learning model evolved, one particular class of transfer learning models stood out and achieved many SOTA results. This class of transfer learning models utilizes a special DL architecture called transformers. The use of transformers ushered in a new era for NLP transfer learning in NLP. BERT and XLNet are key representatives of transformers. BERT is one of the first DL models that utilizes the transformer architecture. Transformer architecture could be intuitively understood as neural networks that draw connections between different parts within a sequence length (e.g. 256 tokens) of texts (Uszkoreit, 2017). By utilizing the transformer, BERT is designed to parse out the input embedding into different functional parts such as grammatical matching and meaning retaining. BERT also incorporates an explicit embedding part including the location of the input token (equivalent to a word but with some special items) to suggest the relative distance between different tokens (Devlin, Chang, Lee, & Toutanova, 2018). By matching and contrasting these functional parts, each layer within the transformers layers could recognize and organize the meanings of words (Alammar, 2018). Intuitively, the end product is a representation of meanings of a sequence of words within a certain context. The model also learned how to make sense of more complex linguistic phenomena such as compositionality, polysemy, long-term dependencies, and negation. XLNet builds on top of the BERT architecture by introducing the idea of *permutation language modeling* that learns to model the dependencies between all combinations of inputs in contrast to traditional language models that could only learn dependencies in one direction. XLNet holds the SOTA results for multiple NLP tasks and is considered the most advanced transfer learning model at the time of this writing.

Frame and Framing Analysis

Framing Analysis and Dimensions of Approaches

Scholarly interests in framing analysis are spread across multiple disciplines such as Communication, Psychology, Sociology, Linguistics and Computer Science. Even within the same discipline, frames may be interpreted differently in different subfields such as policy, political communication, and cultural studies (David, Atun, Fille, & Monterola, 2011; Field et al., 2018; Shah, Watts, Domke, & Fan, 2002). In addition, framing analysis has been conducted on a wide range of topic contexts such as technology (e.g. Gamson & Modigliani, 1989), terrorism (e.g. Powell, 2011; Yousaf, 2015), and public health (e.g. Gerlach, 2016; Shih, Wijaya, & Brossard, 2008). Given the omnipresence of the concept of framing, it has been interpreted in diverse ways. Thinking broadly, there are two school of thoughts regarding frames (Cacciatore, Scheufele, & Iyengar, 2016). The first school of thought is called (logically) equivalency framing. Rooted in the field of psychology, equivalency framing claims that human decision making is contingent upon how the information was contextualized, even though the information presented is logically equivalent. The second school of thought is called emphasis framing which was rooted in the field of sociology. Emphasis framing moves the discussion of framing outside of logically equivalent information into what information is manipulated and presented. This dissertation adopts the emphasis framing approach by defining frames as what is being communicated rather than how a piece of information is presented. I chose the emphasis framing approach because it reveals nuances and details of texts which would provide insights in understanding a certain issue, rather than testing the psychological effects of certain wordings or phrasing to describe information.

With the emphasis framing approach, there are different definitions of frames. For example, Gitlin (1980) interprets frames as "the principles of selection, emphasis, and presentation composed of little tacit theories about what exists, what happens, and what matters(p. 6);" Gamson and Modigliani (1987) conceptualize frames as "a central organizing idea or storyline that provides meaning to an unfolding strip of events" (p. 143); Frames have also been defined as organizing principles that are socially shared and persistent over time, and that work symbolically to meaningfully structure the social world (Reese, 2007). Although these definitions are critical to the conceptual understanding of framing, they all stress the idea of selection and salience. However, these definitions could not be directly translated into an applicable operationalization of frames for conducting framing analysis because there are no inherent guidelines for identifying frames using these definitions (Matthes & Kohring, 2008). Entman's (1993) widely cited definition of frames provides a roadmap to identify frames in an applicable way: “to frame is to select some aspects of a perceived reality and make them more salient in a communicating context, in such a way to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described (p. 52).” This definition reveals what frames generally do, such as defining problems, drawing causal inferences, making moral evaluations, and recommending treatments. Such a definition serves as an operational guideline that distinguishes frames from themes, arguments and other under-theorized concepts (Entman, Matthes, & Pellicano, 2009; Matthes, 2009). This article adopts Entman’s definition to further guide the development of the codebook, or called annotation scheme, of frames.

The diverse set of conceptual understandings of framing have resulted in a rich variety of methods for detecting and defining frames. There is no definite guidance on how to identify frames. Such frustration has been expressed in multiple cases. For example, Reese noted that framing authors “often give an obligatory nod to the literature before proceeding to do whatever they were going to do in the first place” (2007, p. 151). Similarly, “various observers have noted how subtly and unconsciously framing operates” (Gamson & Modigliani, 1989, p. 7), making it hard to find a definitive guide to find frames. Furthermore, it has been declared that a straightforward guideline for identifying frames does not exist (Chong & Druckman, 2007). Drawing upon the extensive literature on framing, Matthes and Kohring summarized five methodological approaches in framing analysis (2008). The first approach is the deductive approach. The deductive approach derives and decides frames from literature a pirori and codes the texts into these frames.

The second is the hermeneutic approach which identifies frames by offering an interpretative account of texts associating frames with broader cultural elements. The hermeneutic approach is grounded in the qualitative paradigms, which are based on small samples of texts derived from the discourse surrounding an event. Third, the linguistic approach identifies frames by analyzing the selection, placement, and structure of specific words and sentences at the level of paragraphs. The linguistic approach makes clear of the structural dimensions of frames that can be detected: syntax, script, theme, and rhetoric (Pan & Kosicki, 1993). The next is called manual holistic approach, in which a frame is derived by a qualitative analysis and coded as holistic variables in a manual content analysis. However, such an approach, together with the hermeneutic approach, have been cautioned with a danger of "long-scholar analysis" (Tankard Jr, 2001) due to the fact that there is no clear criterion for how to find frames. Finally, computer-assisted approaches come to the rescue to find more objective and reliable methods. However, the literature about computer-assisted approaches in communication (e.g. Matthes, 2009; Matthes & Kohring, 2008) i confines understanding of framing to dictionary-based approaches which attach frames to the use of specific words. Dictionary-based approaches cluster words that co-occur with each other to a certain frame or offer corresponding scores to those words. It is worth noting here that given the advancements in NLP, such as deep learning models, computer-assisted approaches have evolved to model tremendously more sophisticated and nuanced syntactical and semantical aspects of texts.

Although the above summarization of framing approaches makes a huge theoretical and practical contribution to guide future framing research, they do not stand as definitive accounts of framing methodology. As Matthes and Kohring (2008) acknowledged, the identified approaches do not serve as an exhaustive list and those five methods are not mutually independent of each other. Moreover, these methods are not measuring the same conceptual level of the framing methodology (R. A. Lind & Salo, 2002). In a more thorough attempt to summarize framing methodology, Matthes (2009) categorized different approaches according to different methodological dimensions. He proposed to know (a) if the analysis is text-based or number-based, (b) whether frames are derived inductively or deductively (c) whether coding is manual or computer-assisted (d) whether data-reduction techniques are used. These methodological dimensions can overlap and result in different approaches.

This article follows this trajectory in categorizing framing methods by their diverse methodological dimensions, rather than trying to categorize each of them into a specific type of method. With the evolvement of computational tools and methods, such as NLP, the framing methods are not black versus white but a mixture of methodological dimensions. For example, in the past, qualitative frame analysis is usually based on a small sample of text and frames are interpreted manually through close reading and synthesizing (David et al., 2011). However, it is a common practice to utilize computational tools, such as Nvivo to annotate data and automate statistical aggregation and inferences in qualitative frame analysis. It is ambivalent to categorize such practice either in qualitatively hermeneutic approach or computer-aided approach. Such a simple categorization of framing methods would yield little conceptual and theoretical rewards to understand the framing methodology until a more dynamic understanding of the framing methodology is acknowledged. Thus, this dissertation serves as a thorough update on synthesizing the different dimensions of framing literature, with a special focus on the use of computational methods. Different dimensions will be compiled and explained below:

Qualitative vs. Quantitative

On the face value, this may correspond to Matthes's (2009) presentation if the analysis is text-based or number based. However, it is hard to find studies that only used text or numbers to conduct framing analysis. Even the most qualitative studies, featuring manual close reading and interpreting frames from a small sample of texts, may resort to some level of numerical analysis such as counting or statistical hypothesis testing (e.g. Gerlach, 2016; Shih et al., 2008). Thus, qualitative vs quantitative may not be a useful axis to evaluate a framing method (Krippendorff, 2004). Closely related to this but more importantly, one might be interested in whether the frame is derived inductively or deductively.

Inductive vs. Deductive

The inductive approach starts with little or loosely defined presumptions about frames. It let the texts “tell” what the frames are through the researcher’s close reading and synthesizing (Semetko & Valkenburg, 2000). The deductive approach makes stronger assumptions by following theoretical guidance or past literature.

Even though qualitative studies tend to be inductive (Shim, Park, & Wilding, 2015), both qualitative and quantitative studies extract frames inductively or deductively (Matthes, 2009). Also, both manual and computational framing analysis use inductive and deductive approaches.

The inductive approach has been criticized for not offering a definitive and objective guide to operationalize and interpret frames, leading to a deficiency in replicable findings (Matthes, 2009). In contrast, a deductive approach is inflexible and cannot identify new frames due to the strong assumptions about frames, thus making it difficult to capture all of the important frames (Matthes & Kohring, 2008).

The inductive vs. deductive dimension becomes more important when utilizing machine learning methods to automatically identify frames because this dimension decides how the training data is going to be annotated. Machine learning is a subfield of artificial intelligence, which uses algorithms to learn from data, identify patterns, and make predictions without explicit human intervention. When using machine learning, researchers need to provide training data, which annotates texts with the correct frame. The training data for an inductive approach is tremendously different from that in a deductive approach. The inductive approach eventually would classify the texts into one or more frames per se, while the deductive approach would classify the texts into many indicator questions. Indicator questions tap into one small aspect of the frame elements. The cluster of these frame elements may represent frames. Such frame elements may be suggested by the previous framing literature conducted in the same context, or by a clear operationalization of a framing definition.

Manual vs. Computational

This dimension does not yield clear-cut difference in terms of framing method as we have witnessed more research that combines the benefits from both approaches. The manual close reading provides a flexible and iterative searching scheme and a better understanding of frames at a higher level of abstraction. The computation methods have much to offer in formalizing and automating the analysis of framing, enabling greater scale and breadth of application across issues and disciplines (Card, Boydstun, Gross, Resnik, & Smith, 2015). Computational methods should not only involve numerical counting and summarizing, as they appear in convention, but more importantly, include the process of building a statistical model to make predictions related to frames.

General Frames vs. Context Specific Frames

Frames have been conceptualized at different levels of abstraction from study to study (Matthes, 2009). Context-specific frames (also called issue-specific frames) denote those ad hoc frames derived through a close reading and relying on n interpretation based on contextual background knowledge (Shah et al., 2002; Tucker, 1998). General frames (alsocalled generic frames) do not tie to any specific context but intend to subsume all specific contextual frames that might appear on any issue of public concern (Card et al., 2015). For example, Shih et al. postulated six generic frames for their studies on epidemics: consequences, uncertainty, action, reassurance, conflict and new evidence (2008).

Supervised Learning vs. Unsupervised Learning

This dimension only applies to methods involving machine learning. In the context of framing, supervised learning methods would provide clearly labeled frame(s), or target frame(s), for each text. The supervised learning algorithm is responsible for discovering the pattern before the texts and the label so that it could predict the frame automatically given the input of texts. Prior studies have achieved success in conducting supervised learning framing analysis (Baumer, Elovic, Qin, Polletta, & Gay, 2015; Field et al., 2018). Common supervised learning algorithms include Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest. Unsupervised learning does not require a priori labeled frame(s), rather, it clusters or categorizes according to inherent patterns within the data itself. Common unsupervised learning methods include Principal Component Analysis, K-means Clustering and Hierarchical Clustering (David et al., 2011; Lin, Hao, & Liao, 2016; Odijk, Burscher, Vliegenthart, & De Rijke, 2013; Semetko & Valkenburg, 2000; Shim et al., 2015; Tian & Stewart, 2005; Yousaf, 2015).

Unsupervised learning algorithms have been predominately used in deductive approaches. Supervised vs. unsupervised learning is also an important dimension because it determines the use of possible algorithms.

Shallow vs. Deep Learning

There is no definition about shallow learning as it is mainly used to contrast with deep learning. Shallow learning usually omits higher-level information and interactions, or it employs naïve assumptions about the data. Shallow learning can be useful but it trades expressivity for efficiency. In the context of NLP, an example of shallow learning is word embeddings, which uses a vector of numbers to represent each word. Word embeddings have been found to increase the classification accuracy by two to three percentage points (Kim, 2014). However, it only captures the meanings of individual words and treats them independently without any interaction of word meanings. Another common shallow learning practice is called the bag-of-words method, which holds strong and naïve presumption that a sentence is merely a combination of words and the order of those words do not matter. The bag-of-words simply take consideration of the counts of words within different documents. It does not perform well in some texts that involve complex language phenomena such as compositionality, polysemy, anaphora, long-term dependencies, agreement, and negation.

