Accelerating the Design Process Through Natural Language Processing-based Idea Filtering

by

Kristen M. Edwards

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Department of Mechanical Engineering
August 12, 2022

Certified by . . . .
Faez Ahmed
Assistant Professor of Mechanical Engineering
Thesis Supervisor

Accepted by . . . .
Nicolas Hadjiconstantinou
Chairman, Department Committee on Graduate Theses

Accelerating the Design Process Through Natural Language Processing-based Idea Filtering

by

Kristen M. Edwards

Submitted to the Department of Mechanical Engineering
on August 12, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Mechanical Engineering

Abstract

The following treatise explores the use of natural language processing to accelerate the design process in various domains by automating idea filtering. During the design of products or programs, a bottleneck often arises when experts need to filter through an exorbitant number of ideas, searching for which ones are most innovative, creative, relevant, or any other number of subjective characteristics. We observe this bottleneck when filtering entrepreneurial ideas for innovation, when filtering early-stage design concepts for creativity and usefulness, and when filtering literature for relevance toward policy-informing evidence syntheses. Motivated by the common challenge of idea filtering in various design domains, my research explores the use of natural language processing (NLP) for accelerating design through automated idea filtering. My team and I investigate the possibility of using machine learning to predict expert-derived creativity assessments of design ideas from more accessible non-expert survey results. We demonstrate the ability of machine learning models to predict design metrics from the design itself and textual survey information. Our results show that incorporating NLP improves prediction results across design metrics, and that clear distinctions in the predictability of certain metrics exist. We go on to explore the effectiveness of using NLP to accelerate literature screening for designing evidence-based policies and programs. In this research, we introduce the use of transformer models for idea filtering and evaluation. Transformer-based models have produced state of the art results in NLP tasks such as language translation, question answering, reading comprehension, and sentiment analysis. Our results show that the fine-tunable transformer-based models achieve the highest text classification accuracy, 79%, therefore accurately evaluating and filtering our textual dataset. Furthermore, we observe that the model accuracy improves with the training data size with diminishing marginal effect. The findings can facilitate informed decision making regarding the trade-off between model accuracy and manual labeling efforts, increasing efficiency. After demonstrating the effectiveness of using NLP to accelerate literature screening, we aimed to next decrease the level of effort required of expert reviewers to generate training data. To train an idea-filtering model, we need a labeled dataset of ideas, however obtaining labeled data is a challenge for engineering and design applications, as experts are expensive and, therefore, expert labeled datasets are as well. We were motivated to explore avenues to decrease the size of training data needed by using active learning (AL). AL is the concept that a machine learning algorithm can perform better with less training data if it is allowed to choose the data from which it learns. We find that data selection techniques that incorporate active learning result in higher F1 scores, a more balanced training set, and fewer necessary labeled training instances. These results suggest that active learning is effective in decreasing expert level of effort for NLP-based idea filtering with highly imbalanced data. We ultimately find that NLP can accelerate design processes in various domains by automating idea filtering and further decreasing the level of effort required of human experts.

Thesis Supervisor: Faez Ahmed
Title: Assistant Professor of Mechanical Engineering

Acknowledgments

I would like to sincerely thank my advisor, Faez Ahmed, for his guidance, expertise, and support throughout my research and graduate studies. I would also like to thank Ryan Sochol, Dan Frey, Jaron Porciello, and Scarlett Miller for their continued guidance and insight. I would like to acknowledge and thank my co-authors Binyang Song, Vaishnavi Addala, and Aoran Pen. I am particularly grateful to Binyang Song for her mentorship and partnership throughout my research. I must close with a sincere thank you to my loving parents, family, and friends who have smiled with me through the good times, laughed with me through the wild times, and bolstered me when I most needed it. Thank you.

My research has been supported by the National Science Foundation Graduate Fellowship, MIT's Ida M. Green Fellowship, the GEM Fellowship, and the USAID Bureau for Resilience and Food Security, U.S. Agency for International Development.

Contents

1 Introduction 17

1.1 List of Contributions 19

1.2 Publications Supporting this Thesis 20

2 Natural Language Processing for Filtering Design Ideas based on Creativity 21

2.1 Introduction 21

2.2 Related Works 24

2.2.1 Creativity Assessment Methods 24

2.2.2 Machine Learning and Creativity Assessment 27

2.3 Methods 28

2.3.1 SVS Data Processing 30

2.3.2 Converting SVS Values to Text Embeddings 31

2.3.3 Universal Sentence Encoder 33

2.3.4 Dimensionality Reduction 34

2.3.5 Design Representations 34

2.3.6 Regression Models and Evaluation Metrics 35

2.4 Results 36

2.4.1 Predicting Design Metrics From One-Hot-Encoded SVS Features 36

2.4.2 Converting SVS Values to Text Embeddings 37

2.4.3 Predicting Design Metrics From One-Hot-Encoded SVS features + Text Embedded Description 40

2.4.4 Predicting Design Metrics From Text Embedded SVS Features and Description . . . 41
2.5 Discussion . . . 43
2.6 Summary . . . 45

3 Accelerating Evidence Synthesis for the Design of International Development Projects Using Natural Language Processing-based Idea Filtering 47
3.1 Introduction . . . 47
3.2 Related Works . . . 49
3.2.1 Natural Language Processing . . . 49
3.2.2 Information Extraction . . . 51
3.2.3 Evidence Gap Maps . . . 52
3.3 Methods . . . 55
3.3.1 Dataset Description . . . 55
3.3.2 Data Pre-processing . . . 56
3.3.3 Universal Sentence Encoder Model . . . 56
3.3.4 Fine-Tuned BERT Model . . . 58
3.3.5 Integrated BERT Model . . . 59
3.3.6 Training Size Variation . . . 60
3.4 Results . . . 60
3.5 Discussion . . . 63
3.6 Summary . . . 65

4 Active Learning for Real-World Imbalanced Natural Language Processing Applications 67
4.1 Introduction . . . 67
4.2 Related Works . . . 69
4.3 Methods . . . 70
4.3.1 Dataset . . . 71
4.3.2 Implementation Strategy . . . 72

4.4 Results 74
4.5 Discussion 77
4.6 Summary 79

5 Conclusion **81**

A Figures **85**

List of Figures

2-1 The overall architecture of our model. From a design, we initially have a dataset of SVS features in numerical form as well as a written text description provided by the designer. We convert the numerical SVS features into text and combine that text with the original description to gain an all-text representation of the design. We encode this text representation in Google's Universal Sentence Encoder to gain a numerical text embedding for each design. We input this text embedding into a regression model to predict five expert acquired design metrics: Usefulness, Elegance, Drawing, Uniqueness, and Creativity. 29

2-2 The process of generating text embedding using Google's Universal Sentence Encoder. Shown on the left, the text descriptions for each design are inputted into the autoencoder. The 1 x 512 numerical embeddings for each description are found, and a heat map of the semantic textual similarity is shown on the right, where a score of 1 indicates maximum similarity. 32

2-3 This graph shows the R^2 score for each of the five CAT metrics with three different design representations. The line marked at 0.3 indicates an R^2 score threshold of a weak positive relationship. These results are generated with the RF regression model. We observe that, overall, adding text embeddings improves the regression results, as seen by the increase in R^2 scores. The figure highlights that despite the improvements, the regression results for all the design metrics are still below the threshold of weak relationship. 38

2-4 A scatterplot of the predicted Usefulness rating vs. the actual Usefulness rating found using a Random Forest Regression Model. A perfect prediction would follow a line with a slope of one and intercept at the origin, represented in the plot by a black line. This prediction has an R^2 score of 0.270. 39

2-5 The left shows three student-generated designs for milk frothers: a Bicycle design, Cattle design, and Rodeo design. Under each design is a handwritten description provided by the designer. We measure the similarity of these designs via the cosine similarity of their representations. We use two different representation methods as shown in the bar chart. Note that when using one-hot-encoded features, the similarity between Bicycle & Rodeo is less than that between Bicycle & Cattle. However, text embeddings highlight their semantic similarity where riding is involved in both the devices to froth the milk. 40

3-1 A high level view of the current EGM creation process. 54

3-2 The architecture of our three transformer-based models. 57

3-3 The average classification accuracy of each model when trained on 70% of the data. We observe that the iBERT, which combines the fine-tuned BERT model with extracted information performs the best. 61

3-4 Model classification accuracy for different training sizes. Both models are influenced by training size, but we note that BERT outperforms our USE model at all training sizes. 62

4-1 The F1 score vs. training data size for the four data selection techniques. 75

4-2 The class breakdown of training data from the initial training set of 1,000 instances to the final set of 5,000 instances. 76

4-3 Class 1 Recall vs. Percent of Training Data that is Class 1. We observe a positive linear relationship between the percent of the training data that is from the positive minority class, class 1, and the recall of the positive class. The R^2 value is 0.412. 77

A-1 Training data breakdown per iteration for the Balanced Initial + Least Confidence Selection Technique. 85

A-2 Training data breakdown per iteration for the Random Initial + Least Confidence Selection Technique. 86

A-3 Training data breakdown per iteration for the Balanced Initial + Random Selection Technique. 86

A-4 Training data breakdown per iteration for the Random Initial + Random Selection Technique. 87

A-5 Class 1 recall vs. Total Training Size. We observe a very weak relationship between class 1 recall and training size, as indicated by an R^2 value of 0.02. This helps confirm that the positive linear relationship we observe between class 1 recall and the proportion of the training data from class 1 is in fact that relationship, and is not confounded by training data size. 88

List of Tables

2.1 R^2 Score by Regression Model using three different design representations. 37

2.2 Confusion matrix for classifying the Usefulness of a design using a random forest regression and percentile classification. The precision values for the low, medium, and high classes are 0.619, 0.387, and 0.645 respectively. The recall values for the low, medium, and high classes were 0.591, 0.462, and 0.580 respectively. 37

4.1 Common imbalanced datasets in NLP and their size, positive class, and the prior distribution of the positive class. Our dataset has a notably low prior distribution, indicating that it is very imbalanced. 72

Chapter 1

Introduction

During the design of products or programs, a bottleneck often arises when experts need to filter through an exorbitant number of ideas, searching for which ones are most innovative, creative, relevant, or any other number of subjective characteristics. We have seen this bottleneck arise in Google's Project 10ˆ100th. This project was launched as a "call for ideas that could help as many people as possible" [25]. However, upon receiving over 150,000 ideas, Google was forced to postpone the selection of winners until two years later due to the time and resources required to evaluate all submissions [127]. We see this bottleneck again arise when designing evidence-based policies and programs. To design such policies and programs, decision-makers must distill key information from a vast and rapidly growing literature base. Identifying relevant literature from raw search results is time and resource intensive, and often requires significant manual screening. Given that one of the main barriers to evidence use among policymakers is the lack of timely research outputs, there is a critical need to reduce the time and effort needed to complete such literature reviews [89].

Motivated by the common challenge of idea filtering in various design domains, my research explores the use of natural language processing (NLP) for accelerating design through automated idea filtering.

In Chapter 2, my team and I investigate the possibility of using machine learning to predict expert-derived creativity assessments of design ideas from more accessible non-expert survey results. When evaluating early-stage design concepts, the aim is to capture a range of information, including usefulness, uniqueness, and novelty of a design. The subjective nature of these concepts makes their evaluation difficult. Still, many attempts have been made and metrics developed to do so, because design evaluation is integral to the creation of novel solutions. The most common metrics used are the consensual assessment technique (CAT) and the Shah, Vargas-Hernandez, and Smith (SVS) method. While CAT is accurate and often regarded as the "gold standard," it relies on using expert ratings, making CAT expensive and time-consuming. Comparatively, SVS is less resource-demanding, but often criticized as lacking sensitivity and accuracy. We utilize the complementary strengths of both methods through machine learning. This study investigates the possibility of using machine learning to predict expert creativity assessments from more accessible non-expert survey results. The SVS method results in a text-rich dataset about a design. We utilize these textual design representations and the deep semantic relationships that words and sentences encode to predict more desirable design metrics, including CAT metrics. We demonstrate the ability of machine learning models to predict design metrics from the design itself and SVS survey information. We show that incorporating natural language processing improves prediction results across design metrics, and that clear distinctions in the predictability of certain metrics exist.

In Chapter 3, we go on to explore the effectiveness of using NLP to accelerate literature screening for designing evidence-based policies and programs. We compare three transformer-based text classification models: the universal sentence encoder (USE), bidirectional encoder representations from transformers (BERT), and BERT integrated with a rule-based model. We further study the influence of training data size on accuracy. Results show that the fine-tunable BERT-based models outperform the USE model by 10% in terms of accuracy; the integrated BERT model surpasses the BERT model by 2%, achieving an accuracy of 79%. The model accuracy improves with the training data size with diminishing marginal effect. The findings can facilitate informed decision making regarding the trade-off between model accuracy and manual labeling efforts, increasing efficiency.