Deep Learning is a collection of techniques and methods that are used to build flexibly differentiable architectures. Deep learning models adopt the basic setup of neural networks but have more sophisticated designs in structuring those neural networks to achieve better prediction accuracy and generalizability. Deep learning models not only encapsulate the word meanings like word embeddings, but also take consideration in the positional sequence of them and the interaction of meanings between words and phrases. Deep learning has been proven to achieve the best results in miscellaneous tasks such as machine translation (e.g.Artetxe, Labaka, Agirre, & Cho, 2017), reading comprehension (e.g.Rajpurkar, Zhang, Lopyrev, & Liang, 2016), and sentiment analysis (Felbo, Mislove, Søgaard, Rahwan, & Lehmann, 2017). However, there is no research that investigates the accuracy and efficiency of deep learning in frame analysis. The second study in this dissertation serves as a thorough experiment on the effectiveness of deep learning in frame analysis.

The next chapter will introduce in detail about the research methods and data for the two studies separately.

CHAPTER II

METHODS AND DATA

This chapter offers detailed explanation about the research methods, research procedures and the datasets for both studies in this dissertation.

Methods and Data of the First Study

This section will provide details regarding research methods and procedures for the study 1, which is the comprehensive experiments of ML/DL model performances on small training sizes. The experiment procedures, the three stages of the experiments, are discussed first. Especially, during each experiment stage, a rationale is provided for the design of the experiments. Second, brief explanations about the different models are provided. Finally, this chapter will provide details regarding each dataset used in study 1.

Experiment Procedures

To date, there has been little research conducted specifically under the constraint of small sample sizes (Hanoz Bhathena, 2019). Evidence regarding the variability of the performances of those ML/DL models is also lacking. Thus, the first study of this dissertation is to conduct comprehensive experiments to investigate the performance of ML/DL models across 5 datasets, utilizing different small training sizes (from 100 to 5000), and under different annotation schemes.

The experiments could be divided into three stages. The first stage is to investigate the model's performances using a relatively large training size. The original plan of the experiments at this level is to use the whole training size to train different ML/DL models and to observe their performances compared with the SOTA results. However, such a plan become infeasible given the bottleneck of the computing power because the first stage experiments involve running at least 5 repetitions per model on each dataset, which would take a tremendously long time to finish if we utilize all of the training data for those datasets (e.g. 3,000,000 for Amazon Review dataset). Therefore, this dissertation took the compromise by truncating the training size to 10,000, which should be considered a moderately large training size for an applied research analysis using DL. The testing data was sampled randomly from the original dataset at the size of 2500. We randomly sampled training and testing data 5 times from each dataset and trained the models on those sets of training and testing data. The accuracy that each model achieved on each dataset during each run was recorded for comparisons.

The second stage of the experiments added complexity on top of the first layer by evaluating the models' performances on different small training sizes. There are 7 small training sizes this research operates on: 100, 200, 400, 800, 1600, 3200, and 5000. For each of the 7 small training sizes, this study experimented 5 or 7 repetitions of each model on each dataset. For each rep within the experiments, the corresponding size of training data was sampled from the original dataset and I set aside a test size of 2000 sampled from the original data as well. This design of the experiments is to mimic the applied research setting where we do not start with large training size. In those cases, the training data researchers acquire may be conceptually considered to be a subsample of a large training data space, like in this experiments we sample 100 or 200 training data from the total training size of datasets. This design of the experiments also reveal the influence of the covariate space of the sampled training data to the covariate space of the testing data, thus revealing the stability of the models' performances. If the model could not handle the discrepancies of the covariate space between the training and testing data, then the variance of the model performances would be large, especially during the small training size setting. The number of repetitions of the experiments could be viewed from the following table.

Table 1 Number of repetitions of experiments in Stage 2

	# Classes	**ML models**	**DL models**
AG News	4	7 sizes * 4 models * 10 reps	7 sizes * 2 models * 5 reps
DBPedia	13	7 sizes * 4 models * 10 reps	7 sizes * 2 models * 5 reps
Yelp Review	5	7 sizes * 4 models * 6 reps	7 sizes * 2 models * 5 reps
Amazon Review	5	7 sizes * 4 models * 6 reps	7 sizes * 2 models * 5 reps
Customer Complaint	18	7 sizes * 4 models * 6 reps	7 sizes * 1 model * 5 reps

Note: The XLNet could not be fit on the Customer Complaint dataset because the model is so large and it takes a great amount of RAM of the GPU. The complaint texts are also too large to be fitted into the rest of the RAM. Thus, this research only tested the performances of BERT and ML models on this dataset, without XLNet's.

In stage 3, the goal of the experiments is to investigate if the compact annotation scheme would yield more accurate and stable model results. This stage was directly inspired by the current experimental results from Yelp and Amazon Review datasets. Even the current SOTA results, trained from the most advanced NLP model, XLNet, from a large training size (3,000,000 for Amazon Review and 650,000 for Yelp Review datasets) are merely 67.74% and 73.20% for Amazon and Yelp Review datasets. There are multiple angles to explain such underperformances but one of the key reasons is regarding the cross entropy loss function. In the cross entropy loss function, the model treats a misclassification of 4, from the correct class of 5, equally as the misclassification of 1. However, in practice, a misclassification of 4, from the truth of 5, is way more understandable and tolerable than a misclassification of 1. This choice of loss function complicates this task thus requiring the DL models to be more distinctive between nuances of reviews to make the right classification. Under the cross entropy loss function, the accuracy would be penalized with minor differences like mentioned above and such minor differences might contribute to the majority of error to the final performance. In order to alleviate the error incurred simply due to the use of cross entropy loss function, we removed the categories of 2 and 4 from both training and testing set, in both Yelp and Amazon Review datasets. This study raises the question about the model's performance on such a compactly annotated dataset where the differences between classes are more distinctive assuming the same training and testing size. For the DBPedia dataset which originally has 13 classes, I truncated the dataset with 5 classes left. For the Customer Complaint dataset which originally has 18 classes, the compactly annotated dataset has 3 classes left. The compactly annotated datasets used in the stage 3 are presented in Table 3.

Table 2 Number of repetitions of experiments in Stage 3

	Compact # classes	ML models	DL models
AG News	-	-	-
DBPedia	5	7 sizes * 4 models * 6 reps	7 sizes * 2 models * 5 reps
Yelp Review	3	7 sizes * 4 models * 6 reps	7 sizes * 2 models * 5 reps
Amazon Review	3	7 sizes * 4 models * 6 reps	7 sizes * 2 models * 5 reps
Customer Complaint	3	7 sizes * 4 models * 6 reps	7 sizes * 1 model * 5 reps

Note: The reason we did not define a compactly annotated version of AG News is that the original dataset only contains 4 categories. The compact annotated version, at least three in this dissertation, probably would not make much difference from 4 to 3 categories.

Models and Training Parameters

There are 6 models used in this study with 4 of them being ML and 2 DL . The 4 ML models are LinearSVC, Logistic Regression, NaïveBayes and Random Forest. The 2 DL models are BERT and XLNet. All ML models follow a paradigm of method for text classification called bag-of-words (BOW), which differs from the transfer learning paradigm.

BOW is a common method to extract features from text documents (e.g. Aaldering & Vliegenthart, 2016; Haselmayer & Jenny, 2017; Y. Zhang, Jin, & Zhou, 2010). These features are used for training machine learning algorithms. BOW creates a vocabulary of all the unique set of tokens (such as words, punctuations or special symbols) occurring in all the documents in the training set. BOW is usually used with another concept called N-grams, which denotes a contiguous sequence of N tokens from a given text. Researchers could fine-tune the number N in N-grams to decide the scale of feature units. BOW method does not feed the raw texts into ML models, instead, it creates and feed a document-term matrix. Within the matrix, each row represents a document and each column represents a certain feature which depends on the choice of N-grams. If N is 1, then the feature unit is usually a word. On top of this basic document-term matrix, further modifications can be made to improve the training of ML models. For example, a common practice is to remove 'stop words' which are those common words without substantial meanings such as "the", "a", or "and." Removing these 'stop words' would help the document-term matrix to be more parsimonious thus improving the efficiency of algorithms. Another technique commonly used is called TFIDF, meaning term frequency–inverse document frequency, which is a statistical measure that evaluates the relevancy of a word to a document relative to other documents. Instead of simply tallying the presence of N-gram features across documents, TFIDF is calculated by multiplying two metrics: the count of a word appears in a document and the inverse document frequency of the word across all documents. TFIDF is especially useful in information retrieval. Under TFIDF, the intuition is that those words that are common in every document, such as 'this', 'that', or 'with' would be down-weighted in importance; however, if a certain word appears many times in some documents, while not appearing many times in others, it is considered important.

In this research, after sampling the training and testing data, I removed the 'stop words' first. The N-grams are set within the range of 1 to 3, meaning that all of individual words, two and three consecutive sets of words are included in the document-term matrix. TFIDF was also applied to the document-term matrix to assist pattern retrieval for the ML algorithms. The BOW method may be useful in some tasks, however, it get greatly outperformed by DL models. One of the key reasons is that each pattern, N-gram, is independent of other patterns. Thus, BOW could not handle the dependence of meanings across patterns.

The DL models were trained from a different mechanism and handled the dependence of meanings better than BOW. For BERT, which serves as a base for more recent advancements in other transfer learning models, each token is represented by a long vector (e.g. of length 256) of numbers called word embeddings. These number were trained from a gigantic corpus of texts thus they encode the meanings of tokens 'learned' from the training materials. Subsequently, after the embeddings are fed into the DL models, the model architecture would match and contrast the meanings of different tokens to extract more abstract meanings from various combinations of tokens. The whole model, including the word embeddings and the model weights, could be transferred to miscellaneous unrelated datasets.

XLNet has made improvements in introducing the idea of permutation language modeling that learns to model the dependencies between all combinations of inputs in contrast to traditional language models that only learn dependencies in one direction. For example, there is a sentence "I like cats more than dogs." The "understanding" of a language model is represented by its capability to predict the next word correctly. This prediction is usually done sequentially from the left of the sentence to the right.

However, this one directional learning only captured the words before the prediction but with no idea about the subsequent words. This one directional could be improved by learning the sentence from two directions with the word you are predicting masked. For example, if you were to predict the third word in the sentence and you saw ‘dogs’ appear at the end, you have a better clue that this third word is an animal, and very likely to be cats. You would not have this insight if you did not know there is the word “dogs” at the end, if you use a one directional learning. XLNet further extends this two-directional learning into a more general learning process. It breaks the order of the prediction, meaning we do not predict the next work following the left-to-right order. We randomly draw a word from the sentence from any location, and predict the next word randomly drawn from a random location. Intuitively, this way of learning puts a higher standard for the model in that it does not only need to know how to predict the next word given prior clues, it also needs to know how to predict other words in the proximity of the random word you drew. To the model, if it saw a word “than” then it should predict that the contents in proximity very likely contain two things to compare and a comparative word. This way of training in XLNet proved to be working and XLNet is the current SOTA holder for multiple NLP tasks.

In terms of model fitting, for both ML and DL models, I followed the default “out-of-box” set of parameters without fine-tuning the parameters for their maximum performance. The specific parameters used, which are necessary to initiate the algorithm since no starting value is provided in the default, are compiled in Table 4.

Table 3 Parameters for different models

	Parameter set	Implementation package
Logistic Regression	max_iter = 200	Scikit-learn
LinearSVC	-	Scikit-learn
NaïveBayes	-	Scikit-learn

Random Forest	n_estimators=200,
	max_depth=3
		Scikit-learn

BERT	Max learning rate = 1e-04, Epochs = 5 Batch size = 64	Fastai and Huggingface
XLNet	Max learning rate = 1e-05, Epochs = 5 Batch size = 4	Fastai and Huggingface

Datasets

For the first study, this dissertation adopts 5 datasets to view the performance of ML/DL models under the small training sizes. These 5 datasets are: AG News, DBPedia, Yelp Review, Amazon Review, and Customer Complaint dataset. These 5 datasets were chosen for 2 reasons. First, they are either the benchmarks to compare model performances within the NLP community, or because they are common datasets NLP practitioners use. Thus, by using these dataset, we would be able to directly compare the performances of our models against the SOTA results under small sample sizes. Second, even though these datasets could all be fit into the category of textual classification, their annotation schemes are different and they represent different applications of textual classification in practice. For AG News and DBPedia, their annotation schemes are to classify each piece of information into distinctive types (e.g. ‘sports’, ‘politics’). For Yelp and Amazon datasets, the annotation scheme is to classify a piece of review text into a 5-point rating where 1 means extremely dissatisfying and 5 means extremely satisfying. These two annotation schemes are common in research and practice. Thus, by experimenting on these datasets, we can get an overview about the models’ performances on different applications of textual classifications.

Table 4 Overview of Datasets in Study 1

	# Classes	Compact # classes	Training sizes	Testing sizes
AG News	4	-	120,000	7600
DBPedia	13	5	560,000	70,000
Yelp Review	5	3	650,000	50,000
Amazon Review	5	3	3,000,000	650,000
Customer Complaint	18	3	600,000	50,000

AG’s news corpus

This dissertation adopts the version of AG’s corpus of news article that has been used consistently as a benchmark for evaluating model performances ("Text Classification," 2020; Yang et al., 2019; X. Zhang, Zhao, & LeCun, 2015). The original AG News corpus contains 496,835 categorized news articles from more than 2000 news sources. The version used here, consistent with other literature, includes the 4 largest classes from the original corpus to construct our dataset, using only the title and description fields/columns. The number of training samples for each class is 30,000 and testing 1900.

DBPedia Ontology Dataset

DBpedia is a crowd-sourced community practice to extract structural information from Wikipedia (Lehmann et al., 2015). This dissertation adopts the version of the DBPedia datset that is consistent with NLP research community (X. Zhang et al., 2015). The DBpedia ontology dataset is constructed by picking 14 non-overlapping classes from DBpedia 2014. From each of these 14 ontology classes, this version includes 40,000 training samples and 5,000 testing samples. The fields/column used for this dataset contain title and abstract of each Wikipedia article.

Yelp Reviews

The Yelp Reviews dataset is obtained from the Yelp Dataset Challenge in 2015 (https://www.yelp.com/dataset/challenge). This dataset contains 1,569,264 samples with a 5-point rating for each text. This dissertation adopts the version of the dataset to predict the number of stars the user has given. The full dataset has 130,000 training samples and 10,000 testing samples for each star.

Amazon reviews

The original Amazon Review dataset was derived from the Stanford Network Analysis Project (SNAP, http://snap.stanford.edu/), which spans 18 years with 34,686,770 reviews from 6,643,669 users on 2,441,053 products. Similar to the Yelp review dataset, this dissertation adopts the version to predict a 5-point star rating from a review test. The full dataset contains 600,000 training samples and 130,000 testing samples in each class. The fields used are review title and review content. The SOTA results used both fields (review title and review content) to make the prediction of the 5-point rating score. In this dissertation, we only use review content instead of including the review title. This is because in general social science research practice, researchers would only have datasets containing individual paragraphs of texts that does not include a brief summary or title. To be consistent with the goal of this research, that is to evaluate the performances of ML/DL models, especially the DL models, for social science researchers, we chose not to use the review title field but only the review text field. Doing so would inevitably lead to less satisfying results when compared to the SOTA results, but these results are more relatable to social scientists or DL practitioners.

Customer Complaint

Each week the CFPB (Consumer Financial Protection Bureau) sends thousands of consumers' complaints about financial products and services to companies for response. By adding their voice, consumers help improve the financial marketplace. This is a common dataset that NLP practitioners use in blogposts to illustrate model performances (Li, 2018). There are different annotation schemes attached to this dataset from the product type, sub-product type to sub-issues. In this research, we used the annotation of the product type which includes 18 categories such as mortgage or debt collection. The data harvested for this study includes 600,000 sized training data and 50,00 sized testing data.

Methods and Data of the Second Study

This section includes the dataset, the methods for annotating the data, the experimental procedure, and the detailed documentation of the annotation scheme of the second study.

Dataset

This study will scrape its own twitter dataset surrounding a controversy about a DL model called GPT2. In February 2019, DL powerhouse OpenAI released a new language model, GPT2, which was trained to predict the next word in a sample of 40 gigabytes of Internet text. GPT2 could generate text that mimic the style and content of the conditioning text, allowing the user to generate realistic and coherent continuations about certain text prompts (Whittaker, 2019). The model is a vast improvement on the prior counterparts by producing longer text with greater coherence in meaning. The release of GPT2 caused tremendous controversy. The GPT2 is indeed the best language model and it could be extremely useful in speech recognition and text generalization. However, OpenAI only released a elementary version of the model due to their fear that the model could be misused to generate fake news, impersonate people, or automate abusive or spam comments (Whittaker, 2019). This practice is contrary to the norm in the DL community where progresses are usually open sourced for replication. There were criticisms that OpenAI is closing off its research. The proponent of OpenAI cited the recent turmoil created by another DL technique - commonly referred to as Deepfake. It uses DL to generate fake videos featuring celebrities. However, criticisms remain that OpenAI's decision was to attract fame and media exposure rather for the goodness of the scientific community.

Given the controversy related to OpenAI's decision, it should be expected there should be adequate frames surrounding this issue. Moreover, GPT2 controversy is still an issue debated within the scientific community, thus Tweets should be long enough to contain useful information. This research uses Tweets as analytical units also because DL is such a fast-evolving field and Twitter is one of the most involved media platforms where discussions and discourses are developed. Therefore, choosing to analyze Tweets over traditional news reports would yield more nuanced findings about frames.

The data harvested in this study include 779 tweets from the time frame of February 15th to April 26th in 2019. February 15th is the first data when OpenAI's decision exposed which stirred debates. April 26th was chosen as the final date in order to have above 800 raw tweets with the hashtag #GPT2. After removing the tweets that are not related to the GPT2 model, there are in total 779 tweets left in our final dataset.

Data Annotation

This study utilizes an iterative approach that commute between the emergent grounded data and the theoretical inspirations offered by the past research (Tracy, 2019). This iterative approach, specific to inductive qualitative research refines and adjusts the coding as the data analysis proceeds. Researchers continually adjusting their analytical framework by drawing insights from the data, leveraging the refined version compared with the former one, delving back into data, referring to the literature and theory, and so on. Under the guidance of this iterative approach, it is not uncommon to identify the pivotal themes inspired from data but that have not been systematically investigated by the literature. As a result, the literature review may be revised based on the emerging themes from post-analysis. Following this process, the annotation scheme, or codebook, was developed to guide researchers to annotate the individual tweets. After consulting to Entman's operationalizable definition "promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described." (1993, p.52) and closely reading the data, this study extracts 7 frames. The annotation scheme is presented in the next subsection.

Annotation Scheme

This section consists of the coding scheme for the frame analysis. Under this annotation scheme, each tweet is categorized to only one frame. The coding scheme and its explanation is detailed below:

Concern Over the Damage

This frame stresses the possible damages the GPT2 could possibly cause. The damages include the spread of fake news and profane language.

The associated words include: "harmful", "threat", "danger", "terrifying", "apocolypse", "malicious", "abusive" and so on.

An example tweet: OPEN-SOURCE A.I. MAY BE TOO DANGEROUS TO RELEASE -- At its core, GPT2 is a text generator. The AI system is fed text, anything from a few words to a whole page, and asked to generate subsequent words.

The frame could also be implicitly implied without the outright use of words with negative connotations. For example: @J_Kom_ another reason why this is such a bad idea. To show what that means, OpenAI made one version of GPT2 with a few modest tweaks that can be used to generate infinite positive or negative reviews of products https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

Questioning the Purpose and Motivation of OpenAI

This frame stresses the inquiry about the purpose of developing GPT2 given its powerful yet limited technological purpose. This frame also focuses on the discourse surrounding the motivation of OpenAI for not releasing the source code. The core of the discourse is whether OpenAI chose not to release the source code as a marketing stunt rather for the benefit the public benefits. If a message mainly discusses the disagreement, disblief or any doubt about the potential harm GPT2 could cause, it is also classified under this frame.

An example tweet: Am I missing something? the big question for me is: What's the purpose of GPT2? It's not solving a real problem eg relieving people from tedious work or enhancing inclusion - it's enhancing productivity as an end in itself. https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction.

Leisure Banter

This frame consists questions or statements that are purely for the purpose of joking, sarcasm, or the curiosity. This frame involves GPT2 generated text. This frame usually appear with the mention of general-sense AI, robot or Sci-fi movies. This frame may also feature the machine generated response from certain human generated prompt. An example tweet: “At some point, if I do my job right, I'll just write some GPT2 emulator to tweet for me; wait, has anyone tried gpt2 for horoscope production?”

Support OpenAI's decision and endeavor

This frame includes arguments supporting OpenAI's decision for not releasing the full GPT2 source code. It also includes accolades for OpenAI's endeavor to develop GPT2.

Note that this frame usually couples with the 'Concern over the damage' frame. However, this is a particular useful frame to be singled out in this research context.

An example tweet: “Amazing work @OpenAI both on the development of GPT2 text generator and consideration of the #Ethics of its release. Glad to see the ethics of #AI models getting genuine consideration by their creators. #NLP”

Informational and technological details

This frame focuses on the technological detail about the algorithm design of GPT2 or the programming tips about GPT2. This frame contains mainly informational messages without any emotional or judgemental words or phrases. It also contains messages where a neutral or undecided stance is expressed. If an informational piece coupled with any sentiment, then it is likely to be classified as other frames.

An example tweet: “Weekend experiment with @OpenAI's GPT2: Generated a dark future for @IDEO Fictions become believable once plausible and meticulous details are baked in; subtle accuracies are the beauty/curse of #GPT2.”

Control or the remedies for the possible misuse

This frame consists tweets that discusses the remedies or solutions for containing the possible misuse of GPT2 or spread of fake news. The frame may also involves precautions that needs to be taken for future artificial intelligence threat.

An example tweet: “The community should sort out patterns that help diagnose AI generated texts.”

Technological Excitement and Approval

This frame includes messages that expresses the prowess or the convenience of GPT2 model. This frame also contains information about the future development of GPT2 related technology. The messages usually stress the vividness of the generated text and that usually bear real world resemblance.

An example tweet: “Congratulations to the @OpenAI team on releasing #GPT2 ! Not only are the reported results a jump in the SOTA, but the ensuing open sourcing conversation has raised some points that are worthy of ongoing discussion.”

Experiment procedure

For the contextual dataset used in this study, both ML/DL models used in study 1 have been applied. The experiment procedure is similar to stage 1 in the first study: we randomly shuffled the dataset and divided the dataset into training and testing set (the testing set proportion is 20%). The model performances are compared and plotted for visual illustrations.

We only have training data of the size about 623 (80% of the whole dataset of size 779), however, we have 7 imbalanced classes. The class distribution is presented in Table 6.

Table 5 Distribution of frames in GPT2 dataset

Frame	Count
Leisure_banter	238
Concern	170
Informational	166
Tech_excitment	63
Question_purpose	58
Control	51
Support_decision	33

CHAPTER III

RESULTS

This chapter presents the results from both studies: the first study is the comprehensive experiments about the DL/ML's performances on textual classification under small training sizes; the second study is to put ML/DL into the task of frame analysis, on the GPT2 dataset.

Study One Results: Comprehensive Experiments

In this section, we present the results from the first study. As explained before, there are three stages for the experiments. The first stage is to investigate the marginal performances of different ML/DL models on the benchmark datasets, as compared to the current SOTA results. The second stage is to add another layer on top of the first stage by observing the models' performances under different small training sizes (from 100 to 5000). The third stage is to verify if more condensed annotation scheme would encourage better model performances under the same level of small training sizes. We present the results of stages accordingly in the next three subsections.

We only present results in terms of accuracy as a key metric to evaluate the model performances. This is because the datasets we chose are balanced in their distribution of classes/clusters. For example, in AG News dataset, each of the 4 categories of news have approximately the same amount within both the training and testing datasets. Also, the goal of the experiments is to contrast and compare the model performances to guide our future model selection and annotation strategies, rather than showing how to achieve the best results through parameter fine-tuning or reweighting of the training data. Therefore, we do not review additional performance metrics such as sensitivity, specificity, ROC-AUC curve, or confusion matrices. This practice is common in literature that evaluates algorithm performances.

Stage 1 Results

Model Performances Vary Depending on Datasets

From the results, the first theme we observe is that model performance varies depending on the dataset. AG News and DBPedia datasets are generally considered easier datasets to classify because their tasks are to classify news items into general topics. All the models generally perform well on these datasets. In AG News dataset, 5 out of the 6 models achieved approximately 90% accuracy levels. The Random Forest model is the only one that underperformed. The SOTA result have been reported as 95.51% ("Text Classification," 2020) and surprisingly our XLNet's performance even slightly surpassed the SOTA result. This might happen due to our choice of the hyperparameters (such as learning rates, weight decay etc.). In DBPedia dataset, 5 out of 6 models perform, on average, above the 95% accuracy level.. Random Forest once again yields the worst performance. The SOTA result was reported as 99.38% ("Text Classification," 2020). Both the performance from BERT and XLNet are very close to the SOTA result. The results could be viewed from Figure 3.

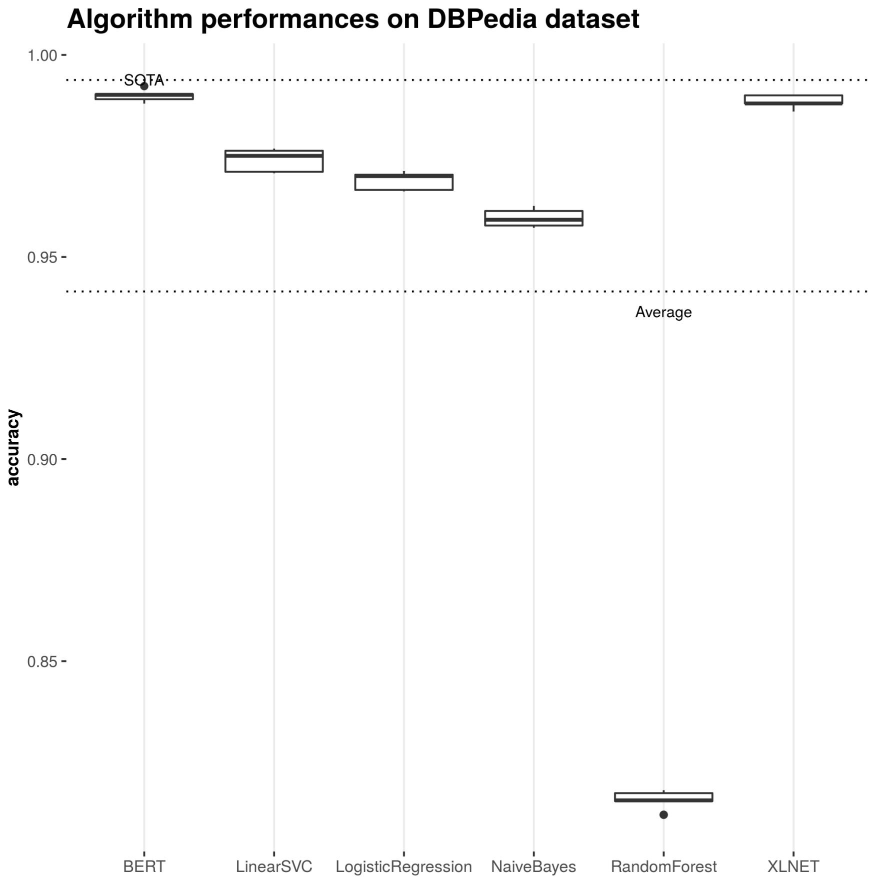


Figure 1 Models' performances on DBPedia dataset.