After demonstrating the effectiveness of using NLP to accelerate literature screening, we aimed to next decrease the level of effort required of expert reviewers to generate training data. In Chapter 4, we were motivated to explore avenues to decrease the size of training data needed by using active learning (AL). AL is the concept that a machine learning algorithm can perform better with less training data if it is allowed to choose the data from which it learns. We experiment on classifier models with four different data selection scenarios with our highly imbalanced real-world dataset. We find that data selection techniques that incorporate active learning result in higher F1 scores, a more balanced training set, and fewer necessary labeled training instances. These results suggest that active learning is effective in decreasing expert level of effort for NLP-based idea filtering with highly imbalanced data.

1.1 List of Contributions

The contributions from our three studies are listed below:

- Chapter 2
  - Incorporating natural language processing in the representation of a design consistently improves a model's ability to predict expert design metrics.
  - Different design metrics vary distinctly in their predictability. Usefulness and Elegance perform the best, and Creativity and Uniqueness perform the worst.
- Chapter 3
  - For classifying documents for inclusion in evidence syntheses, the fine-tunable BERT-based models outperform the USE model by 10% in terms of accuracy; the integrated BERT model surpasses the BERT model by 2%, achieving an accuracy of 79%.
  - Model accuracy improves with the training data size until a size of 10,000 with diminishing marginal effect.
- Chapter 4

- Incorporating active learning in data selection results in an F1 score of 0.78, which is 7.7% higher than the baseline data selection technique.
- Incorporating active learning in data selection results in more balanced, and consequently more diverse, training datasets.The two techniques incorporating AL end up with training sets comprised of 30% and 33% of minority class instances, while the two techniques without AL result in 7% and 16% minority class training sets.
- If an active learning technique is used to identify new samples to query, then a random initial training set will perform *just as well* as a balanced initial training set. Since obtaining a balanced initial training set requires non-trivial expert labeling effort, an equally high performance from the random initial training set means we can propose a technique that maximizes model performance while minimizing human expert level of effort.

We ultimately find that NLP can accelerate design processes in various domains by automating idea filtering and decreasing the level of effort required of human experts.

1.2 Publications Supporting this Thesis

- If a Picture is Worth 1000 Words, Is a Word Worth 1000 Features for Design Metric Estimation? *Journal of Mechanical Design* [45]
- If a Picture is Worth 1000 Words, Is a Word Worth 1000 Features for Design Metric Estimation? *International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, 2021* [46]
- Accelerated Evidence Synthesis in International Development using Integrated Transformer-based Language Models (Submitted)

Chapter 2

Natural Language Processing for Filtering Design Ideas based on Creativity

2.1 Introduction

A picture is often said to be worth a thousand words because of the amount of information it can transmit. A picture captures not only the object or concept of interest, but also potentially embedded interactions with the environment and, possibly, the preferences of the picture owner [118]. The same can be said with words in design concept evaluation due to the nature of languages being complex and context-dependent. For example, the same word may have different meanings and different words may have the same meanings when used in different contexts. As a result, the number of features used to capture the dynamic state of a word increases. This is also what makes creative design evaluations difficult. The need for creative design evaluations stems from the increased attention in research on creativity and innovation in engineering, as they are crucial in providing novel solutions to new and existing problems [12, 80, 7]. It has even been said that creativity is mankind's most valuable resource, as innovation and progress rely heavily on creativity [24]. Creativity and innovation mark an individual's ability to produce new ideas, a skill that is crucial in the production of novel technology [14, 122, 87, 72]. As a result, there has been a surge in research that examines possible methods to boost student creative and innovative behaviors [78, 125, 106].

Methods to engage students in creativity require separate methods that assess the outcomes. Assessments can help identify creative ideas and individuals, and facilitate improvements in both [105, 124, 35, 48]. Assessments can also serve as a way of evaluating the design metrics in terms of their effectiveness at aiding the process of creative idea generation [112]. A lack in assessment strategy is not only a setback to assessing the suitability and effectiveness of these creativity-enhancing techniques in specific projects, but also may bring into question their overall effectiveness [72]. Therefore, there is a need to measure not only if a concept is creative, but also to what degree it is creative [105].

There are a plethora of metrics that aim to measure creativity today [39, 24, 112]. These metrics include, but are not limited to, expert panels [9, 33, 32, 20, 51], the Consensual Assessment Technique (CAT) [10, 14, 21], the Shah, Vargas-Hernandez, and Smith (SVS) method [112], and the Comparative Creativity Assessment (CCA), which is built upon the SVS method [73]. Among all of the metrics created, the most common are the CAT [10, 14, 21] and the SVS method [112]. Despite the fact that many metrics exist, measuring creativity is still difficult. One of the reasons could be due to the multi-faceted nature of creativity and what it entails [101]. In addition, the unique characteristics of the measurement methods can also result in increased challenges for researchers to establish an assessment standard [72, 85, 17]. The abundance of metrics available has resulted in great variability between the methodology of different studies, which makes comparing findings increasingly difficult [85]. For example, SVS results have been found to not match expert ratings in design variability [74], while the heavy dependence of CAT assessments on the experience, number, and subjectivity of the experts can result in the experiment being significantly restricted by their time and financial budget [39].

To address these limitations with the creativity assessment metrics, this study tries to uncover how machine learning methods can enable automated assessment of creativity. Specifically, we investigate how regression models can be used to predict the design metrics for unseen designs using SVS features. A total of five design metrics relating to creativity will be measured, including creativity itself, since in addition to measuring the different aspects of creativity, it is also important to gauge the overall creativity of the design. Creativity can be defined here as a measure of the capacity to generate original work that is useful [14, 122, 87]. In the rating process for CAT, the ratings of the experts rely heavily on their own internal heuristics and prior experience. At the beginning of the rating process, ideas will be picked out from the group that represent high, medium, and low creativity [102]. The rating of each expert will be different, as this is heavily based on the individual expert's own definition of what creativity is, and how they perceive the ideas. The expert raters will then be asked to use the low, medium, and high creativity concepts as the baseline, and use them to evaluate all the other concepts [54]. This comparative evaluation will be captured through a 7-point Likert scale [54], and can be an accurate assessment of how relatively creative each concept is. Then, the expert will examine the concepts further in terms of its usefulness (quality and utility of the idea), uniqueness (originality of the idea), and elegance (well-crafted), which are CAT sub-metrics [26]. It will also be used to evaluate the drawing of the concepts as it has been found to be correlated with design outcome [133].

In this study, we will further show how natural language processing (NLP) based models, which capture semantic relationships between words, can help overcome the issue of attribute dependencies in SVS features and improve the prediction results. NLP is a subset of speech and language processing that aims to train the computer to interpret the text in a more naturalistic, human way [28]. Applications of NLP include text sentiment detection and response generation [28], all of which could be helpful in completing the goal of this study. More specifically, this study can help to examine the plausibility of using SVS ratings to predict and produce design metrics including CAT ratings; CAT ratings are very resource demanding, whereas utilizing SVS ratings would be faster, easier, and cheaper to gather. Our code and additional information about our work are available at http://decode.mit.edu/projects/nlp-design-eval/

2.2 Related Works

Creativity and innovation are traits that are greatly valued in the current market, as they are crucial in the formation of new ideas [12, 80, 7]. As a result, there has been an abundance of research on methods that aim to encourage and improve creativity in students [78, 125, 106]. An important step following the promotion of creativity is the assessment of creativity, which is important in the identification and evaluation of progress [105, 124, 35, 48, 112]. The two most commonly used metrics for creativity measurement are the CAT [10, 14, 21] and the SVS method [112]. Although the CAT method is valued as more accurate in the measurement process, it is also very resource consuming [61, 39, 64] due to the time needed to have expert raters code hundreds or thousands of ideas, plus the time needed to train novice raters if experts are unavailable. By comparison, the SVS method is faster and cheaper, as novice raters can achieve high levels of inter-rater agreement. However, it lacks sensitivity and accuracy, and does not match the rating of expert reviewers [74]. Therefore, this study was constructed to investigate the plausibility of using SVS data to predict and produce design metrics that accurately assess the creativity of the concepts.

2.2.1 Creativity Assessment Methods

Creative assessment methods are a crucial after-step of engineering design research, and are often seen as a ”means to an end” to researchers studying engineering design creativity. They are an effective tool in identifying and evaluating the effectiveness of creativity-enhancing techniques, and helping assess the novelty of ideas generated [105, 124, 35, 48, 112]. There have been many metrics designed that aim to measure the creativity of concepts [62] or aspects related to creativity. Research has classified these assessment methods into two categories: process-based and outcome-based [88]. More specifically, process-based methods focus on the cognitive processes that are involved during the concept generation process [88]. Comparatively, outcome-based methods examine the outcomes of the ideation process [88, 112]. Of these two categories, the outcome-based methods are more commonly known and used [88]. This is primarily due to the complexity and difficulty associated with process-based approaches [88, 112]. Although research has been done to classify the different methods, there has not been one method that is denoted as the standardized method [62]. This is because each of the metrics available have their respective advantages and disadvantages. For example, although CAT has been praised as the "gold standard of creativity assessment," [23] it is still flawed because its methods are very time and resource consuming [61, 39, 64]. Therefore, it is up to the researchers to determine, based on their needs and their resources, which method their study will employ. For this study, the focus will primarily be on CAT and SVS, two of the most commonly used metrics.

The CAT was first conceptualized and developed by Teresa Amabile to assess creativity in a subjective manner [10, 12, 13, 85], and is often regarded as the best rating method [61, 39, 23, 68]. It measures creativity by employing a panel with appropriate expertise in the field of interest and asking them to provide their own ratings on the products or ideas generated based on a Likert scale [68, 23, 22, 10, 14, 21, 85, 39]. The process for attaining CAT ratings is as follows: (1) a group of creative concepts are gathered [68, 23, 39, 85] and (2) raters are then asked to provide ratings from 1 (low in a factor) to 7 (high in a factor) based on the definition of each factor[85]. During the assessment, it is stressed that the experts should make their assessments independently, subjectively, and take into consideration other products under review [68, 23, 39, 85].

The basis for this method is that an idea is only creative to the extent that experts agree, independently, that it is creative. [85, 61]. Because of this, the accuracy of these assessments heavily depends on the expertise of the reviewer in that field [11, 10]. This is supported by previous research that found expert ratings to have a higher agreement (higher inter-rater reliability) than non-expert ratings [11, 10, 39, 65]. Therefore, it could be said that there is no standard scoring system available for

CAT, as it is entirely based on subjective comparison within a certain group [64]. This is because concepts designed in the same environment, same situation, and same predisposition can be evaluated with respect to each other [54].

Although CAT ratings are accurate because they are based on the opinion of experts in the field, they are also relatively difficult to gather [39]. Experts are often very expensive, relatively hard to find, and extremely busy [39]. This aspect of CAT makes it more difficult to implement than some of the other metrics. Human raters have also been shown to be inconsistent between each other, which can be a result of different expertise levels, as well as differences in their beliefs about creativity [77]. These are among the reasons why current researches are looking into alternatives, such as using novice raters [65], quasi-novice raters [66, 67], or in the case of this chapter, machine learning.

Another common metric that is used is the SVS method [112]. This method is an example of a model using the genealogical tree approach [7] and is more commonly used and accepted in engineering [112, 91]. Here, the focus is more on using effectiveness to quantify creativity of ideation [85]. When using the SVS approach, the role of the human rater is replaced by predefined components as an attempt to increase repeatability and reduce subjectivity [6]. This type of metric usually breaks down the concepts into components, and quantitatively measures the creativity of each component based on relative frequencies [112, 85, 111]. More specifically, the concepts will be broken down based on the function of the components [85, 17]. However, this metric has been criticized as lacking in its sensitivity and accuracy [6]. For example, Linsey's research reported that SVS results were inconsistent with the ratings produced by experts in terms of variety of concepts [74]. Other studies have reported that SVS results can have decreased accuracy as a result of an increase in sample size [115, 94]. In addition, one study by Sluis-Tiescheffer *et al.* [115] found the SVS approach was unable to provide comparison between different attributes, only allowing for comparison between the same attribute.

SVS encompasses four sub-metrics: (1) novelty, (2) variety, (3) quality, and (4) quantity of ideation [112]. In this case, novelty can be defined as how different the concept is from other concepts; variety is how different the concept is from other concepts generated by the same designer; quality is a subjective measure of feasibility and degree of success at meeting desired requirement; and quantity is the number of concepts generated [88, 85, 112]. Of these four sub-metrics, quality and novelty are usually the more focused factors, as novel and appropriateness of ideas being part of the definition of creativity [85, 14]. The novelty component of SVS examines how similar the idea is with other ideas from the same group [112, 88, 85]. Through the genealogical tree, also known as the feature tree approach, SVS proposes that novelty can be calculated based on the type of features the concept includes, as well as how each feature is satisfied [112]. Therefore, concepts that have features in categories with lower overall frequency will indicate that not many other concepts share their idea, which would then indicate higher novelty for that idea [85]. Comparatively, the quality component of SVS measures the feasibility of the concepts in terms of how successful they are at meeting the desired design requirements [85, 88, 112].

2.2.2 Machine Learning and Creativity Assessment

Data-driven research approaches and methods, like machine learning, are advantageous for analyzing large amounts of data for meaning, patterns, relationships, and even development and formation of theories [86]. In recent years, machine learning algorithms can even be used to construct models from data that are not necessarily linearly related [86]. This characteristic is what gives machine learning the potential to be used in the analysis of subjective measures. This is supported by prior research, where machine learning has been successfully used to predict subjective measures, such as mental workload [86]. It has also been used in objective quality-assessments of videos that would otherwise demand a much larger-scale and a more expensive experiment [8]. The result of this study shows that machine learning was able to significantly reduce the amount of work needed without compromising the conclusion [8]. Machine learning also shows promise in contributing to opinion-based data-mining [123]. It has already been found that opinion mining can be used to determine whether a sentence or a document is expressing positive or negative sentiment [123]. Machine learning and opinion mining has also been used with natural language processing to assess online reviews [123, 60], where opinion spam detection would be used to separate out reviews while usefulness measurements [123, 60] can be used to identify the usefulness and subjectivity of a review [123, 52, 75, 79, 53]. Therefore, the ability of machine learning to assist in the assessment of subjective measures, like creativity, is investigated in this study. More specifically, this study focuses on using the natural language processing.

NLP is a type of machine learning process that attempts to teach computers to interpret language in a natural, "human," way [28]. More specifically, it seeks to explore how computers can be used to understand and manipulate natural language to accomplish useful tasks [34]. Therefore, it could be said that the ultimate goal of NLP is to achieve "human-like language processing" abilities [71]. NLP has been found to be helpful in the completion of tasks like understanding the sentiment behind text and generating responses to questions [28]. For example, NLP has been used to extract information from narrative text in the medical field, and has found the results to be reasonable [82]. In another study by Li et al., NLP was used to assess Chinese subjective answers [70]. Design researchers have used NLP based methods on a plethora of tasks ranging from identifying manager interventions [55], predicting contest winners [5], sentiment analysis of conversations [41] to understanding product reviews [63]. In this chapter, we show how these models, when combined with design feature information, can also enable prediction of multiple design metrics.

2.3 Methods

Our goal is to predict design metrics including expert CAT ratings of a design using SVS features and a description gathered from that design. Details on the rating process can be found in the work by Miller et al. [85]. More specifically, for the CAT ratings, two expert raters were used. One rater has a graduate degree and the other has completed graduate coursework. Both raters were in engineering-related fields. Both raters had at least four years of experience in design and assessment, and

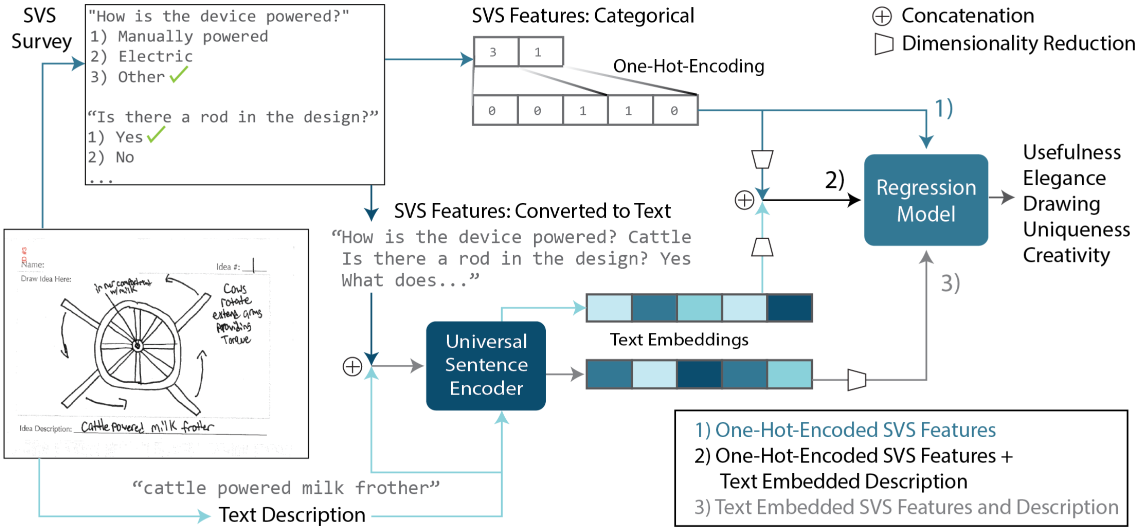
SVS Survey
"How is the device powered?"
1) Manually powered
2) Electric
3) Other ✓
"Is there a rod in the design?"
1) Yes ✓
2) No
...
SVS Features: Categorical
3
1
One-Hot-Encoding
0
0
1
1
0
⊕ Concatenation
Dimensionality Reduction
1)
2)
3)
Regression Model
Usefulness
Elegance
Drawing
Uniqueness
Creativity
SVS Features: Converted to Text
"How is the device powered? Cattle
Is there a rod in the design? Yes
What does..."
Universal Sentence Encoder
Text Embeddings
ID #3
Name:
Idea #: 1
Draw Idea Here:
inner compartment w/milk
Cows rotate extend arms providing Torque
Idea Description: Cattle powered milk frother
"cattle powered milk frother"
Text Description
1) One-Hot-Encoded SVS Features
2) One-Hot-Encoded SVS Features + Text Embedded Description
3) Text Embedded SVS Features and Description

Figure 2-1: The overall architecture of our model. From a design, we initially have a dataset of SVS features in numerical form as well as a written text description provided by the designer. We convert the numerical SVS features into text and combine that text with the original description to gain an all-text representation of the design. We encode this text representation in Google's Universal Sentence Encoder to gain a numerical text embedding for each design. We input this text embedding into a regression model to predict five expert acquired design metrics: Usefulness, Elegance, Drawing, Uniqueness, and Creativity.

had published papers in the related fields. The rating process followed the guidelines established by Besemer [26] and Besemer and O'Quinn [27]. The raters received around 20 hours of training sessions that included a history of CAT and how to use it, potential biases that might arise when rating, and the meaning of the different ratings. The raters then engaged in a practice round of rating and discussed their different ratings. Following the training session, the raters independently rated all the ideas. Their independent ratings were then aggregated. For the SVS ratings, two raters were recruited. Both raters are engineering students, with one graduate student and one undergraduate student. One of the raters had previous experience using the CAT process. Both raters received training on the design process and the rating process. The interrater agreement for this approach was 0.85.

We experiment with different methods to predict the five different design metrics: Usefulness, Elegance, Drawing, Uniqueness, and Creativity. Our different methods stem from using three different representations of a design, as well as three different regression models. The different design representations are derived from data available in the design itself and the unprocessed SVS features. The three design representations are (1) One-hot-encoded SVS Features, (2) One-hot-encoded SVS Features + Text Embedded Description, (3) Text Embedded SVS Features and Description.

Figure 2-1 shows how we created each of the three design representations and we will explain what each of the design representations is in detail in the first five subsections of the Methodology section. In addition to the three distinct design representations, we explore the use of three different regression models: (1) Linear regression, (2) Gradient Boosting (GB) regression, and (3) Random Forest (RF) regression.

As shown in Figure 2-1, we originally start with a design. Initially, we have both a text description and a completed SVS Survey about the design. In the following sections, we will discuss our methods of processing the data, our motivation for converting data into different design representations, which design representations provided the most predictive power, and how this varied for different expert-gathered design metrics.

2.3.1 SVS Data Processing

The original SVS data comes from a survey where raters are asked to examine different designs of milk frothers and answer questions such as: " How is the device powered?" [121] There are 91 questions total. This survey was developed based on the feature-tree approach, where the designs are first reviewed and categories of features that are common in the current designs are identified. These categories were then used to design the SVS survey used in this study, where each question and its respective response serves as an initial feature for the design. An example of two of these survey questions is shown in Figure 2-1, labelled by SVS survey. As the figure shows, the survey responses are categorical. Therefore, the responses have been pre-processed such that each distinct response to a question is given a number. For example, for the question "How is the device powered?" the provided response options are (1) Manually powered, (2) Electric, or (3) Other, and the corresponding number represents the response.

For example, Figure 2-1 shows that the responses for the two questions can be mapped to a two-dimensional vector, where the first value of “3” corresponds to the third category “Other” and the second value “1” corresponds to the first category “Yes” for the second question. Although the responses were given numerical values, their relationships were not numerical, i.e. the difference between “1” and “3” is not necessarily greater than the difference between “1” and “2.” Consequently, we converted these categorical features through one-hot-encoding into vectors.

One-hot-encoding is the most widely used method for converting categorical data into numerical data that can be used for a machine learning model [98]. This method is commonly used when there is not ordinal relationship between categories, since giving each category a varying numerical value would add an ordinal relationship that does not exist [98, 56]. Categorical data often contains labels rather than numbers, for example, an image might be labeled “cat” or “dog”. One-hot-encoding converts these labels into binary features. If the variable were “Type of pet” and the options were “cat” and “dog,” then “cat” can be given a value [0,1] and “dog” can be given a value [1,0].

In the example shown in Figure 2-1, the first question had three possible answers, designated “1,” “2,” and “3.” This single question became three binary questions, where response “1” is now [1,0,0], response “2” is now [0,1,0], and response “3” is now [0,0,1]. Each value is considered a feature, so this overall process increased the number of features from the initial 91 to 522. These 522 dimensional one-hot-encoded SVS features serve as our first design representation.

2.3.2 Converting SVS Values to Text Embeddings

We were motivated to convert our SVS values into text embeddings in order to utilize the relationships between certain features. In our first design representation, all of the features are one-hot-encoded.One-hot-encoding creates features that are all assumed to be completely independent of one another, so a design that is labeled “bolt” is as far away from a design labeled “nail,” as it is from a design labeled “hammer.” However, humans recognize that the first two words are semantically closer to each other than

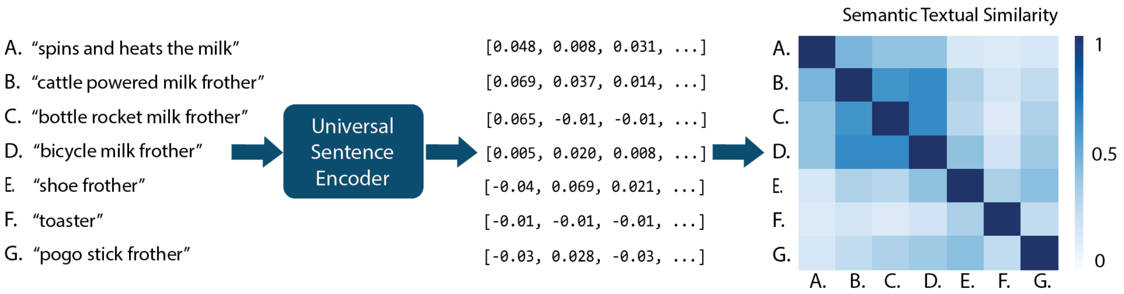

Figure 2-2: The process of generating text embedding using Google's Universal Sentence Encoder. Shown on the left, the text descriptions for each design are inputted into the autoencoder. The 1 x 512 numerical embeddings for each description are found, and a heat map of the semantic textual similarity is shown on the right, where a score of 1 indicates maximum similarity.

the first and third word. In this sense, one-hot-encoding all of our features forced us to lose many relationships between designs.

To address this, we converted all information about a design into text and we combined all of the text such that each design was defined by a long text string. We demonstrate this process in Figure 2-1. Our method used the original SVS survey responses in numerical form. For each question that had a response other than "No," we included both the question and the response as text in the representation of a design. For example, for the second question in Figure 2-1, "Is there a rod in the design?" the response is "Yes," therefore, the question and the response are both included in the string of text that represents the design.

If a question's response was "No," we chose to exclude the question and the response from the text string representation. This was an algorithmic choice. We decided to only represent the features which a design has, and not the features which it did not have, in the representation scheme.

For the first question in Figure 2-1, "How is the device powered?," the response is "Other." Whenever a respondent selected "Other," the respondents provided a description themselves. In this case, the description was "Cattle," so both the question "How is the device powered?" and the response "Cattle" are included in the string of text that represents the design.

Lastly, we add the original text description to the text string representation. The handwritten text descriptions were manually read and entered as strings into the SVS feature dataset during pre-processing [121]. The process of adding the text description to the overall text string representation was automated. With this new design representation, all of the features are in text format. The next step is to convert this textual design representation into numerical continuous embeddings that can be used in regression models.

2.3.3 Universal Sentence Encoder

Figure 2-1 shows that we run both the text description and the SVS features as text through a sentence encoder in order to get text embeddings. A text embedding is a vector representation of text in which text with similar meanings are represented with similar vectors.

We use Google's Universal Sentence Encoder, which maps text into 512-dimensional space [31]. Figure 2-2 demonstrates how we use the Universal Sentence Encoder. We input a string of text such as "spins and heats the milk," then the Universal Sentence Encoder maps the text into 512-dimensional space and outputs a text embedding: a 512-dimensional vector of numbers that represents the inputted text.