Contrasts in model performance are more discernable on Yelp review and Amazon review datasets. Contrary to a simple categorization of news into news topics, the task on the Yelp and Amazon datasets is to classify review texts into 1-5 scale sentiment scores with 1 denoting extremely dissatisfied and 5 extremely satisfied. Such datasets are conceptually more difficult to classify for two reasons. First, the difference between sentiments scores (e.g. 3 and 4) is more nuanced than the distance between the topic "politics" and "sports." The ML/DL models need to discover subtler textual cues to make the classification in such cases. Second, the dataset itself in nature is less structured in patterns because it is not annotated by a set group of annotators. In practice, the annotation or labeling of the training data is under the supervision of a predefined protocol to make sure the annotated data reflects the same underlying conceptual classification scheme. In practice, regardless of whether the annotation is conducted by a group of specified researchers or by the large group of crowdsourcers (e.g. MTurkers), there should be an explicit annotation protocol guiding the individual's annotation practice. However, such explicit annotation protocols do not exist in Yelp and Amazon review datasets. Different reviewers, on Yelp or Amazon, have different satisfaction thresholds to offer a rating score. This situation complicates and is blurred the decision boundary between scores, for example, a score of 4 vs 5. For the ML/DL models, they are likely learning how to give rating scores from different 'teachers' and each individual reviewer is one of those teachers. The model performances corroborate the inconsistent nature of the datasets as we can see that they are not at the same accuracy level as the AG News or DBPedia datasets. For the Amazon review dataset, the best performing model is XLNet with accuracy at around 0.59. The SOTA result is 0.6774 and it was achieved with XLNet (Yang et al., 2019) as well. We did not achieve the SOTA result by using the same XLNet was because they used the full training size but our training size is only 10000. Secondly, the SOTA result was trained on both the title and the contents from the training data. In our experiment, we only used the contents as training data without incorporating the title of each review to train the model. Thus, it is understandable that our models do not perform equally well as SOTA results without the key components of title of each review, even though they all used XLNet as the transfer learning architecture. The results of model performances on Amazon Review and Yelp Review datasets are presented in Figure 4 and 5.

DL models unanimously outperformed ML models

The second key finding is that our DL models consistently outperformed the ML models. The pattern is clear from Figure 5 . Across all of 5 datasets, DL models outperform ML models with larger differences between models in Amazon and Yelp review datasets and smaller differences in the AG News and DBPedia datasets. Within the DL category, XLNet significantly outperformed BERT across all of the datasets (except the Customer Complaint dataset where we did not run XLNet due to technical constraints with respect to RAM of the GPU). This finding is consistent with XLNet's original experiments when comparing XLNet's performances with BERT's.

Within the ML camps, LinearSVC, Logistic Regression and NaïveBayes perform at equivalent levels in most datasets (with the exception that NaïveBayes performed less satisfying on the Customer Complaint dataset). Random Forest underperformed compared to other ML counterparts across all datasets. Random Forest's subpar performance might be attributed to our selection of a set of parameters: n_estimators=200, max_depth=3. We could have improved the performance of Random Forest if we fine-tune those parameters instead of applying 'out-of-box' parameters. However, this practice belies the purpose of this research, that is to examine the model performances for the sake of applying them without an expertise of DL/ML in fine-tuning. Thus, if we were to start with some baseline models in practice, we might want to start with linear based models or NaiveBayes rather than Random Forests.

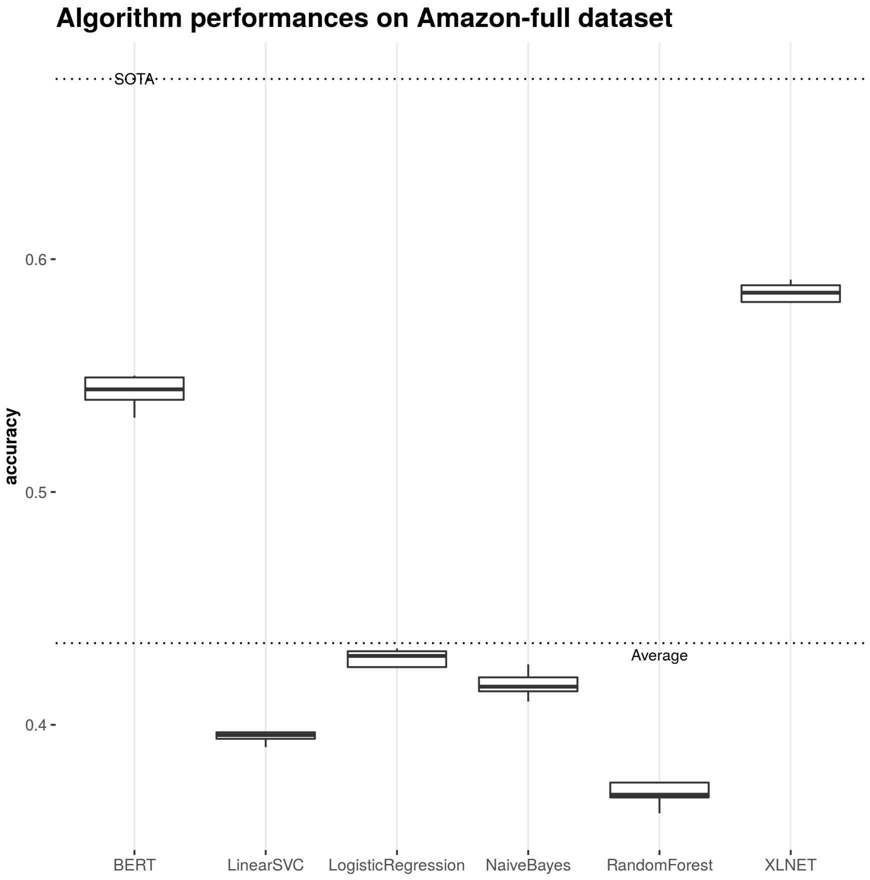


Figure 2 Models' performances on Amazon Review dataset.

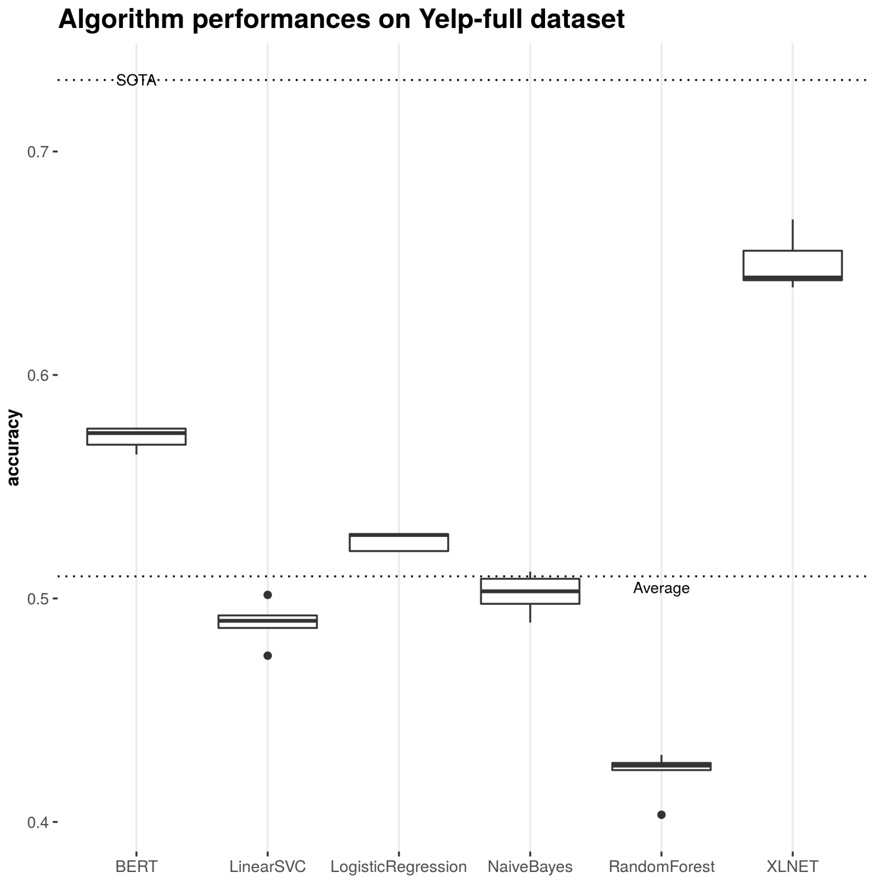


Figure 3 Models' performances on Yelp Review dataset.

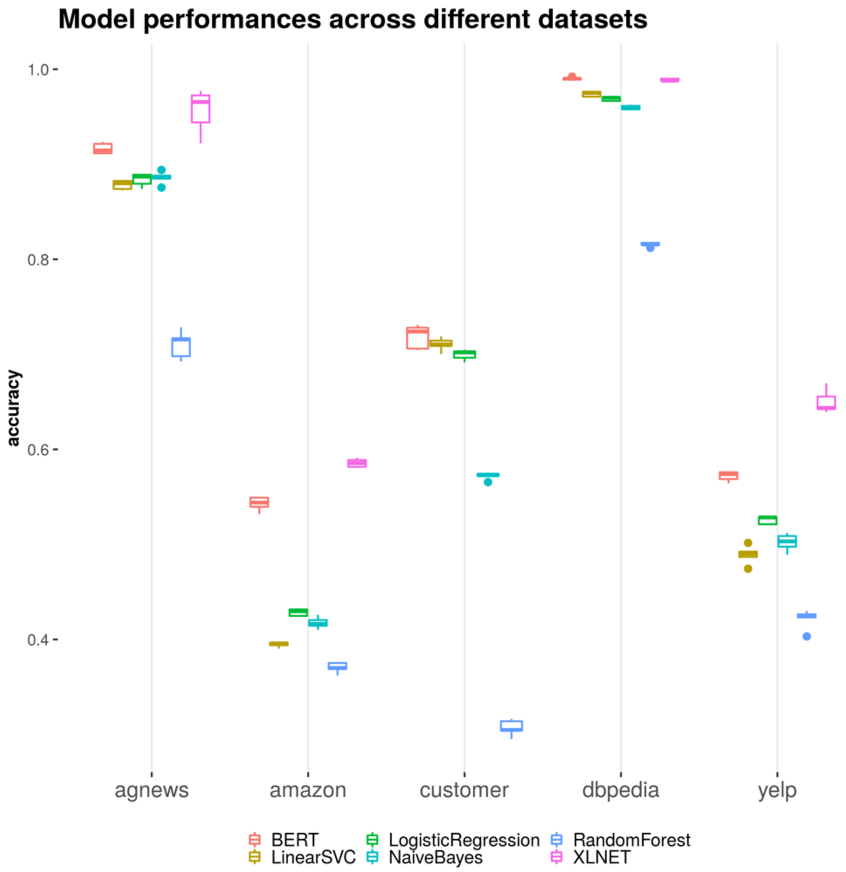


Figure 4 Models' performances across datasets in Stage 1 experiments.

Stage 2 Results

Cold Start for BERT in Performance

The finding that DL models outperform ML models is consistent. However, there were a few exceptions. . The first finding is that BERT has a cold start problem. BERT's performances were the one of the lowest when compared to other ML/DL models across 5 datasets, at the training size of 100. For example, see Figure 7 about BERT's performance on small sample sizes on Yelp Review dataset. This happens because our transfer learning strategy is to train a custom linear neural network ‘head’ on top of the ‘body’ of loaded weights. The weights were trained by the author of BERT at Google. Such a practice is prone to optimization issues from connecting the ‘body’ and the ‘head’, in this case, the transferred “fragile co-adapted” layers and our own custom linear layer (Yosinski, Clune, Bengio, & Lipson, 2014). However, such malperformance of BERT should not worry us much because it starts to pick up and excel as the training size increases. For AG News and DBpedia data (see Figure 8 and 9), BERT gained a phenomenal boost in performances even at the training size of 200, and BERT can excel or equal t ML models at training size of 5000. For the Yelp and Amazon Review dataset (see Figure 10), BERT does not have an explosive growth in accuracy as it does on AG News and DBPedia datasets, but it gradually climbs up in accuracy as the training size increases and tops the performances of ML models, on average, at the training size of 800 on the Yelp Review dataset and 1600 on the Amazon Review dataset.

It should also be noted that BERT’s DL counterpart, XLNet, does not suffer the cold start problem as BERT does. XLNet’s starting performance at the training size of 100 on average is the highest in AG News, DBPedia and Yelp Review datasets and at the comparable level with other models in Amazon Review dataset. This improvement over BERT may reflect the architectural design of the XLNet.