The Semantic Textual Similarity plot shown in Figure 2-2 illustrates the effectiveness of the Universal Sentence Encoder. In the plot, the original sentences are represented with letters A-G. After calculating the text embeddings, we use cosine vector similarity to find which sentences are more similar. The similarities between the seven sentences are shown by the 7x7 matrix at the right. A darker square indicates a higher similarity between two sentences. For example, sentences B-D all include the word "milk frother," and these sentences are shown to have some of the most similarity (darkest squares). In contrast, sentences E-G are quite different from each other and consequently have low similarity scores and light squares. We also notice deeper relationships being captured by the Universal Sentence Encoder. For instance, the most similar item in this list to "pogo stick frother" is a "shoe frother," which relates to the deeper connection between a "shoe" and a "pogo stick."

2.3.4 Dimensionality Reduction

In the above sections, we discussed using both one-hot-encoding and text embeddings to represent our designs. Both of these methods output vectors of over 500 dimensions, or features. The higher the dimensionality, the greater number of datapoints, designs in this case, are needed for a model to learn effectively [120]. Because we have 934 designs, we need to reduce the number of dimensions in our feature space.

We utilize principal component analysis (PCA) to reduce the dimensions of our feature space. PCA is a tool from linear algebra that projects our original features into a lower-dimensional space. PCA accomplishes this by taking linear combinations of the original features, thereby creating new features. These new features are principal components, and they are independent of one another and also retain most of the information from the original features. PCA aims to put the most information possible in the first component, and the second most in the second component, and so on. Finally, PCA drops the principal components that have the least information, thereby reducing the number of features we have.

2.3.5 Design Representations

The three different design representations come from applying various combinations of the data processing methods mentioned above. Ultimately, the first design representation is a one-hot-encoding of all of the SVS features - resulting in 522 features total. We illustrate the process of generating this representation in Figure 2-1, where it is denoted as "1."

The second design representation combines one-hot-encoding, text embedding, and PCA, as illustrated in Figure 2-1. This design representation results in 30 features total, 15 from the one-hot-encoded SVS features after PCA, and 15 from the embeddings of the text descriptions after PCA.

The third design representation converts all of the information available about a design into text - the SVS data as well as the text descriptions. From that complete text representation, the third design representation finds a text embedding. This text embedding is reduced via PCA from 512 dimensions to a final form of 30 dimensions.

2.3.6 Regression Models and Evaluation Metrics

We experimented with generating design metric predictions with three different regression models: linear regression, gradient boosting (GB) regression, and random forest (RF) regression.

Linear regression attempts to model the relationship between independent variables and a dependent variable through a linear equation. We train the linear regression by minimizing the sum of squared differences between our predicted and actual values. This method is known as the least-squares method.

The gradient boosting and random forest regression models are both ensembles of decision trees. An ensemble means that these models are aggregations of other models. The motivation behind an ensemble method is to combine predictions from multiple base models into a prediction that is better than that of any single model [108]. GB trains multiple decision trees sequentially, which is an ensemble method called boosting. The first decision tree is able to predict the majority of the data, and the following trees work to capture areas of the data that have been missed.

RF utilizes bagging, or training individual models in parallel on a randomized subset of the data, as the ensemble method [108]. RF takes the average of the predictions. This combination of decision trees yields better predictions because it has lower variance compared to a single decision tree. We train both GB and RF using the least-squares method.

We perform supervised learning with an 80-20 train-test split for each of our models. The train-test split procedure is typically done to avoid the phenomena of overfitting in a model [132]. The train-test split values of 80% and 20% describe what percentage of our data we use for training our model vs. what percentage we keep out to test how well our model works on data it has never seen before. We have 934 designs total, so we train using 747 designs, and test using the remaining 187.

We use the R^2 value as the evaluation metric of our regression models. For each model we predict design metrics for the designs in the test set. We compare these predictions to the actual design metrics and use this comparison as a means for evaluating how effective our model is. The R^2 value, or coefficient of determination, quantifies the degree to which our predicted values are linearly correlated with our actual values. A perfect R^2 score is 1, indicating perfect linear correlation, and an R^2 score of 0.3 is generally accepted as indicating a weak positive correlation.

2.4 Results

In this section, we discuss how effectively we predicted CAT expert ratings, with three varying parameters: (1) which design metric we are predicting, (2) which design representation we are inputting, and (3) which regression model we are using. We found that all three of these parameters impact the effectiveness of our predictions, as evaluated by the R^2 metric. The following sections are divided based on design representation, and, within each section, we explore results for predicting each design metric with each regression model.

2.4.1 Predicting Design Metrics From One-Hot-Encoded SVS Features

The left three columns of Table 2.1 shows the R^2 score obtained using the first design representation: one-hot-encoded SVS features. Overall, linear regression performs the worst, with consistently large and negative R^2 scores. Gradient boosting and random forest regressions perform comparably, with gradient boosting slightly outperforming random forest for every CAT metric. The trends along the design metrics are also distinct. Due to the poor nature of the linear regression results in this particular experiment, we will only discuss trends seen within the gradient boosting and random forest regressions. For the GB and RF models, the Usefulness design metric shown in the final row of Table 2.1 has the highest R^2 value, with Elegance following behind. Creativity and Uniqueness show some of the worst R^2 scores, Creativity being the worst with negative R^2 values across both the GB and RF models.

	One-hot encoded SVS features			One-hot encoded SVS features + Description text embeddings			Text-based SVS features + Description text embeddings		
	Linear Regression	**Gradient Boosting**	**Random Forest**	**Linear Regression**	**Gradient Boosting**	**Random Forest**	**Linear Regression**	**Gradient Boosting**	**Random Forest**
Creativity	-9.54E+25	-0.033	-0.263	-0.023	-0.114	-0.098	-0.014	-0.085	**0.034**
Uniqueness	-7.24E+25	**0.070**	-0.020	-0.109	-0.099	-0.043	0.016	-0.159	0.057
Drawing	-5.57E+24	0.117	-0.031	**0.163**	0.114	0.113	0.150	0.064	0.099
Elegance	-2.53E+25	0.162	0.100	0.140	0.058	0.139	0.179	0.079	**0.211**
Usefulness	-4.05E+26	0.171	0.162	0.149	0.131	0.214	0.209	0.180	**0.263**

Table 2.1: R^2 Score by Regression Model using three different design representations.

		Actual Class		
Predicted Class		Low	Medium	High
	Low	39	20	7
	Medium	13	24	15
	High	11	18	40

Table 2.2: Confusion matrix for classifying the Usefulness of a design using a random forest regression and percentile classification. The precision values for the low, medium, and high classes are 0.619, 0.387, and 0.645 respectively. The recall values for the low, medium, and high classes were 0.591, 0.462, and 0.580 respectively.

2.4.2 Converting SVS Values to Text Embeddings

Figure 2-5 demonstrates the motivation and effectiveness of using NLP in this context. The figure displays three designs for milk frothers, from left to right: a Bicycle design, a Cattle design, and a Rodeo design. The Bicycle design involves attaching containers of milk to the spokes of a bicycle wheel and frothing the milk through the motion of riding a bike. The Cattle design involves putting milk inside a horizontal wheel, cattle will push the spokes of the wheel as they walk in circles, which froths the milk. The Rodeo design involves attaching a container of milk to the back of a mechanical bull, and the motion of riding that bull will froth milk. For context, many other designs in the dataset are a variation of whipping milk with a whisk, so these three designs are qualitatively unique and similar to each other.

Our goal is to create an objective model (as opposed to subjective expert ratings) that captures relationships among designs in order to more effectively predict their design metrics. The results shown in Figure 2-5 illustrate that using NLP in design representation captures these relationships much more effectively than using

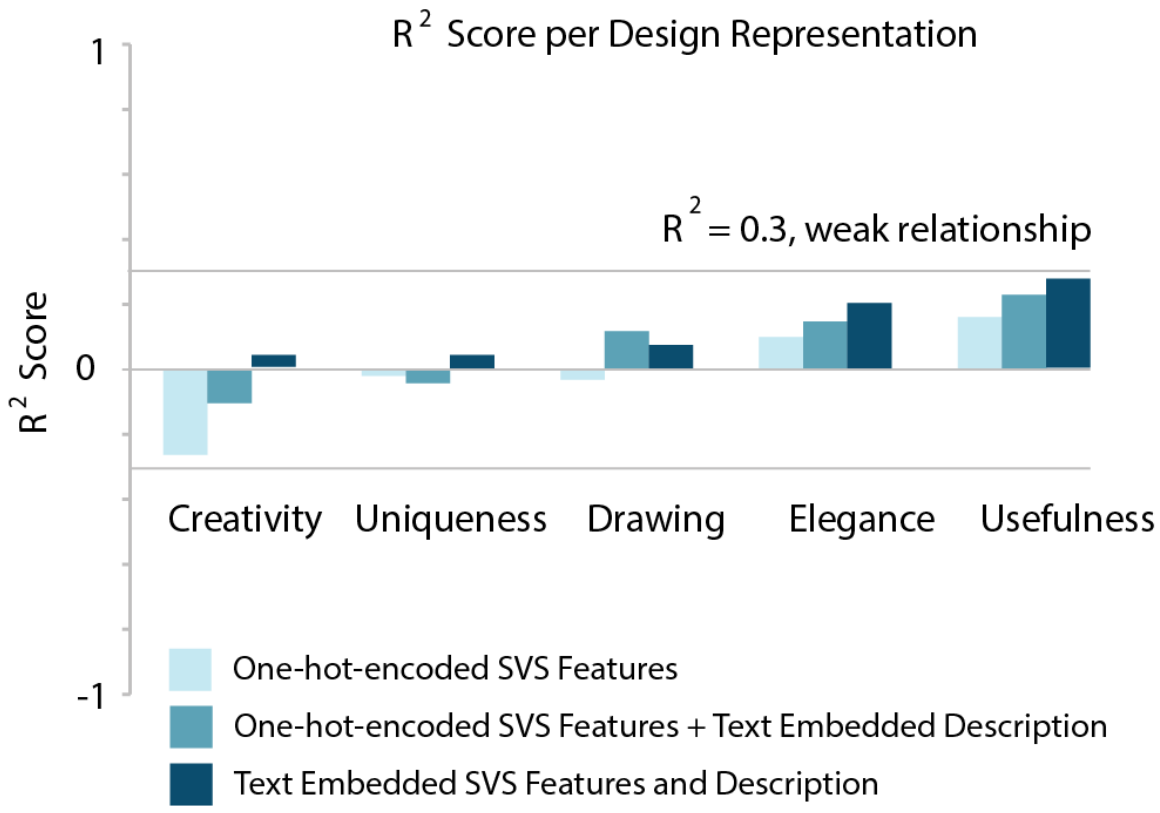

Figure 2-3: This graph shows the R^2 score for each of the five CAT metrics with three different design representations. The line marked at 0.3 indicates an R^2 score threshold of a weak positive relationship. These results are generated with the RF regression model. We observe that, overall, adding text embeddings improves the regression results, as seen by the increase in R^2 scores. The figure highlights that despite the improvements, the regression results for all the design metrics are still below the threshold of weak relationship.

discontinuous one-hot-encoding representations.

The bar chart in Figure 2-5 displays the cosine similarity between designs for two design representations: One-Hot-Encoded SVS Features, and Text Embedded SVS Features and Description. Cosine similarity is a measure of similarity between two vectors, and ranges from 0 to 1. For two vectors A and B, it is defined as:

$$Sim(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{2.1}$$

In the example of the Bicycle and Rodeo designs, the cosine similarity between their one-hot-encoded representations is 0.333, whereas the cosine similarity between

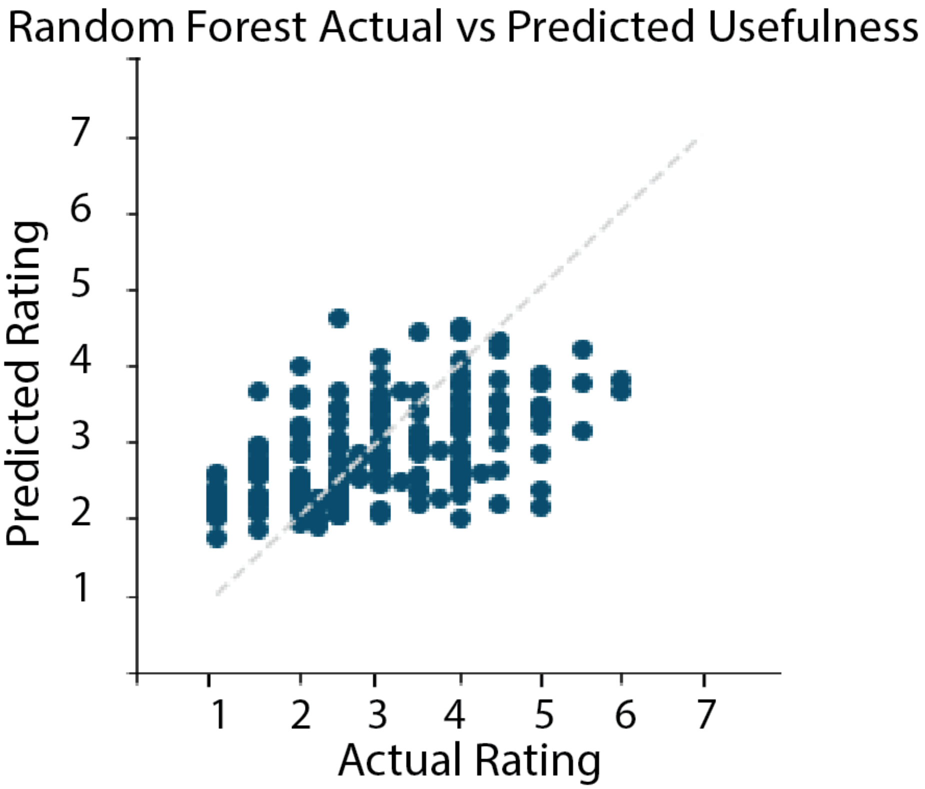

Figure 2-4: A scatterplot of the predicted Usefulness rating vs. the actual Usefulness rating found using a Random Forest Regression Model. A perfect prediction would follow a line with a slope of one and intercept at the origin, represented in the plot by a black line. This prediction has an R^2 score of 0.270.

the text embedded representations is 0.810. This may suggest that the text embedded representations maintained relationships between the designs that are lost when using one-hot-encoded representations. These results comply with our understanding of the two types of representation. One-hot-encoded vectors ensure that each feature has a cosine similarity of 0 with any other feature, since each feature that exists has a value of 0 in all dimensions except one, where it has a value of 1. Each one-hot-encoded vector of a feature is orthogonal to all others. In contrast, text embeddings found through NLP, such as those we used through Google's Universal Sentence Encoder, map all of the features in a continuous space, which preserves deeper relationships between them.

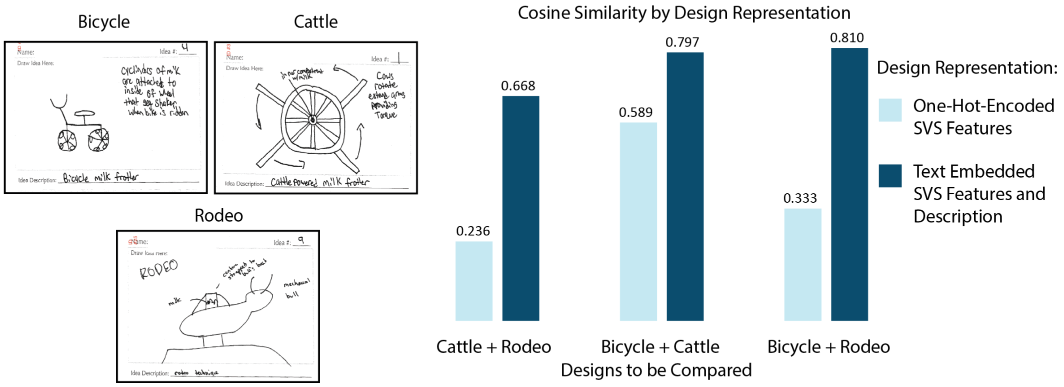

Figure 2-5: The left shows three student-generated designs for milk frothers: a Bicycle design, Cattle design, and Rodeo design. Under each design is a handwritten description provided by the designer. We measure the similarity of these designs via the cosine similarity of their representations. We use two different representation methods as shown in the bar chart. Note that when using one-hot-encoded features, the similarity between Bicycle & Rodeo is less than that between Bicycle & Cattle. However, text embeddings highlight their semantic similarity where riding is involved in both the devices to froth the milk.

2.4.3 Predicting Design Metrics From One-Hot-Encoded SVS features + Text Embedded Description

Our second design representation utilizes NLP and represents designs using one-hot-encoded SVS features and a text embedding of the design's description. As described in the Design Representations section of Methodology, we perform PCA dimensionality reduction on both of these vectors, resulting in 30 features for the second design representation.

The middle three columns of Table 2.1 shows the R^2 scores for each of the design metrics with predictions from each of the three regression models. As compared to Table 2.1, the linear regression model performs significantly better with this design representation. All three regression models perform comparably. The performance of the design metrics matches the general trend seen in the first design representation. Usefulness and Elegance are the best predicted. With the new design representation, Drawing is more accurately predicted in both linear regression and random forest, while gradient boosting's prediction is comparable to its prediction from the first design representation.

The overall trend across different design representations is illustrated in Figure 2-3 for the random forest regression. The figure indicates an overall improvement in R^2 scores from the first design representation: One-Hot-Encoded SVS Features, to the second design representation: One-Hot-Encoded SVS Features + Text Embedded Description.

2.4.4 Predicting Design Metrics From Text Embedded SVS Features and Description

Our third design representation is a text embedding of both SVS features and the text description. All of the information from the SVS Features is converted to text, combined with the original text description, mapped to a continuous space, and then encoded as text embeddings that preserve the semantic relationships from the continuous mapping.

Regression Results: The results from using this design representation are shown in the right three columns of Table 2.1, which shows the R^2 score found for each combination of design metric and regression model. The results are consistent with the overall trend seen in both design representation one and two. Usefulness and Elegance are the design metrics that are best predicted, as demonstrated by the highest R^2 scores across all three regression models. For both of these metrics, all three regression models perform the better with this NLP based design representation than with the other two design representations. Additionally, for both the linear regression and random forest models, the performance of Usefulness and Elegance have consistently improved from design representation one to design representation three. This overall trend can be observed in Figure 2-3 for the random forest model.

In agreement with the trend we have observed in the other two design representations, the Creativity and Uniqueness metrics have the lowest R^2 scores. We will discuss our hypotheses explaining why certain CAT metrics are consistently more or less accurately predicted in the Discussion section to follow.

We show a visual representation of the predicted design metrics vs. the actual design metrics in Figure 2-4. This scatter plot is generated using predictions from the random forest regression model using the third design representation: Text Embedded SVS Features and Description, and the Usefulness metric. A set of perfect predictions would follow the black line with a one-to-one slope between the Actual and Predicted Ratings.

The plot reveals that most of the Predicted Ratings range between 2 and 4. This trend appears for other CAT ratings as well. Remedying the models to more successfully predict values on the high and low extremes, perhaps through increasing the weight of designs with extreme Actual Ratings, could be an area of future exploration.

Classification results: In response to the low R^2 values we found with a regression model, we tested how effective our model is at predicting the relative class of a design metric. Our classes are determined by the percentile of a design's rating with regard to all other designs in a test set. Designs with ratings between the 1-33 percentiles are in the low class, designs with ratings between the 34-67 percentiles are in the medium class, and designs with ratings between the 67-100 percentiles are in the high class.

The results are demonstrated in the confusion matrix in Table 2.2. The confusion matrix is a machine learning concept that contains information outputted from a classification system about actual and predicted outputs [104]. Confusion matrices typically have two dimensions, where one dimension categorizes the information by the actual outputs and the other categorizes the information by the predicted outputs [42]. Together, these two dimensions can illustrate agreement between the predicted and actual outputs, where perfect classification is when the matrix only has numbers along the diagonal and zeroes everywhere else. Not all errors are equal in confusion matrices. Predicting a high value when the true value is low (or vice versa) is an extreme error, which we aim to minimize the most.

In Table 2.2, we see that we predict a low class when the real class is high 7 times, and do the opposite 11 times. This compares to correctly predicting the low class 39 times and the high class 40 times.

We also observe the precision and recall metrics. Precision indicates how many positive predictions are true, and recall, also known as the true positive rate (TPR), measures how many of the positive cases our model is able to correctly predict. They are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

For the low, medium, and high classes the precision values are 0.619, 0.387, and 0.645 respectively.

For the low, medium, and high classes the recall values were 0.591, 0.462, and 0.580 respectively. The high recall values show that our regression model is able to effectively classify designs into low, medium and high categories, which can enable human raters to use these models for initial filtering of ideas and then focusing on individual categories to identify the top ideas.

2.5 Discussion

The goal of this study was to identify how to take advantage of the distinct strengths of both the SVS and CAT method through machine learning. More specifically, this study sought to investigate the possibility of using machine learning to facilitate automated creativity assessment. Our results revealed two major and consistent trends:

Trend 1 is displayed visually in Figure 2-3. Each design representation, from one to three, incorporates more text and NLP in the design representation. On the figure this is visually displayed with the lightest bar incorporating the least NLP and the darkest bar incorporating the most. Across all design metrics a general trend emerges. In design representation three, Text Embedded SVS Features and Description tends to outperform both other design representations, while design representation two tends to outperform design representation one.

We propose that this trend is due to Universal Sentence Encoder's ability to capture relationships between features by mapping them in a continuous space. Universal Sentence Encoder's ability to do this is in contrast to one-hot-encoding vectorizations, which assume all features are independent. We note that the overall R^2 values are low. Our predictions have much room for improvement, as visualized in Figure 2-4. However, the emerging trends still provide insight that we find valuable in how we can predict classically subjective ratings of designs objectively.

The other emerging trend that we found compelling is the consistent difference in the ability to predict certain design metrics. We found that Usefulness and Elegance metrics were consistently the most predictable. This trend could result from the objective nature of Usefulness and Elegance. More raters may agree on what characteristics of a design qualifies as useful. Furthermore, elegance is tied to the simplicity of a design, which can be more objectively agreed upon than, say, creativity. To that note, Creativity and Uniqueness were the least predictable across all regression models. We attribute this to the subjectivity of these ratings. In fact, Creativity is often not agreed upon by experts, and expert ratings are the basis of our regression models.

This finding opens another discussion centered around the availability of expert vs. novice design ratings. Novice design ratings are less expensive and more easily acquired than expert design ratings. However, expert design ratings naturally hold more weight - novices must be trained to be 'expert-proxies' using some sample set from expert ratings. Even when you can train novices to be 'expert-like,' they still lack the mental models and experiences of experts which ultimately impact rating performance. However, some areas do exist where experts and novices tend to agree such as in the novelty of design ratings [85]; thus, one could argue that there are instances where novice ratings can supplement or replace expert ones.

Following the discussion of expert vs. novice design ratings is that of the type and level of automation our study has achieved. We note that the initial SVS survey needs to be done for all designs. This survey, however, is objective and does not need to be performed by experts. None of the questions directly asks the surveyor to assess the creativity of the design. So while this survey does need to be done manually, we have explored whether we can automatically predict an expert creativity assessment rating from objective, non-expert survey results.

Our findings have inspired work that combines sketch understanding via computer vision methods and text understanding via NLP to predict design metrics [119]. Future work in this direction may be be training models to identify design features from the sketch and text description. Such a model can augment our understanding of sketches and help in automated categorization and assessment of design sketches.

2.6 Summary

Creativity and innovation are important steps in the development of novel solutions to existing and new problems and for making important technological progress [12, 80, 7]. As a result, both creativity and innovation have often been viewed to be mankind's most valuable resources [24]. There have been many attempts to help boost creativity. One important step that follows the implementation of improvement methods is creativity evaluation, which helps identify whether progress has been made. Currently, the most common evaluation methods used by researchers are the CAT and the SVS method. However, both techniques have their advantages and disadvantages. While CAT has been widely accepted as the "gold standard," it relies heavily on the subjective judgement of human experts in that domain, and is highly resource consuming. The SVS method, on the other hand, is less resource intensive but has been criticized for lacking sensitivity and accuracy. In light of these complementary strengths and weaknesses, this study was created to investigate the possibility of using machine learning to facilitate automated creativity assessment. More specifically, this study seeks the possibility of taking advantage of both methods by incorporating machine learning to use SVS ratings, which are easier to collect, to predict design metrics. This is done by using regression models in the prediction process, and also by exploring the possibility of using NLP based models to improve the results.

Our results, although preliminary, show that incorporating NLP in the prediction process can improve the model's prediction of design metrics, including CAT ratings. This study also found that the predictability of different aspects of the design metrics vary, with Usefulness and Elegance having the best predictability. These findings can serve as empirical evidence supporting the investigation of novice vs expert usage in creativity assessment. In addition, the results can also serve as empirical evidence of the plausibility of using machine learning to facilitate creativity assessment. The preliminary success in using NLP to predict design metrics can help to show the wide application of NLP, and also support its usage in modeling subjective ratings like creativity.

Being able to model and evaluate design metrics provides basis for automatically filtering design ideas based in these metrics. We hope that the results of this study can shed some light on the ongoing debate of using novices vs experts in creativity assessments. The potential of utilizing a small number of expert assessments to train a model to then predict future design idea assessments can decrease the level of effort of expert reviewers. This work addresses two common challenges in idea filtering, that manual filtering by experts is expensive and time consuming, and that large labeled datasets are not frequently available and are expensive to generate.