XLNet Distinctively Outperforms Others in Performance

XLNet is the current record holder for the miscellaneous NLP tasks ("Text Classification," 2020) and this pattern echoes throughout our experiments. Not only did

XLNet perform the best across datasets, they also do not require a large training size to get to an equivalent accuracy level as other models. For example, in the Yelp Review dataset, XLNet model yielded better results at a training size of 400 than other models at training sizes of 5000. Similarly, in the Amazon Review dataset, XLNet's performance at a training size of 800 is better than other models' performances at training sizes of 5000.

The DL models' Variances of Performance Worth Attention

One key contribution of this dissertation is to conduct experiments/simulations to investigate the variance of models' performances. This aspect of model performance is often overlooked in the literature as the research paradigm in ML/DL is to showcase the improvements of an algorithm by using a SOTA result they observed. A measure of the stability of model's performance is lacking.

The overall trend for all ML/DL models is that the larger the training size, the variance of model's performance would be smaller. However, there are exceptions for the DL models for moderately small training sizes (in the case of the experiments, the training size at 200, 400 and 800). At the smallest training size of 100, both DL models did not have enough data to learn from thus the performance tends to be primitive with a relatively small variance. As the training data grows to the moderately small sizes (e.g. 200, 400 and 800) the variances of both DL models tends to increase, and the variance change is more significant in BERT's performances than those of XLNet's. There is little research that taps into this situation in terms of variances of DL models under small training size. There are possible explanations for the increase of variance of DL model performances at a moderately small training size. The first has to do with the initialized model weights in the final linear layer. A lucky initialization could lead to faster and more stable downward direction in the loss function thus leading to a better accuracy. The second explanation has to do with a phenomenon called covariate shift. Covariate shift denotes a situation where the training input points and test input points follow different probability distributions but the conditional distributions of output values given input points are unchanged (Sugiyama, Krauledat, & MÃžller, 2007). For example, consider if we were to train a classifier to distinguish if an image object is a dog and the images in our training data are all featured with a black dog while in the testing/validation data the dogs are all white and brown. The covariate, in this case, the color of the dog, is different in training data and testing/validation data. Such a covariate shift might confuse the model/classifier into classifying the white dogs into objects other than dogs because there were no white dogs in our training data.

In a moderately small training size, the models have slightly more training data to 'learn' than the minimal training size (e.g. 100), which tends to increase the accuracy level on average. However, the moderately small training size could not cover the covariate space well thus causing the discrepancies between the training covariate space and the testing/validation covariate space. When we are fortunate to acquire a training set that resembles the covariate space of the testing set, our model could perform exceptionally well. But if we are unfortunate to have a 'good' covariate space to start with in our training data, our model performance might be compromised. At the moderately small training size, the discrepancies among the possible covariate space we get from our training set lead to a relatively larger variance in model performances. As the training size continue to increase, the covariate space tends to become wider to match the covariate space of the testing set thus leads to a smaller variance.

The Marginal Performance of the Model Diminishes

The first overall pattern we observe about models across different datasets is that gains in performance diminish as the training size increases and eventually plateaus. We also observe that the plateau arrives relatively later for DL models, suggesting that DL models could be further improved with more training data. This finding is consistent with the literature in DL/ML. This pattern could be viewed in Figure 9 about the summary of the Stage 2 experiments results.

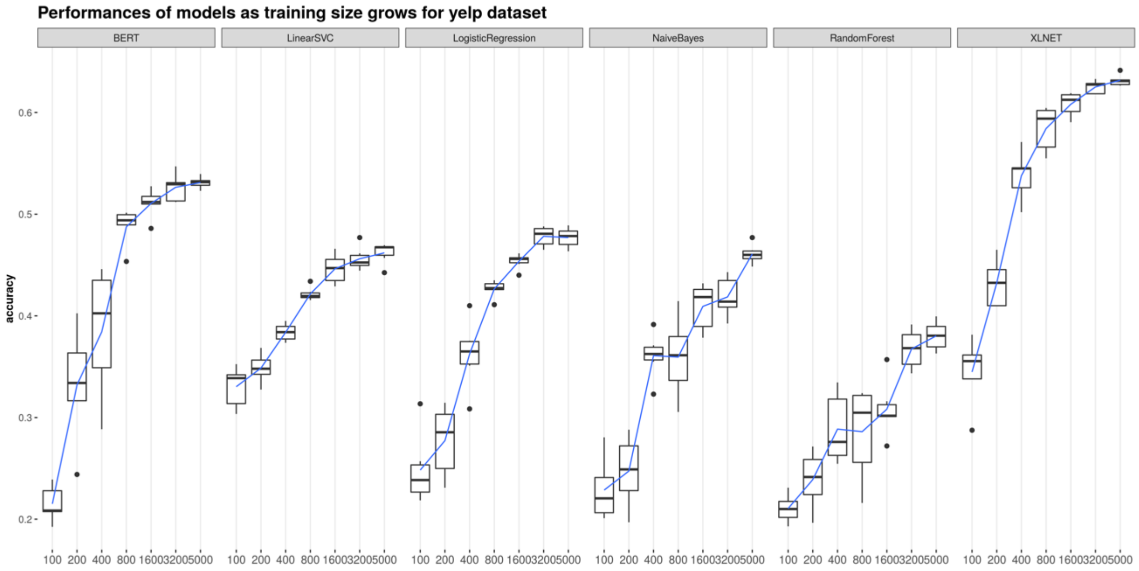


Figure 5 Model performances as training size grows on the Yelp Review dataset.

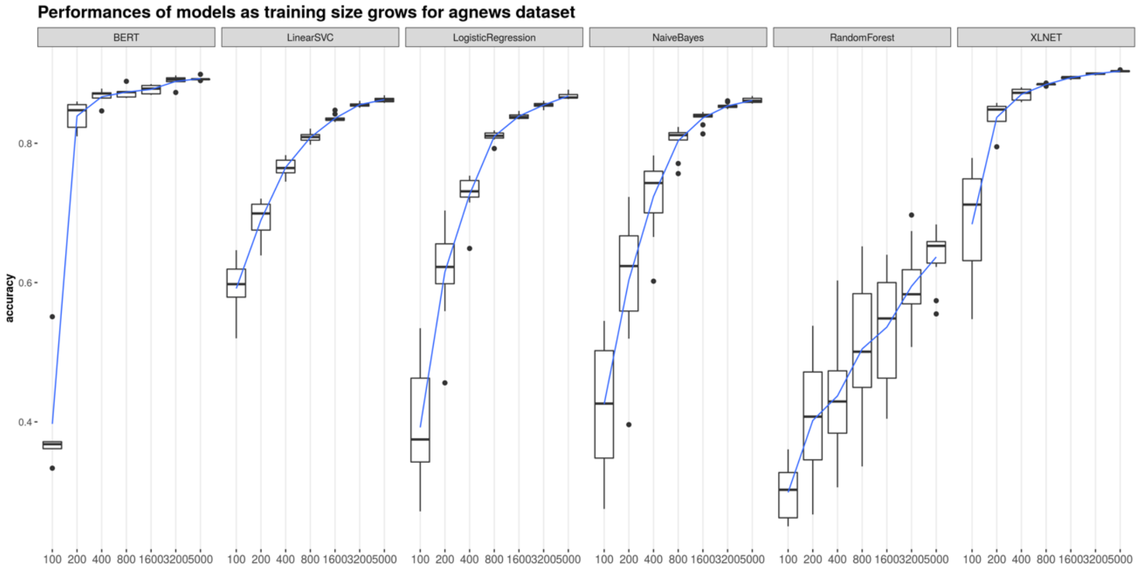


Figure 6 Model performances as training size grows on the AG News dataset.

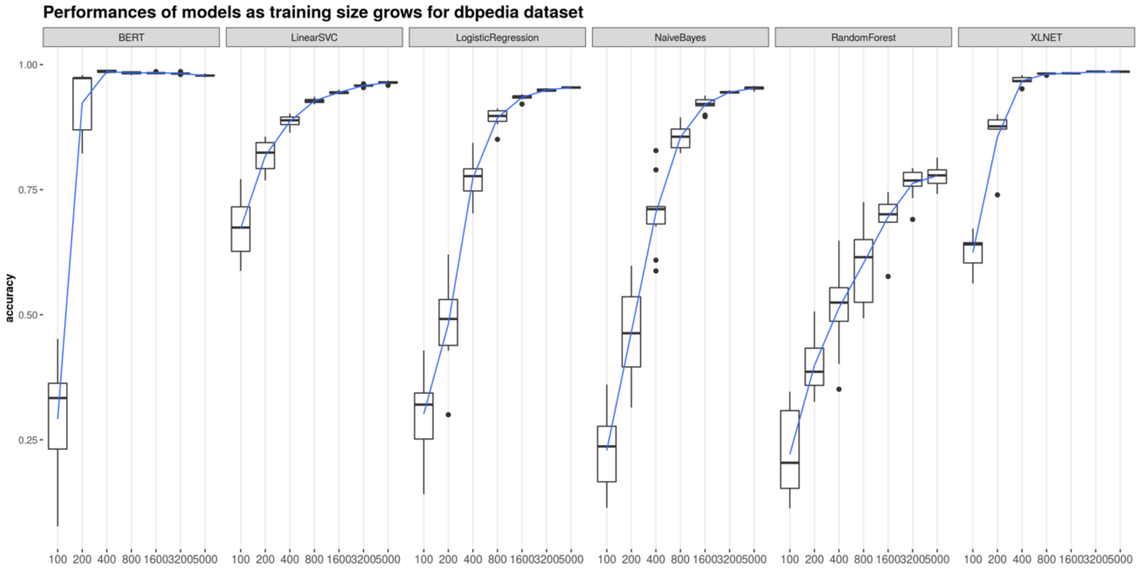


Figure 7 Model performances as training size grows on the DBPedia dataset.

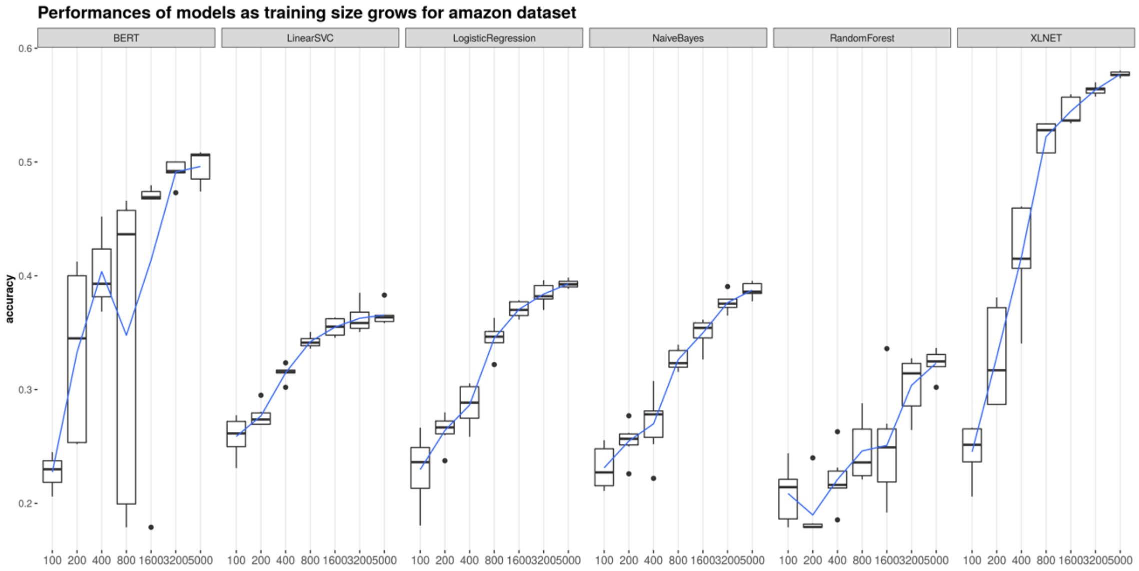


Figure 8 Model performances as training size grows on the Amazon Review dataset.

Stage 3 Results

The Compact Annotation Scheme Yields Better Results

Across all datasets, we have observed improved performances for the compactly annotated version of data. For example, in Yelp Review data with the original annotation from 1 to 5, even the best performing model XLNet could not exceed the 60% threshold. However, in the compactly annotated version, which only contains training data that were 1, 3 or 5, multiple DL/ML models approached 80% accuracy and XLNet even close to 90%. This result has been plotted as shown in Figure 12. This pattern also emerges in the Amazon Review dataset which has model performances at around 50% accuracy level. After removing the category of 2 and 4, the model performances escalated with BERT jumping to 70% accuracy and XLNet nearly 80%. The results for Amazon Review data could be viewed in Figure 13.