In this chapter we have explored the potential for NLP to accelerate idea filtering of creative early-stage designs with a small dataset of 934 documents. In the following chapters, we will look at automated idea filtering of text documents and datasets of over 50,000 documents.

Chapter 3

Accelerating Evidence Synthesis for the Design of International Development Projects Using Natural Language Processing-based Idea Filtering

3.1 Introduction

Our goal is to use state of the art natural language processing (NLP) methods to perform automatic document inclusion classification to aid the creation of evidence gap maps (EGMs), a form of evidence synthesis.

We are currently in an information explosion. The National Science Foundation reports that in 2018 alone, global research output in science and engineering was 2.6 million articles, and that it grew at a rate of 4% annually from 2008-2018 [49]. A person's capacity to understand all available research is limited. This poses a problem in scenarios like designing policies, in which decision-makers seek a full understanding of all available research. To make informed decisions and understand the growing corpus of research available, researchers have turned to evidence synthesis. Evidence synthesis refers to the process of compiling information and knowledge from many sources and disciplines to inform decisions [44, 117]. However, creating evidence synthesis products like EGMs requires extensive time and effort from human experts.

We are motivated to reduce the manual level of effort and human hours required to create EGMs, while maintaining or improving their robustness and quality. By automating aspects of the EGM creation process, we can reduce the amount of work human experts have to do, and in turn produce EGMs quicker. These EGMs can be used to inform policy with all relevant evidence.

NLP has long been used to promote evidence based decisions in the medical field. One example of this is BioMedICUS, which is a system for large-scale text analysis and processing of biomedical and clinical reports being developed by the Natural Language Processing and Information Extraction Program at the University of Minnesota Institute for Health Informatics [99]. NLP based models have since been used in other domains like the legal field to automatically extract information from text [19]. However, there are particular challenges to applying NLP to literature in the social sciences because social science papers, including titles and abstracts, are less likely than those in medicine or the "hard" sciences to adopt standardized structures and terminology [92, 116]. These challenges motivated us to apply NLP to understanding and classifying text in the international development domain. We propose incorporating state of the art transformer-based NLP models in order to understand, extract information from, and classify international development text will allow for the quicker creation of EGMs.

Many information extraction techniques already build off of NLP in order to interpret and identify important parts of documents. These automated information extraction techniques can reduce the time required to create EGMs by identifying key information about a document and clustering similar documents. This reduces the time needed to provide EGMs to policy makers and researchers. However, recent improvements in NLP have yet to be incorporated in information extraction for the field of EGMs. We propose incorporating state of the art transformer-based NLP models in order to extract information such as study type, population, intervention, and outcome from unstructured text and ultimately aid and quicken the typically manual creation of EGMs.

Exploring the fields of NLP and EGMs reveals a gap in the literature. While there has been great success in the field of NLP, few studies have specifically focused on NLP for text within the policy and international development domain. The work that has successfully done so may be improved upon by incorporating the latest transformer and transfer learning based methods. By doing so, we can specifically aid the creation of EGMs, which are valuable tools for decision-makers in policy and research funding.

Through collaborations with members of 3ie we have identified the most useful ways to aid EGM creation through machine learning. We aim to create a classification model that determines whether a document should be included in an EGM based on its title and abstract. We also use rule-based models and transformer-based models to perform information-extraction for entities that are relevant to EGM creation in the international development field: country of study, language of document, and type of study. To improve the performance of the classification model, we will explore whether merging the information extracted by the rule-based model into the transformer-based model helps.

3.2 Related Works

Here we discuss the fields of natural language processing (NLP), information extraction (IE), evidence gap maps (EGMs), and active learning (AL). NLP and IE have developed hand-in-hand, while EGMs and AL have each grown largely independently for many decades. We explore existing and potential work that lies at the intersection of these fields in the following sections.

3.2.1 Natural Language Processing

NLP is a field of machine learning in which computational machines are trained to understand text and spoken language. Historically in NLP words are represented as vectors where similar words are located near each other in continuous space [84, 83, 95]. Early research in NLP built off of these word representations and recurrent neural networks (RNNs) [96]. RNNs take in sequential inputs, which proved useful with data like language where the ordering of words and sentences is important. However, RNNs face the problem of vanishing and exploding gradients, which limited the length of sequences, or the number of words that a model could intake. This problem was solved by the development of Long Short-Term Memory (LSTM) which applies additional computation in the hidden layers of RNNs [58]. RNNs already took a long time to train, and while LSTMs solved issues in vanishing and exploding gradients, the cost was even longer training times. The need for sequential inputs in LSTMs does not utilize the computational power of modern GPUs which allow for the parallelization of tasks.

In 2017, Vaswani et al. introduced the transformer, which allowed inputs to be fed in in parallel and had state of the art results in many NLP tasks [128]. Since then, many NLP models have built off of the transformer and have produced state of the art results in NLP tasks such as language translation, question answering, reading comprehension, and sentiment analysis [43, 30, 76, 134]. One of the most well-known transformer-based language models is BERT, which stands for Bidirectional Encoder Representations from Transformers [43]. From BERT, many other researchers have made various improvements by adjusting aspects of the model like the pre-training objective, training dataset, architecture, and transfer learning technique used. In response to the large corpus of research done to make small adjustments to the BERT architecture, researchers at Google produced Google T5, a model which explores the landscape of transfer learning techniques to find the optimal combination of NLP techniques [100].

NLP has long been used to promote evidence based decisions in the medical field. Models such as BioMedICUS perform large-scale text analysis and processing of biomedical and clinical reports [99]. The success of NLP in the medical field has led to its use in other fields, with models like LexNLP which automatically extracts information from legal text [19]. NLP has also been utilized in the design field, for example to predict expert design creativity ratings from a design and its descriptions [45].

3.2.2 Information Extraction

Information extraction (IE) is a high level task within NLP [114]. IE specifically aims to automatically retrieve (recognize and extract) certain types of information from natural language text [131]. In this sense, IE can also be understood as taking unstructured language data and producing structured information specified by certain criteria [114]. IE subtasks include:

- Named entity recognition
- Temporal information extraction
- Coreference resolution
- Named entity linking
- Relation extraction
- Knowledge base reasoning

In order to perform these tasks, some kind of model that specifies what to look for (e.g. country name, study type, population, outcome) is needed to guide the process [131].

In 2018, Singh et al. produced a survey of state of the art IE methods. They present three main categories for IE approaches: pattern matching (identifying speech patterns like Inc. or Co. coming after a company name), gazetteer (pre-defining a complete dictionary of words, like all country names, and searching for words included in that dictionary), and machine learning (algorithms automatically learn IE patterns from a set of training data) [114].

Machine learning has produced state of the art results for many IE subtasks, and produced models like BERT that are capable of performing multiple subtasks at once [43, 76, 100, 131]. Researchers have successfully produced IE models that are specific to certain industries. For example, LexNLP was developed to perform NLP based IE on legal and regulatory texts [19]. LexNLP's key functionality is to take unstructured legal text and a) segment documents, b) identify key text such as titles and section headings, c) extract over eighteen types of structured information like distances and dates, d) extract named entities such as companies and geopolitical entities, e) transform text into features for model training, and f) build unsupervised and supervised models such as word embedding or tagging models.

LexNLP is specific to the legal field, but it was largely inspired by success in applying IE to the medical field [19]. Such examples include BioMedICUS, which is a system for large-scale text analysis and processing of biomedical and clinical reports being developed by the Natural Language Processing and Information Extraction Program at the University of Minnesota Institute for Health Informatics [99]. These domain specific IE programs demonstrate the success and exigence of NLP. However, many of these tasks were trained specifically for their domain. Additionally, many of these programs do not yet incorporate state of the art NLP techniques, such as transformer-based models.

3.2.3 Evidence Gap Maps

EGMs are one form of evidence synthesis - the process of compiling information and knowledge from many sources and disciplines to inform decisions [44, 117]. Evidence synthesis provides more reliable information about a topic than a single study by systematically collecting, categorizing, and analyzing a broad range of studies [29]. Thus, evidence synthesis is an incredibly valuable tool for decision-makers seeking to design policies and fund research [44].

The International Initiative for Impact Evaluation (3ie) has pioneered the use of EGMs, which present a visual overview of completed and ongoing impact evaluations and systematic reviews in a specific sector [3]. 3ie creates these EGMs via the "thematic [collection] of information about impact evaluation and systematic reviews that measure the effects of international development policies and programme" [2]. The final product is a matrix, organized by "intervention" categories on the vertical axis and “outcome” categories on the horizontal axis. Interventions are the action taken in the study, and outcomes are the result of the action. Each cell of the matrix contains studies that rigorously evaluate the impact of a specific intervention on a specific outcome.

3ie sets the global standard for EGMs, and the mapping method has been adapted by organizations including the Campbell Collaboration, the World Bank Independent Evaluation Group, and USAID [3]. Like other forms of evidence synthesis, EGMs begin with an expansive and systematic search of scholarly databases and ”grey literature” sources (such as repositories of government documents or websites of think-tanks) to identify potentially relevant studies. EGM teams then screen these search results to identify studies that meet the EGM’s criteria for interventions evaluated, outcomes measured, implementation setting, and study design. Once eligible studies are identified, the EGM team extracts information on interventions, outcomes, and other key characteristics of each study to determine its placement in the EGM matrix and to allow for analysis of trends in the literature.

3ie uses a software called EPPI-Reviewer which aids in the creation of EGMs. While EPPI-Reviewer has some machine learning functions that can accelerate screening [92], most EGM tasks are still performed manually. Thus, each EGM requires significant human effort and expertise, with typical EGMs taking nearly six months to complete [117]. Given that one of the main barriers to evidence use among policymakers is the lack of timely research outputs [89], there is a critical need to reduce the time and effort needed to complete EGMs and other evidence synthesis products.

The creation of an EGM is shown in Figure 3-1. Our work focuses on step 3, in which reviewers screen documents for inclusion in an EGM based on their title and abstract. Selected documents will move on to full-text review. We create three transformer-based NLP models that automatically classify documents for inclusion at this step.

In 2020, Porciello et al. introduced a machine learning model, Persephone, that aids humans in the creation of EGMs [97]. Persephone analysed 500,000 unstructured text summaries from prominent sources of agricultural research, and determined with

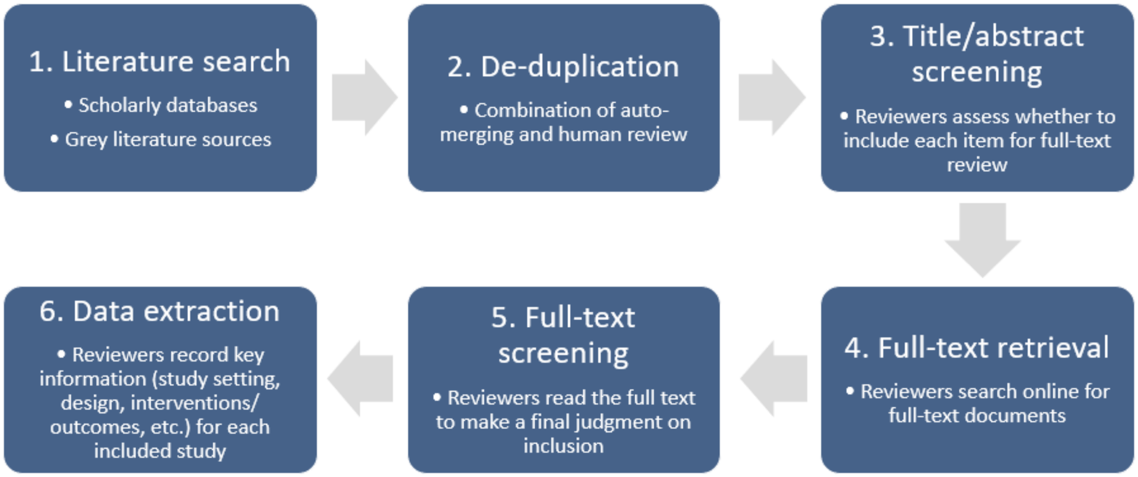
1. Literature search
• Scholarly databases
• Grey literature sources
2. De-duplication
• Combination of auto-merging and human review
3. Title/abstract screening
• Reviewers assess whether to include each item for full-text review
4. Full-text retrieval
• Reviewers search online for full-text documents
5. Full-text screening
• Reviewers read the full text to make a final judgment on inclusion
6. Data extraction
• Reviewers record key information (study setting, design, interventions/ outcomes, etc.) for each included study

Figure 3-1: A high level view of the current EGM creation process.

90% accuracy the subset of studies that would eventually be selected by expert researchers [97]. This work builds from past success in incorporating machine learning in evidence synthesis within the health sciences community [81, 59].

The results of this paper exemplify the effectiveness of machine learning models in reducing the human effort to create EGMs. The research showed that machine learning model assistance allows human reviewers to screen and extract relevant information from large datasets more quickly. Human and machine learning teams could reduce datasets by at least 50% in three days, allowing experts to dedicate their time to fewer documents and ideally hasten the creation of an EGM [97]. Additionally, the results showed that document inclusion decisions made using information extracted from machine learning models match those decisions made from an entirely manual process with an average accuracy of 90% [97].

This work demonstrates the great potential of machine learning aided EGM creation. We can build upon this work by incorporating the most novel and effective NLP models. For example, we will move away from some of the rule-based information extraction techniques used in this work, and incorporate state-of-the-art NLP models instead.

Additionally, the model created by Porciello et al. automatically extracts information that a human reviewer can then observe to decide if a document should be included. However, the model does not provide a probability of inclusion itself. We can expand upon this work by developing an inclusion vs. exclusion classification model based on the text and information we extract.

3.3 Methods

In the following sections we discuss our methodology including the dataset we used, our data preprocessing, and the three NLP-based classification models we created. We also describe how we tested our models with different training data sizes in order to gain insight on how many documents must be manually labeled in real-world applications of our process.

3.3.1 Dataset Description

Our data is provided by 3ie and is derived from manually labeled documents from 3ie's Development Evidence Portal (DEP) [1], an expansive repository of impact evaluations and systematic reviews in international development across a wide range of sectors. For the classification model, we utilize a dataset of 53,521 documents from 3ie's DEP. For each document we have the following information: title, year of publication, abstract, and inclusion decision.

For the information extraction tasks that support the integrated model, we utilize a dataset derived from previously created EGMs. For each document in the dataset we have both the title/abstract and the manually labeled information relevant to the model we are training. There are approximately 75 fields in the dataset, encompassing:

1. Bibliographic metadata
2. Thematic keywords
3. Country of study
4. Study methodology

5. Funding sources for the program and research
6. Populations targeted
7. Interventions and outcomes
8. Whether the study reports on the cost of the intervention or performs cost-effectiveness analysis

3.3.2 Data Pre-processing

To guarantee the performance of the classifier models, the original documents were pre-processed for noise removal.Two types of noise were identified and removed in this step. The first type is non-English texts. A part of the documents provide titles and abstracts in multiple languages. Since the classier models considered in this paper only take text in English as input, all the sentences in other language than English are noise to the models and should be removed. A substantial portion of the the original documents contain content not relevant to the topic of the the document, such as copyright statement. The pre-processing consists of three steps: (1) Each document is parsed into sentences; (2) the sentences from 500 documents were labeled manually as relevant content or irrelevant content; (3) a BERT classifier model was trained on the manually labeled data to identify irrelevant sentences automatically. The accuracy of the model is higher than 0.99. After the irrelevant sentences were identified and removed, the relevant sentences in English were integrated back to the abstracts of the original documents.

3.3.3 Universal Sentence Encoder Model

We utilize Google's Universal Sentence Encoder (USE) embeddings and a neural net classifier as our base model. USE is a pre-trained model that encodes text into embedding vectors [31]. An embedding vector is a numerical vector representation of text in which text strings with similar meanings are represented by similar vectors.

Google's USE is pre-trained on a variety of web sources including Wikipedia, web news, web question-answer pages, and discussion forums [31]. In the 2018 paper in

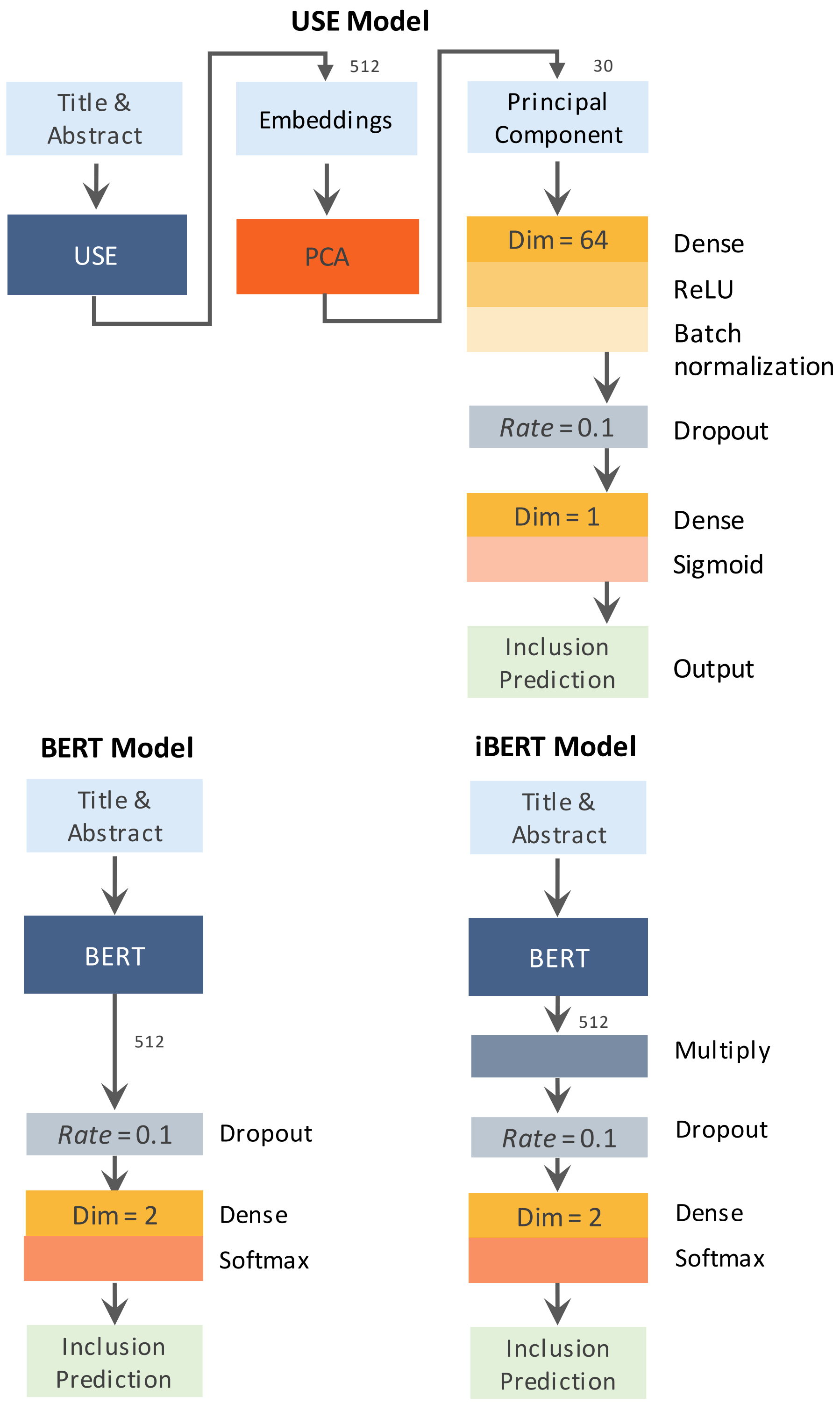

Figure 3-2: The architecture of our three transformer-based models.

which USE was introduced, Google provided two models: one that is transformer-based and targets high accuracy at the cost of greater model complexity, and one that uses a Deep Averaging Network which targets efficiency with reduced accuracy. We utilize the transformer-based USE model. This model builds from the encoding portion of the transformer architecture [128] in order to produce text embeddings.

The USE model's architecture is shown in Figure 3-2. For each document, we input a concatenated string of the document's title and abstract. The USE model outputs a 512 dimensional vector that represents the inputted text string. We perform principal component analysis (PCA) reduction from 512 dimensions to 30 in order to decrease our feature space and therefore the number of parameters our model must learn. Lastly, we built a neural network with one hidden linear layer, batch normalization, dropout, and a rectified linear unit ReLU activation function. This neural network serves as a binary classifier, it produces the probability ($\mathcal{P}$) of including a document, and from this determines the class: included if $\mathcal{P} > 0.5$, and excluded if $\mathcal{P} < 0.5$.

3.3.4 Fine-Tuned BERT Model

The second model we utilize is a pre-trained BERT model for sequence classification that we then fine-tune using our dataset. We show the architecture of this model in Figure 3-2. Our expectation is that this model's performance will be higher than the USE model since it learns domain specific information when being fine-tuned with a corpus of international development text.

We specifically use the uncased BERT base model [43]. This model is comprised of twelve modules of transformer blocks, each transformer block has sixteen layers, and the model has a total of 110 million parameters. Because our implementation of the BERT model is almost identical to the original, we encourage interested readers to learn more about the model in the original paper [43].

A the key strength of the BERT model is that it enables the use of transfer learning that transfers language patterns learned from very large dataset to smaller target datasets. The base BERT model is pre-trained with unlabeled data on two unsupervised tasks (1) masked language modeling, and (2) next sentence prediction. This trains the 110 aforementioned parameters. Compared to the original BERT model, the pre-trained BERT model specifically for sequence classification additionally include a dropout layer and a dense layer was the output layer. In addition to the parameters of the BERT model the weights of the output layer are also pre-trained and transferred to the downstream tasks. Next, we fine-tune a subset of the parameters with our dataset of text on a supervised classification task. We use the manual inclusion and exclusion classifications, and train a binary classifier to predict probability of inclusion from the title and abstract of a document. Similar to our USE model, we input a concatenated string of a document's title and abstract, and we output the predicted class.

3.3.5 Integrated BERT Model

The fine-tuned BERT model is purely trained on the pre-processed texts of the documents. NLP models, such as our fine-tuned BERT, typically ignore domain knowledge, however research has shown that expert domain knowledge greatly improves model performance in NLP tasks [93, 130]. Domain knowledge is particularly critical for idea filtering, in which many ideas within a domain must be evaluated based on criteria that is often nuanced and specific to that domain. Furthermore, in the manual screening process humans decide whether a document should be included or not based on multiple domain specific criteria, such as the study type, country of study, and presence of an intervention.

Given the importance of domain knowledge for idea filtering in the manual process, we were inspired to integrate the information extracted for assessing whether a document satisfies a certain inclusion criterion into the BERT model. On this basis, we developed an integrated BERT (iBERT) model by integrating the information regarding country of study and study methodology into the BERT model.

In the human screening process, all documents only focusing on high-income countries and all documents using laboratory study methods are excluded. Two models have been developed to respectively assess theses two criteria, including: (1) a rule-based country name search model was has been developed to extract country information from document texts and detect whether a document should be excluded due to only involving high-income countries with an accuracy of 93%; (2) a BERT classifier model to detect detect whether a document reports a laboratory study or not with an accuracy of 97%. The outcomes from these two models are integrated into the BERT models as two indicators that are multiplied with the text embeddings from the BERT model before the embeddings are input into the dropout layer. Specifically, if a document only involves high-income countries, it get an indicator of 0 for this criterion, otherwise the indicator is 0; if a document focuses on a laboratory study, the indicator for it is 0, otherwise it is 1. In such an approach, any document only evolving high-income countries or focusing on laboratory studies gets an embedding of all 0s as input to the last classifier layer. The architecture of the iBERT model is shown in Fig. 4.

3.3.6 Training Size Variation

We test the effect of training size on the accuracy of our classification models. Our complete dataset is comprised of 53,521 documents, so we try different train-test splits of our data. We explore the effect of training with different training data sizes ranging from 100 datapoints to 40,000, as shown in Figure 3-4. For each training data size, we run the model five times with different and randomly selected training data each time. We also present the average accuracy for each model when training on 70% of the dataset, or 37,472 documents. These results are shown in Figure 3-3.

3.4 Results

In this section we demonstrate our results, namely comparing the three models we introduced in the Methods section and exploring the effect of training size on the accuracy of the USE and fine-tuned BERT model.

Figure 3-3 shows the model accuracy for our three models: USE, fine-tuned BERT, and the iBERT model. We find that the iBERT model performs the best, slightly outperforming the fine-tuned BERT model from which it is largely derived.

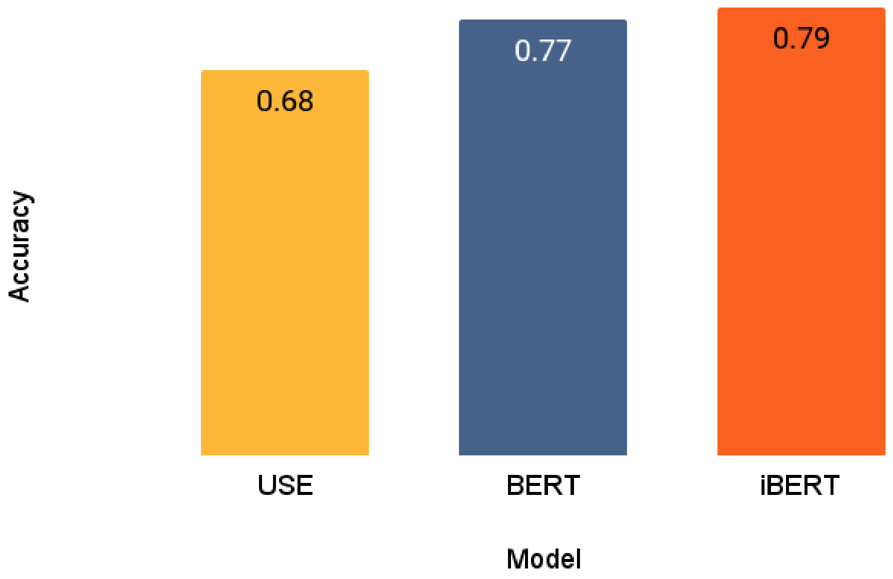

Figure 3-3: The average classification accuracy of each model when trained on 70% of the data. We observe that the iBERT, which combines the fine-tuned BERT model with extracted information performs the best.