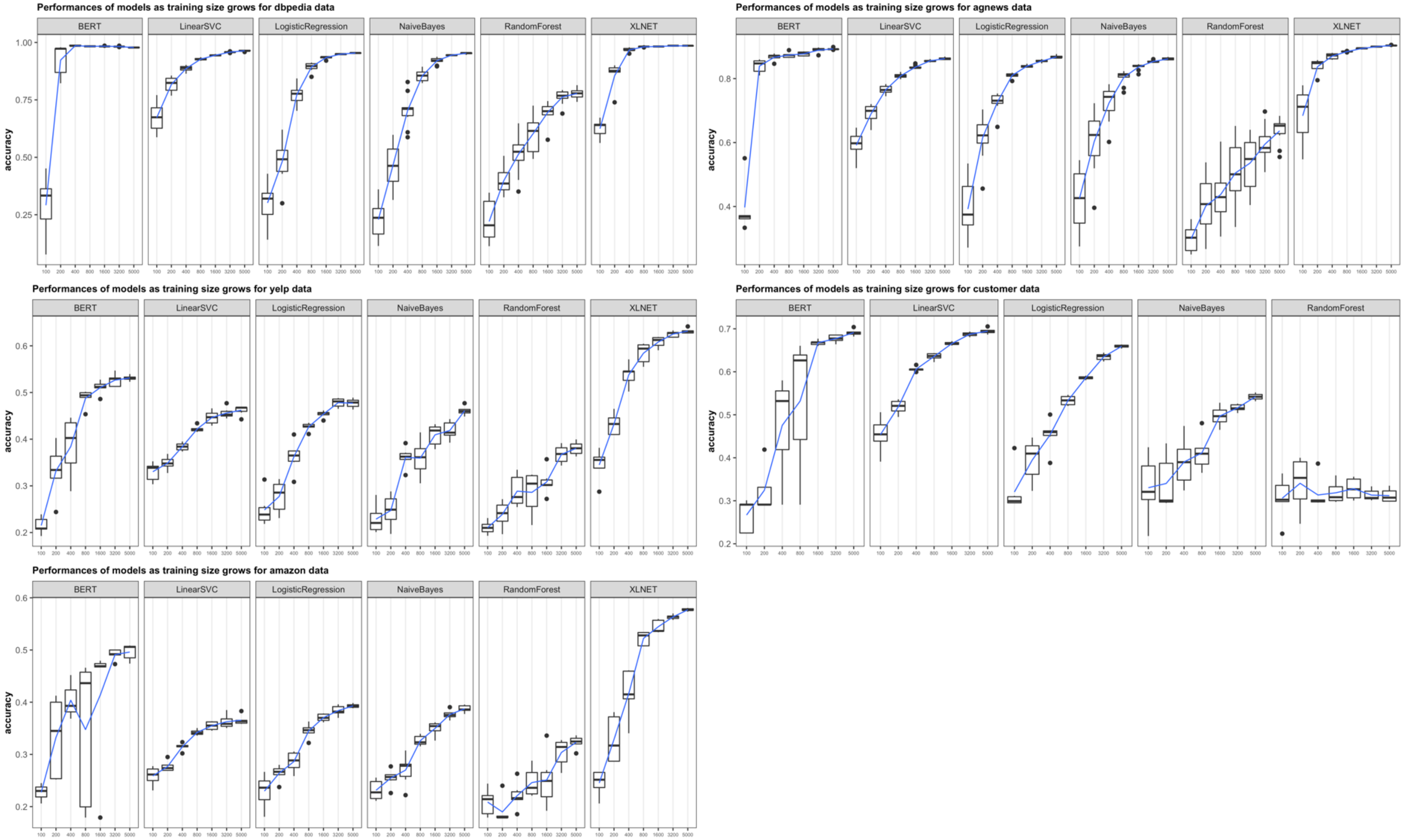


Figure 9 Model performances as training size grows across datasets.

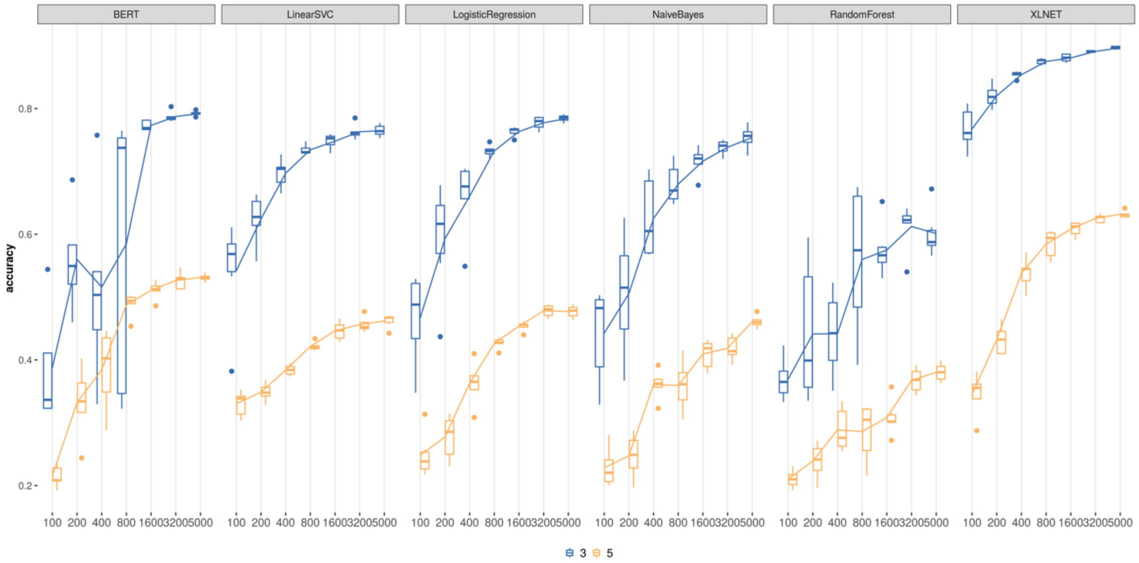


Figure 10 Model performances in compactly annotated scheme compared to the original annotation scheme in Yelp Review dataset.

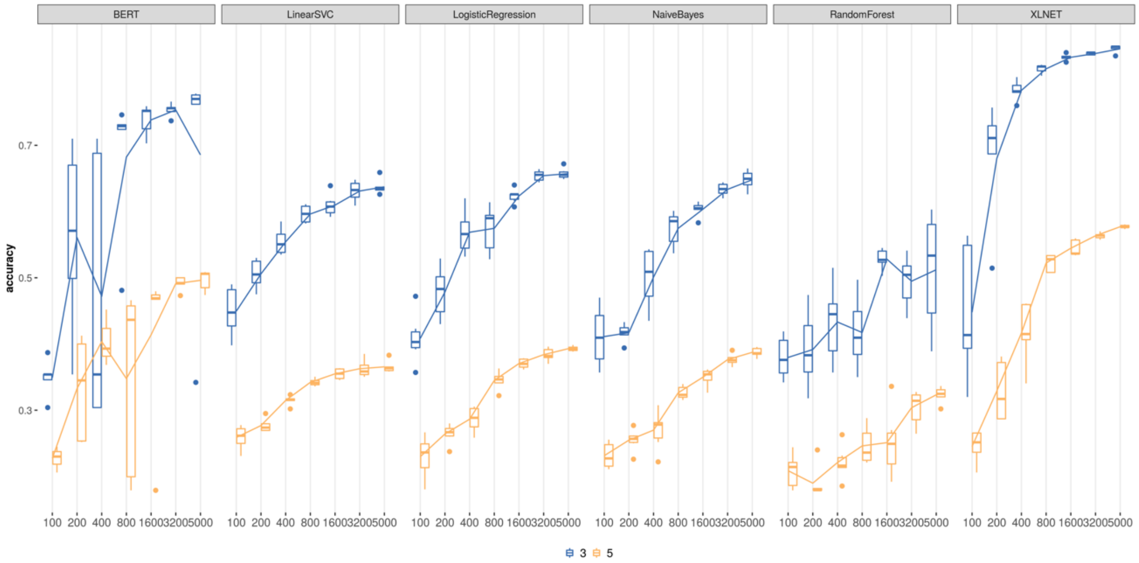


Figure 11 Model performances in compactly annotated scheme compared to the original annotation scheme in Amazon Review dataset.

Summarizing Results Across 3 Stages

Across the three stages of experiments, we have discovered that in terms of overall performances, DL models, BERT and XLNet outperform the ML models, with a minor exception that BERT has a cold start problem that it underperforms when the training size is extremely small like 100. As training sizes increases, BERT escalates rapidly and surpasses the ML models. Thus, our RQ2 has been answered by our experimental results.

The model performances are not uniformly the same across datasets. In cases where the classification task is relatively easier with more distinctive annotation schemes (such as AG News and DBPedia datasets to classify news categories), the difference between ML and DL's performances is relatively smaller. The differences become larger between DL and ML models when the datasets are harder in a way that demands more textual understanding rather than simply referencing the cues of key words, such as Yelp and Amazon Review data.

For RQ2, we have observed that XLNet consistently produces the best results when compared with the rest of models in all datasets, and surprisingly, almost across all training sizes. This means that unlike BERT, which may not be stable in its performances at small training sizes, XLNet outperforms the rest of models in almost all of the small training sizes, even though its variance may slightly increase in the moderately small training sizes (e.g. 200, 400). This suggests us that we should consider using the XLNet as our top priority in model selection and further fine-tuning even though when we have training size as small as 100. In contrast, we might be cautious in using BERT as our only resort when the training size is extremely small of moderately small (e.g. below 800) because of its problem of cold start and relatively large variance in small training sizes. What we might want to do is to ensemble BERT model with a XLNet model when the training size is not so small (e.g. larger than 800) to achieve better generalizability.

For RQ3, we found significant increases in model performances from all DL/ML models under the compactly annotated version of datasets. Such a pattern is consistent across all sizes of training data from 100 to 5000. XLNet saw a large escalation and still yielded the best results when compared with other models. The variance also decreases as the training size increase from 100. In contrast, BERT still suffers the variance increase during the moderately small training size.

Finally, what this dissertation reveals about XLNet's performance contrast sharply to the general impression about DL models. DL models are usually considered to be powerful if only provided with large size training data. Even though some studies have discovered positive cases where transfer learning models would perform well under small training sizes (e.g. Hanoz Bhathena, 2019; Howard & Ruder, 2018), few have investigated the stability of their performances. BERT has been found to be unstable given a small or moderately small training size while XLNet consistently generates the best accuracy compared to other models across all levels of small training sizes. What the results suggest is that, in practice, XLNet could be directly used as a baseline model for future fine-tuning. This carries great practical implementation for social science researchers in that they could use transfer learning models to achieve great accuracy without annotating a large training set, also without understanding much knowledge or experience about model training in DL.

Results about Study Two: Case Experiment in Frame Analysis

Frame Analysis: Harder Textual Classification Problem

Frame analysis is considered a harder task than general textual classification problems. We provided two main reasons in the introduction why frame analysis is conceptually more complicated than textual classification tasks. First, frames are more contextually dependent which injects more human interpretation. One key frame in the current analysis is ‘Leisure banter’ which uses GPT2 to generate texts from text prompts for fun or making fun of the GPT2’s possible usages. For example, a tweet such as “Assuming there are bots that trade based on sentiment analysis of social networks, I wonder how many deep learning script kiddies out there are trying to manipulate cryptocurrencies with GPT2-jr-generated fake news” is deemed a ‘Leisure banter’ because it presents a whimsical view on the usages of GPT2. However, to correctly categorize such an example to the Leisure Banter frame, we need background knowledge about the function of GPT2 (to generate texts) and how cryptocurrencies work. Such background knowledge is hard to capture because it involves human interpretation and the training data could not sufficiently transmit the background knowledge to the model., The result would be a poor performance on this example tweet. The second reason is that the annotation schemes in frame analysis are not usually captured by a a single dimensional as is the case with general textual classification tasks. This pattern has been reflected in the annotation scheme in this study. The frames we extracted are: “Question purpose”, 'Informational', 'Tech excitement', 'Concern', 'Control', and 'Support decision.' These frames are multi-dimensional as they are focusing on different subjects or aspects of the event. For example, “Question purpose” and “Support decision” are specific frames targeting the subject OpenAI, the organization that developed GPT2, while “Tech excitement” or “Informational” mainly focus on the technology aspect of GPT2. The multi-dimensional nature of frames in frame analysis make it more nuanced in differences thus requiring deeper understanding of meaning of texts.

Not surprisingly, this study found that the models do not perform as well as in our earlier analyses of ‘easy’ datasets such as DBPedia or AG News (see Figure 14). The accuracy is about the same level as the Yelp and Amazon review datasets. Even though the reason why the accuracy is not that satisfying is different across these datasets, it suggests that these datasets are more challenging for algorithms to understand and classify.

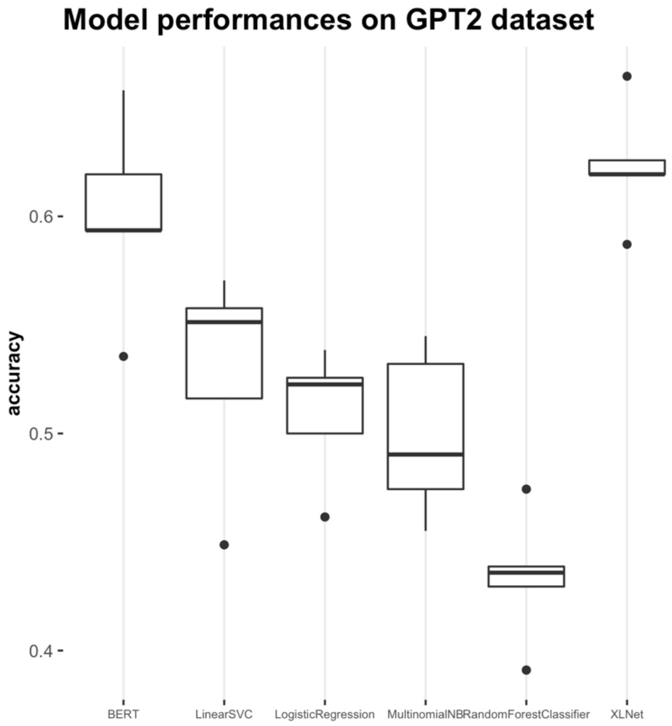
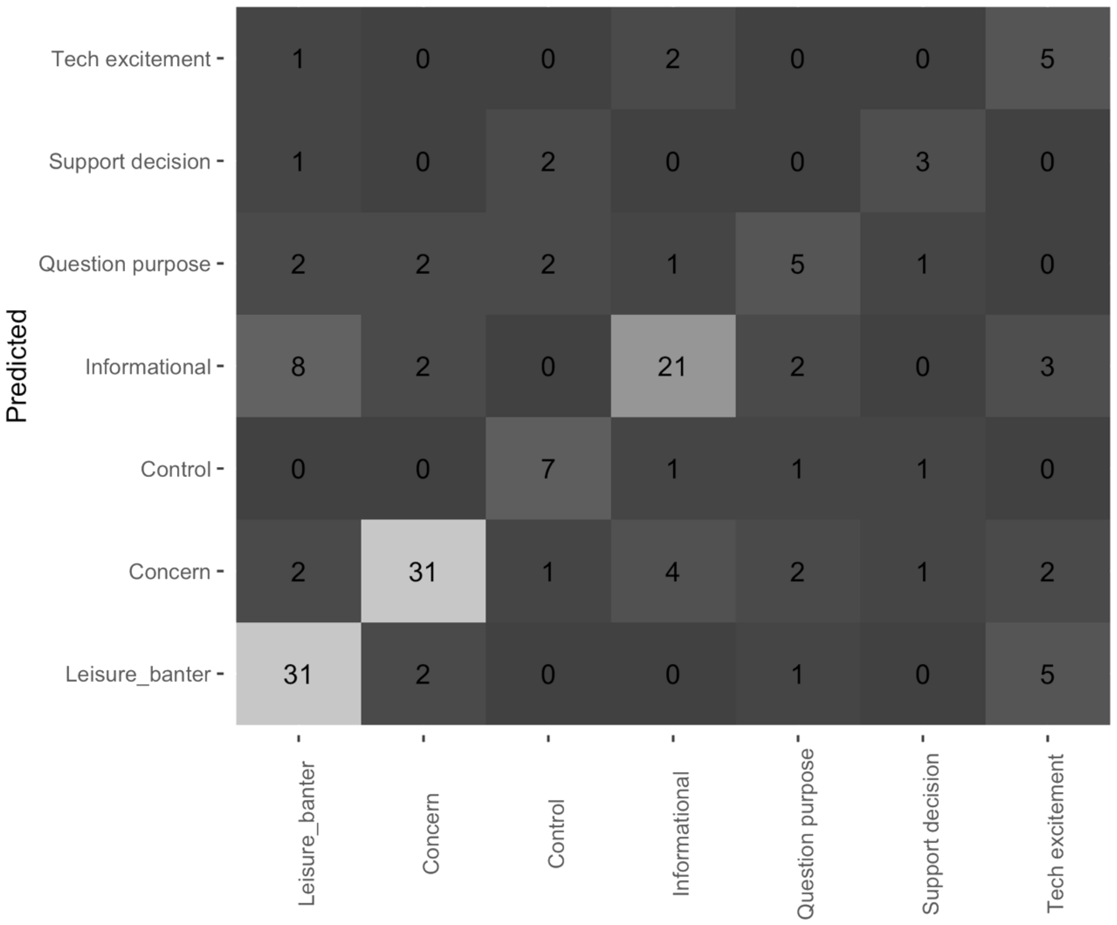


Figure 12 Model performances on frame analysis

The best accuracy level is selected from XLNet to illustrate more details about the classifier. Table 7 presents the micro and macro indices. The corresponding confusion matrix for the 7-frame version could be viewed in Figure 15.

Table 6 Classification report of XLNet on GPT2 dataset (7-frame)

	Precision	Recall	F1-score	Support
Leisure_banter	0.79	0.69	0.74	45
Informational	0.58	0.72	0.65	29
Concern	0.72	0.84	0.77	37
Control	0.70	0.58	0.64	12
Question_purpose	0.38	0.45	0.42	11
Support_decision	0.50	0.50	0.50	6
Tech_excitement	0.62	0.33	0.43	15
Accuracy	-	-	0.66	155
Macro average	0.62	0.59	0.59	155
Weighted average	0.67	0.66	0.66	155



Predicted	Leisure_banter	Concern	Control	Informational	Question purpose	Support decision	Tech excitement
Tech excitement	1	0	0	2	0	0	5
Support decision	1	0	2	0	0	3	0
Question purpose	2	2	2	1	5	1	0
Informational	8	2	0	21	2	0	3
Control	0	0	7	1	1	1	0
Concern	2	31	1	4	2	1	2
Leisure_banter	31	2	0	0	1	0	5

Figure 13 Confusion matrix of XLNet's performance on GPT2 frame analysis

Model Performances Improved from a Compact Annotation Scheme

Due to the small training size, some of our frames have too few presences in the training dataset (80% of the whole dataset) due to the imbalance of frames. The algorithms may not ‘learn’ enough from those training data to extract patterns about all 7 frames. Thus, we created a compact annotation scheme for the frame analysis dataset collapsing 5 frames into 4 frames 4 frames ("Question purpose" and “Support decision”, "Informational" and 'Tech excitement') into 2 frames ("Decision related" and "Tech information"). The reason for choosing these frames to collapse is because they were not prevalent in the dataset (e.g. Support decision only has 33 cases) and these frames operate on the same aspect. For example, Support decision and Question purpose frames relate to moral judgements of OpenAI's motive for developing such a powerful model but not releasing the full model. After collapsing the frames in to 5, the model performances have improved compared with the original annotation scheme (7-frame), as illustrated from the following figure.

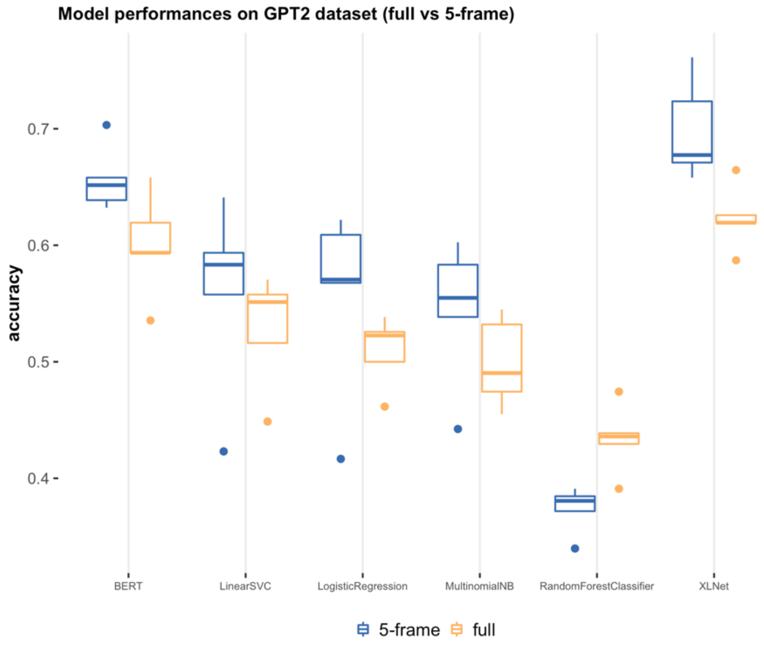


Figure 14 Model performances on GPT2 dataset comparing two annotation schemes

Further, the more detailed micro and macro indices of the best performing XLNet model for the 5-frame version of GPT2 dataset is provided in Table 8. The corresponding confusion matrix is presented in the following figure.

Table 7 Classification report of XLNet on GPT2 dataset (5-frame)

	Precision	**Recall**	**F1-score**	**Support**
Tech information	0.80	0.80	0.80	50
Leisure_banter	0.78	0.82	0.80	51
Concern	0.79	0.76	0.77	29
Control	0.71	0.71	0.71	14
Decision related	0.44	0.36	0.40	11
Accuracy	-	-	0.76	155
Macro average	0.70	069	0.70	155
Weighted average	0.76	0.76	0.76	155

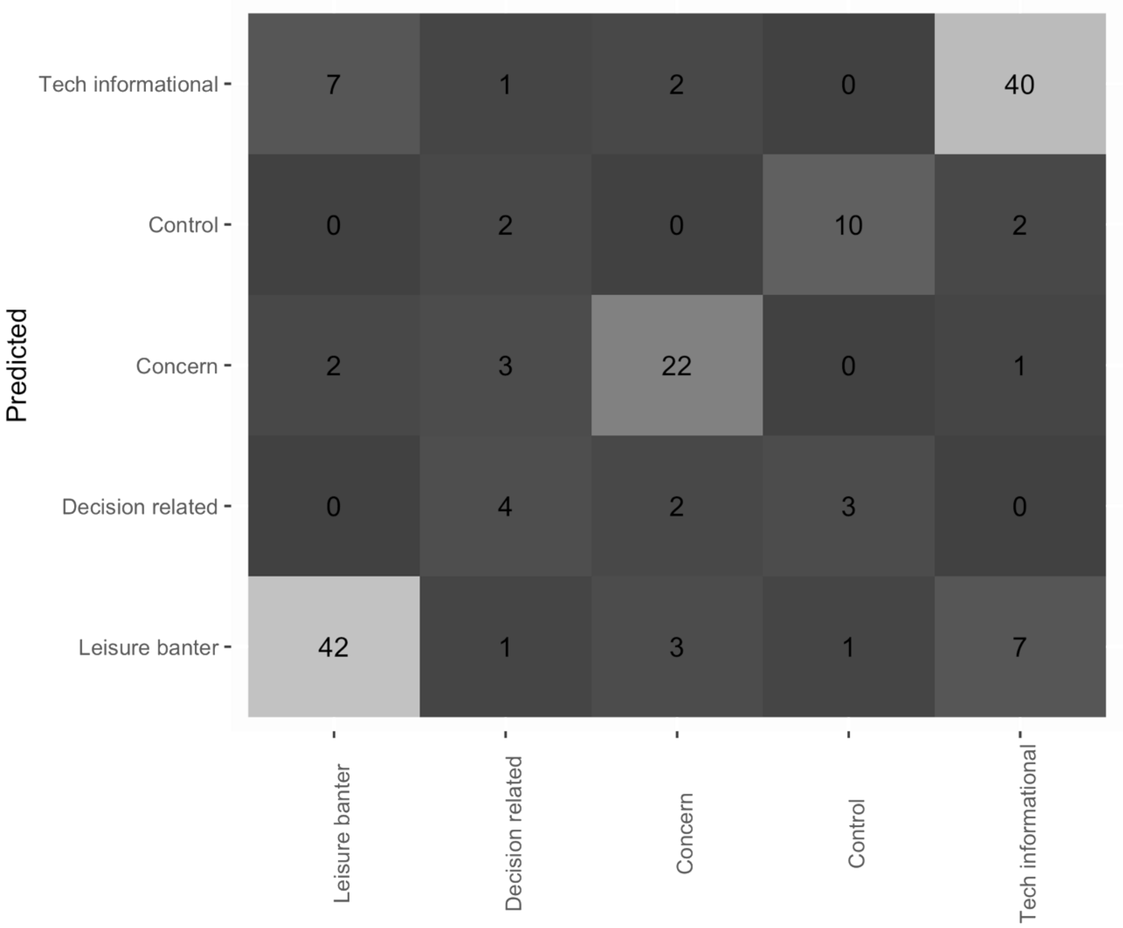


Figure 15 Confusion matrix of XLNet on GPT2 (5-frame version)

XLNet Still Leads the Performance

This result is consistent with our findings in study 1 that XLNet performs better than other models. XLNet leads the performances in both the 5-frame and 7-frame datasets. The variance is larger on the 5-frame version for XLNet but this may reflect the fact that the random sampled testing data have include less frequent frames, such as ‘Control’ or ‘Decision related.’ The classifier still suffers when it encounters these less frequent frames thus if by chance more of them appear in our testing set, the corresponding accuracy would be lower and vice versa. This phenomenon, the larger variance is not comparable to that in study 1 where we have abundant testing data (e.g. of size 2000 in the second stage) with balanced categories. However, further investigations would still be valuable to see if this trend continues in a larger number of experiments or simulations or other datasets.

This study answers the RQ1 which asks the performances of transfer learning models on a frame analysis task. From both conceptual and empirical perspectives, frame analysis is a more complicated task than general textual classifications. This has been supported by our finding that the ML/DL algorithms achieved the same level as other hard textual classification datasets. With a training size of 624, the best performing model XLNet achieved 66% accuracy. When we collapse 4 of frames into 2 larger frames (resulting in 5 frames overall), we get greater accuracy (76%). This result is promising for practical reasons. Even with only around 600 training data, we could get above 70% accuracy on 5 frames that do not operate on single dimensional meanings. Such performances of the algorithm could be iteratively used in building a better classifier for the practical purposes. For example, the classifier we ended up with could serve as the first generation model put into practice to make the classification. For the frames with high confidence (e.g. with high precision, recall, sensitivity, specificity etc.) we could place more trust in the model's classification results; for the frames with low confidence, we can manually recheck the model's classification and fix the mistakes by relabeling those data cases. We can incorporate these results into our training data and retrain the algorithm again as the second-generation classifier for future tasks. The process could be iterated, with more training data, until it achieves a satisfying level.

The results from this study suggests that transfer learning models, such as BERT and XLNet greatly surpassed BOW ML models and should be used as a benchmark before further parameter fine-tuning or model ensembling to achieve better results.

CHAPTER IV

CONCLUSIONS AND LIMITATIONS

This chapter summarizes the results of two studies, illustrates the significance and contribution of this dissertation, and finally addresses the limitation of the study which shed light on future research.

Summary of Results

This dissertation aimed to answer two interrelated questions. The first is how does transfer learning perform on textual classifications on small training sizes and the second is how could transfer learning help communication scholars conducting frame analysis. These two questions are closely related because in practice, frame analysis datasets are annotated from scratch and frame analysis is no different from other textual classification tasks from the perspective of algorithms. Solving these two questions would offer us holistic guidance for conducting textual classification, or more specific frame analysis in situations involving model selection, performance evaluations and annotation strategies.

Study 1 reveals that ML/DL model performances vary greatly depending on the specific classification task and datasets. Most ML/DL models achieve above 95% accuracy on balanced datasets with training sizes of 800. However, even the SOTA does not exceed 70% accuracy on more complicated datasets such as the Amazon review dataset even with training sizes of 3,000,000. When we evaluate model performances, we should consider the nature of the dataset and how the categories were annotated. This leads to another finding that different annotation schemes lead to different model performances. We have achieved better results on all datasets, under the same training size but with smaller number of categories. This boost of performance is especially distinctive in the Yelp and Amazon datasets when the nuanced category of 2 and 4 are waived from the data.

XLNet, the recent model holding SOTA records, proves to be the first choice in terms of model selection. XLNet has been found to provide the best performances for all datasets, even across small training sizes, from 100 to 5000, with comparable variance in performance. In contrast, even though BERT would provide better results than ML models, it suffers the cold start problem, meaning that its performance were worse than even ML models at extremely small training sizes. Also, the variance of BERT tends to increase (rather than decrease) when moving to moderately small training sizes (e.g. 400, 800). This increase in variance could possibly be attributed to the initialization of random weights or covariate shift. As the training size increases, BERT's performance climbs up and exceeds those of ML models but no better than XLNet across different training sizes.

The results of experiments in combination counters the general impression that DL models could not achieve satisfying results without a gigantic training size. The results showed that with transfer learning models, we could get better results from a training size as small as 100 across different datasets with different annotation schemes. Given that transfer learning is still in fast development, we would expect to see more powerful and stable models for applied researchers to use in the near future.

In study 2, we put ML/DL models into a test in frame analysis. The comparison of model performance is consistent with our findings in study 1. At a training size of 600, our transfer learning models outperformed ML models and XLNet again yields the best accuracy. The nature of frame analysis, that it is more dependent on human interpretation and annotated in multiple dimensions of meaning, make frame analysis more challenging than general textual classification tasks. This has been reflected in the accuracies of models in that our best model from XLNet only produces around 60% accuracy on average. With a change into a more compact annotation scheme, the performances increased to above 70% for XLNet. This result could be considered promising given a training size of merely 600 with 5 imbalanced frames, and more importantly, with each training data point less than 280 characters. Without a deep understanding of the contextual knowledge of the dataset, a human classifier might not perform better than the transfer learning models.

Transfer learning could be iteratively used to achieve better performances, with the help of relabeling from researchers. This practice could greatly expedite the research process without having researchers blindly annotating a gigantic training set without knowing the marginal gain of the model performances.

Contribution of the Dissertation

This dissertation is, to the best of my knowledge, the first comprehensive study that experiments and investigates the transfer learning models' performances under small training sizes and across different benchmark datasets, and that applies transfer learning to frame analysis. The results will not only benefit research in frame analysis, but textual analysis in general in terms of research procedure such as model selection, performance evaluation, and annotation strategies. The annotated frame analysis dataset, GPT2 dataset, will also be open sourced for future research to refer to as a benchmark. The scripts used to conduct the comprehensive experiments is also open sourced for peers to replicate and reuse for future research. Finally, the visualization app for this dissertation is also hosted for future researchers to refer to. All the results could be seamlessly applied to communication research when researchers need to apply large volume of textual data to discover insights.

Limitations

This dissertation has successfully answered the research questions. However, there are some notable limitations in terms of the methods and research procedures.

First, we used the default, out-of-box, configurations of those ML/DL models. Different ML/DL models are equipped with different parameters which in combination yield different performances. Certain combinations of parameters might provide better model performances than other combinations. Because a goal of this research is to offer guidance for applied researchers who may not possess ML/DL knowledge, thus this dissertation adopted those DL/ML models in their default states as most applied researchers would, without further fine-tuning which requires programming skills and ML/DL field knowledge and experience. Thus, the optimal performances of those models may not be the same as the results of this dissertation. For example, for our

Random Forest model, we set its default as n_estimators=200, max_depth=3. Random Forest model performance might be improved if we increase the number of estimators (n_estimators) or the depth of each tree (max_depth). To get a more thorough understanding of the models' performances, future research might find out an optimal set of parameters for each model and then resample the training and testing datasets to get an estimate about the accuracy and stability of performances.

Second, this research only resampled a small number of times for each model, on each dataset, on each training size. For example, due to the constraint of time and computing resources, we only resampled 5 times for each transfer learning model on each dataset on each training size. Such a small number of repetitions would not offer us a robust statistical test for the comparison of model performances, nor a robust estimation of the variance of model performances. Further, due to the hardware constraint, we did not apply XLNet on the Customer Complaint dataset. In future research, a more thorough experiment with larger repetition size across more datasets is highly expected to further reveal the pattern of transfer learning models on small training sizes.

What is more, this research adopts the GPT2 controversy as the context of frame analysis. However, we did not harvest as much valid data as planned in the earlier stage. The moderately small data size does not afford the opportunity to further break down model performance across different small training sizes similar to what we did in the second stage in study 1. The small size of data is also compounded by imbalances of frames, affecting the final accuracy estimates. Future research on transfer learning on frame analysis might adopt a larger dataset to investigate in greater detail the application of transfer learning models.

Finally, transfer learning research is growing at a very fast pace and there are already numerous variants of BERT based models such as Roberta, DistillBERT, XLM, and XLNet. This research only adopts two of its most representative kinds: BERT brought the most revolutionary change in transfer learning architectures and XLNet being the SOTA record holder in multiple tasks. However, little is known about other transfer learning models' performances on small training sizes, or frame analysis. In the future, we would expect to see a more comprehensive research which put more transfer learning models in experiments.

REFERENCES

Aaldering, L., & Vliegenthart, R. (2016). Political leaders and the media. Can we measure political leadership images in newspapers using computer-assisted content analysis? *Quality & Quantity, 50*(5), 1871-1905.

Alammar, J. (2018). Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention). Retrieved from https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data *The semantic web* (pp. 722-735): Springer.

Baumer, E., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (2015). *Testing and comparing computational approaches for identifying the language of framing in political news.* Paper presented at the Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Boon, M. L. (2019). *Computational Approaches to Detection of Narrative Frames in News.* Northwestern University.

Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The end of framing as we know it… and the future of media effects. *Mass Communication and Society, 19*(1), 7-23.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2015). *The media frames corpus: Annotations of frames across issues.* Paper presented at the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci., 10*, 103-126.

Copeland, M. (2016). What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?. Retrieved from https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

David, C. C., Atun, J. M., Fille, E., & Monterola, C. (2011). Finding frames: Comparing two methods of frame analysis. *Communication Methods and Measures, 5*(4), 329-351.

Dettmers, T. (2018). TPUs vs GPUs for Transformers (BERT). Retrieved from https://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51-58.

Entman, R. M., Matthes, J., & Pellicano, L. (2009). Nature, sources, and effects of news framing. *The Handbook of Journalism Studies*, 175-190.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., & Tsvetkov, Y. (2018). Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. *arXiv preprint arXiv:1808.09386*.

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology, 95*(1), 1-37.

Gerlach, N. A. (2016). From outbreak to pandemic narrative: Reading newspaper coverage of the 2014 Ebola epidemic. *Canadian Journal of Communication, 41*(4).

Gitlin, T. (1980). *The world is watching: Mass media in the making and unmaking of the new left*. Berkeley: University of California Press.

Hanoz Bhathena, R. M. M. (2019). *Deep (Transfer) Learning for NLP on Small Data Sets.* Paper presented at the NVIDIA GTC conference.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity, 51*(6), 2623-2646.

Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification.* Paper presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research, 30*(3), 411-433.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., . . . Auer, S. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*(2), 167-195.

Li, S. (2018). Multi-Class Text Classification with Scikit-Learn. Retrieved from https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f

Lin, F.-r., Hao, D., & Liao, D. (2016). *Automatic Content Analysis of Media Framing by Text Mining Techniques.* Paper presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS).

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction. *International Journal of Communication, 13*, 4000-4020.

Lind, R. A., & Salo, C. (2002). The framing of feminists and feminism in news and public affairs programs in US electronic media. *Journal of Communication, 52*(1), 211-228.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254-277.

Matthes, J. (2009). What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly, 86*(2), 349-367.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication, 58*(2), 258-279.

Mcneela, D. (2017). The Universal Approximation Theorem for Neural Networks. Retrieved from https://mcneela.github.io/machine_learning/2017/03/21/Universal-Approximation-Theorem.html

Merity, S., Keskar, N. S., Bradbury, J., & Socher, R. (2018). Scalable Language Modeling: WikiText-103 on a Single GPU in 12 hours. Retreived from https://systemsandml.org/Conferences/2019/doc/2018/50.pdf

Nielsen, M. (2019). A visual proof that neural nets can compute any function. Retrieved from http://neuralnetworksanddeeplearning.com/chap4.html

Odijk, D., Burscher, B., Vliegenthart, R., & De Rijke, M. (2013). *Automatic thematic content analysis: Finding frames in news.* Paper presented at the International Conference on Social Informatics.

Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication, 10*(1), 55-75.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Perez, C. E. (2016). Why Deep Learning is Radically Different from Machine Learning. Retrieved from https://medium.com/intuitionmachine/why-deep-learning-is-radically-different-from-machine-learning-945a4a65da4d

Powell, K. A. (2011). Framing Islam: An analysis of US media coverage of terrorism since 9/11. *Communication Studies, 62*(1), 90-112.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Reese, S. D. (2007). The framing project: A bridging model for media research revisited. *Journal of communication, 57*(1), 148-154.

Rother, K., & Rettberg, A. (2018). Ulmfit at germeval-2018: A deep neural language model for the classification of hate speech in german tweets. Retrieved from https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/8490/file/Ruppenhofer_Siegel_Wiegand_GermEval2018_Proceedings.pdf#page=119

Ruder, S. (2018a). NLP's ImageNet moment has arrived. Retrieved from http://ruder.io/nlp-imagenet/

Ruder, S. (2018b). Tracking the Progress in Natural Language Processing. Retrieved from http://ruder.io/tracking-progress-nlp/

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication, 50*(2), 93-109.

Shah, D. V., Watts, M. D., Domke, D., & Fan, D. P. (2002). News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly, 66*(3), 339-370.

Shih, T.-J., Wijaya, R., & Brossard, D. (2008). Media coverage of public health epidemics: Linking framing and issue attention cycle toward an integrated theory of print news coverage of epidemics. *Mass Communication & Society, 11*(2), 141-160.

Shim, J., Park, C., & Wilding, M. (2015). Identifying policy frames through semantic network analysis: an examination of nuclear energy policy across six countries. *Policy Sciences, 48*(1), 51-83.

Sugiyama, M., Krauledat, M., & MÃžller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research, 8*(5), 985-1005.

Tankard Jr, J. W. (2001). The empirical approach to the study of media framing *Framing public life* (pp. 111-121): Routledge.

Terman, R. (2017). Islamophobia and media portrayals of Muslim women: A computational text analysis of US news coverage. *International Studies Quarterly, 61*(3), 489-502.

Text Classification. (2020). Paper with code*.* Retrieved from https://paperswithcode.com/task/text-classification

Tian, Y., & Stewart, C. M. (2005). Framing the SARS crisis: A computer-assisted text analysis of CNN and BBC online news reports of SARS. *Asian Journal of Communication, 15*(3), 289-301.

Tracy, S. J. (2019). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact*. John Wiley & Sons.

Tucker, L. R. (1998). The framing of Calvin Klein: A frame analysis of media discourse about the August 1995 Calvin Klein jeans advertising campaign. *Critical Studies in Media Communication, 15*(2), 141-157.

Uszkoreit, J. (2017). Transformer: A Novel Neural Network Architecture for Language Understanding. Retrieved from https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

van der Burgh, B., & Verberne, S. (2019). The merits of Universal Language Model Fine-tuning for Small Datasets--a case with Dutch book reviews. *arXiv preprint arXiv:1910.00896*.

Walter, D., & Ophir, Y. (2019). News Frame Analysis: An Inductive Mixed-method Computational Approach. *Communication Methods and Measures, 13*(4), 248-266.

Warstadt, A. S., Amanpreet; Bowman, Samuel R. (2018). Neural Network Acceptability Judgments. *arXiv preprint arXiv:1805.12471*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). *Xlnet: Generalized autoregressive pretraining for language understanding.* Paper presented at the Advances in neural information processing systems.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). *How transferable are features in deep neural networks?* Paper presented at the Advances in neural information processing systems.

Yousaf, S. (2015). Representation of Pakistan: A Framing Analysis of the Coverage in the US and Chinese News Media Surrounding Operation Zarb-e-Azb. *International Journal of Communication, 9*, 23.

Zeiler, M. D., & Fergus, R. (2014). *Visualizing and understanding convolutional networks.* Paper presented at the European conference on computer vision.

Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level convolutional networks for text classification.* Paper presented at the Advances in neural information processing systems.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1-4), 43-52.