This result matches our intuition, as the iBERT model builds from the fine-tuned BERT model, but also directly incorporates information automatically extracted from the text. We chose to extract information that human experts look towards to decide inclusion and exclusion decisions: the country of study and the study type. Our results suggest that including this extracted information improves the model's ability to decide whether a document should be included or excluded in and EGM.

We also observe how training size impacts the accuracy of the USE and fine-tuned BERT models. Our results are shown in Figure 3-4. We observe a logarithmic relationship between training size and accuracy. Accuracy increases with training size, but the rate of improvement slows past a training size of approximately 10,000 documents.

Overall, we see that the fine-tuned BERT model has a higher classification accuracy than the USE model at all training sizes. This suggests that the fine-tuned BERT model is better able to predict whether a document should be included in or excluded from an EGM based on the document's title and abstract.

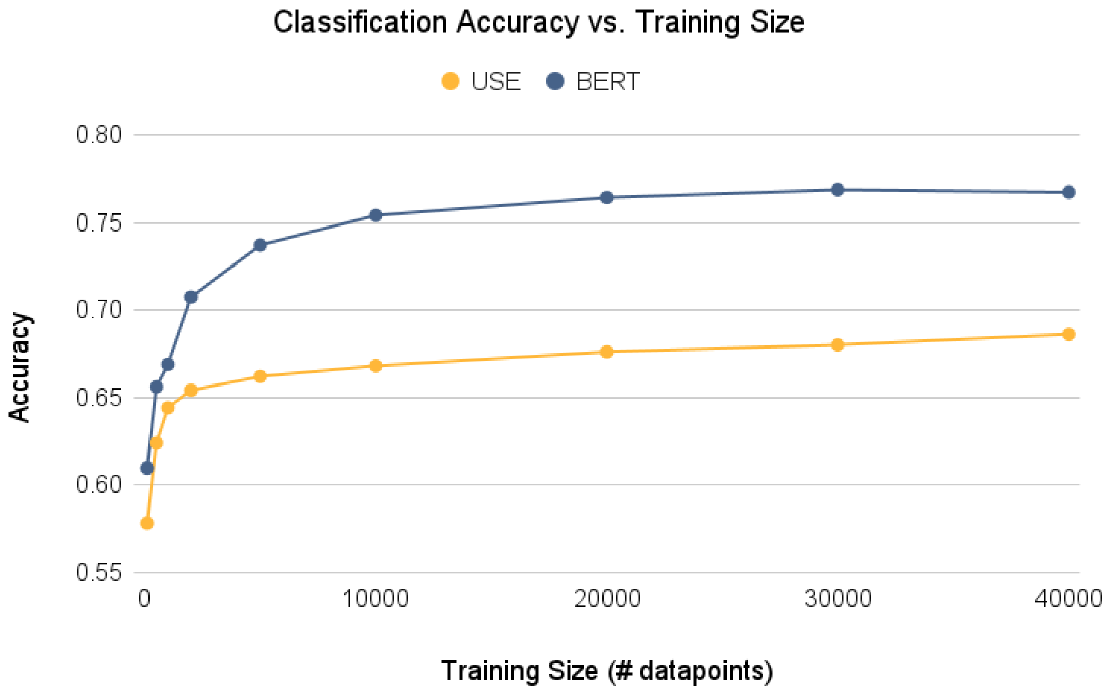

Figure 3-4: Model classification accuracy for different training sizes. Both models are influenced by training size, but we note that BERT outperforms our USE model at all training sizes.

We only study this trend for the fine-tuned BERT and USE models, and not the iBERT model because the iBERT model requires information extraction models that are trained on another dataset. Therefore it is inaccurate to present iBERT trained on smaller datasets, since it would require at least the larger information extraction training data. We note, however, that iBERT is built off of and tends to slightly outperform the fine-tuned BERT model, as shown in Figure 3-3. We therefore hypothesize that a training size that is successful for the classification model of the fine-tuned BERT will also be successful for the classification model portion of iBERT.

Our motivation for exploring the influence of training size on accuracy was to strike a balance between improving the accuracy with more training data and the level of effort required to generate the manual training data. We can utilize these results in real-world applications of our models, in order to ensure that the amount of training data needed to achieve the desired accuracy is feasible.

3.5 Discussion

Our research reveals two overall findings:

1. The iBERT model which combines the fine-tuned BERT model with additional extracted information has the highest accuracy.

2. Accuracy increases with training data size, but the rate of increase slows significantly after 10,000 documents. This demonstrates that just a small training dataset of 10,000 documents can produce very high accuracy, leading to saved effort in training dataset creation and in inclusion screening.

These results align with our intuition. Both the USE model and the BERT models transfer patterns learned from very large datasets to our own data by using the pre-trained embedding modules to encode the titles and abstracts of the documents.

Why the BERT-based models outperform USE The key differences between the USE model and the BERT models lies in two aspects. First, the multi-head attention mechanism and the deep bidirectional learning adopted in BERT enables it to capture more delicate features from texts. BERT is also trained on a larger set of data. These factors make BERT a better model than USE in general for NLP tasks. Second, the BERT models in this study are fine-tuned, which allows the models to learn features from our own dataset, while the USE embedding model are totally frozen with the embeddings being fixed. These two aspects explain why the BERT models perform better than the USE model.

Why iBERT outperforms fine-tuned BERT The iBERT model exhibits better performance than the BERT model alone, since more information of the documents are encoded into the embeddings. For example, the rule-based model for detecting high-income countries first extracts country information from the text and then assesses whether a country is a high income country based on the data from World Bank. The second part of the information cannot been learned from the texts directly. Similarly, since the model for laboratory study detection was trained on different labels, the model learned different features from the texts. When both indicators are encoded to the BERT model, the iBERT model receives more information of the texts, improving its performance.

Effects of training data size The marginal effect of increasing training data size decreases as the training data size gets larger. This is because when the training data size is small, it is more likely that the newly added data can bring data that have not been seen by the model; when the training data is large enough, the newly added data largely repeats the data that have been seen by the model and hardly bring brand new data.

Our findings about the influence of training size on model accuracy can be used during real world applications. Since effort is required to generate a manually labeled training dataset, human-AI EGM-creation teams face a tradeoff between maximizing model accuracy while minimizing level of effort to create a training dataset. Our results suggest that model accuracy increases greatly with training size until about 10,000 datapoints. After this point we notice diminishing returns. We anticipate this being a very useful result when applying these models, since users can make informed decisions regarding accuracy vs. training data size.

Applying NLP models to evidence synthesis tasks has significant untapped potential to increase efficiency and reduce the time and human effort required to produce high-quality evidence syntheses. Preliminary testing with the models described in this paper suggest that by automatically excluding documents with inclusion scores $< .2$, a synthesis team could reduce the screening load by 27% while still achieving over 99 percent recall. Although experienced screeners can screen 50-75 abstracts per hour, EGM searches routinely retrieve 50,000-100,000 documents. Thus, the reduction in screening load represents a significant reduction in person-hours required to identify relevant documents.

3.6 Summary

The aim of this thesis is to accelerate the design process and lower expert level of effort through automated idea filtering. One challenge of manual idea filtering is that it is both time and resource intensive for experts to filter through a plethora of ideas. In this chapter, we explore idea filtering with a dataset of over 50,000 textual documents, which would entail significant effort from experts. Our research explores the effectiveness of three NLP classification models on evaluating documents for inclusion in an international development evidence synthesis. We find that an iBERT model comprised of a fine-tuned BERT model with additional extracted information performs the best with an accuracy of 79%. We further explore the influence of the amount of training data on the accuracy. As our intuition suggests, accuracy improves with the amount of training data used. However, the rate of improvement slows down as the amount of training data increases, and after about 10,000 training datapoints the improvements are small. Our results can inform decisions between increased accuracy and the level of effort to manually create training data in real-world applications of this work. Our work is motivated by the potential of NLP to automate time-intensive aspects of evidence synthesis models like filtering through tens of thousands of text documents. By automatically and accurately filtering documents for inclusion in an evidence gap map, we can quickly provide decision-makers with the tools to design evidence-based policies.

In the next chapter, we will investigate the creation of a training dataset used to train a model for automated idea filtering, and how we might decrease the expert effort required to label such a training set.

Chapter 4

Active Learning for Real-World Imbalanced Natural Language Processing Applications

4.1 Introduction

After demonstrating the efficacy of using NLP to accelerate literature screening in Chapter 3, we aimed to next decrease the level of effort required of expert reviewers to generate training data. The NLP models that we utilized must be fine tuned with labeled documents. In the engineering and design fields, obtaining large labeled datasets is difficult as there is a lack of large publicly available design datasets [103, 93]. Generating a labeled training set is no trivial task, and is often expensive and resource intensive as it requires labels from expert reviewers [93, 39]. To add to that, in most real-world evidence synthesis scenarios, the positive class of included documents will represent a small percentage of the total literature base. Therefore, obtaining a training set that includes instances from the positive class may require labeling many more documents than desired. Given our goal of decreasing human effort and accelerating the design process through automated idea filtering, we turned to a strategy that was developed to increase model performance while decreasing the number of required training points: active learning.

While most machine learning models involve the learner passively receiving training data, this is divergent from human learning tasks. In many human learning tasks, the learner interactively makes queries, takes actions, or performs experiments on specific data to best gain understanding [38]. Active learning is the concept that a machine learning algorithm can perform better with less training data if it is allowed to choose the data from which it learns [38, 109]. Interested readers can find a survey of classical active learning approaches from Burr Settles in [109].

We experiment on classifier models with four different data selection scenarios with our highly imbalanced real-world dataset. The four data selection scenarios have varied methods of selecting the initial training set and selecting additional instances to add to the training set. We compare the model performance as measured by the F1 score for scenarios that begin with a class-balanced vs. random initial seed, and for scenarios that utilize active learning to select additional training data vs. scenarios that do not. Our contributions are as follows, we demonstrate that:

- Incorporating active learning in data selection results in an F1 score of 0.78, which is 7.7% higher than the baseline data selection technique.

- Incorporating active learning (AL) in data selection results in more balanced, and consequently more diverse, training datasets.The two techniques incorporating AL end up with training sets comprised of 30% and 33% of minority class instances, while the two techniques without AL result in 7% and 16% minority class training sets.

- If an AL technique is used to identify new samples to query, then a random initial training set will perform *just as well* as a balanced initial training set. Since obtaining a balanced initial training set requires non-trivial expert labeling effort, an equally high performance from the random initial training set means we can propose a technique that maximizes model performance while minimizing human expert level of effort.

4.2 Related Works

Active learning has been a prominent research area for nearly three decades, but recent research in the field continues to explore the benefit of applying active learning to deep learning problems such as image classification [4, 50], speech recognition [126], data exfiltration detection [36], and many NLP tasks [90].

Active learning (AL) is desirable because, in many instances, obtaining labeled training data is expensive and time consuming [47, 109]. Training data often come from human experts. For example, accurate speech labels for speech recognition problems require trained linguists, and [135] reports that annotation at the word level can take ten times longer than the actual audio being annotated. Furthermore, manual information extraction can take half an hour for even brief news articles [110]. Settles notes specifically that for tasks like our own, classifying documents as “relevant” or “not relevant” to specific groups may classically require thousands of annotated instances, and procuring these can be both “tedious and even redundant” [109].

AL combats the problem of data scarcity by asking “queries,” generally in the form of unlabeled instances, to be answered by an “oracle,” often a human annotator or an existing labeled dataset [109]. There are three main problem setups, or scenarios, in which a learner may be able to ask queries: membership query synthesis [15, 16], stream-based selective sampling [18, 37], and pool-based sampling[69]. I will focus primarily on pool-based sampling, which assumes there is a small set of labeled data $\mathcal{L}$ and a large pool of unlabeled data $\mathcal{U}$ available. Queries are selectively drawn from $\mathcal{U}$ in a “greedy” fashion, according to some informativeness measure that has been used to evaluate all of the instances in the pool. The key attribute of pool-based sampling is that in this scenario, the model evaluates and ranks the entire unlabeled pool for informativeness, and then selects the best queries [109].

Just as there are different scenarios for AL, there are also several different methods for choosing which unlabeled instances from $\mathcal{U}$ to query. The decision method is called the *query strategy.* The most commonly used query strategy is uncertainty sampling, introduced in [69]. In this strategy, the learner queries the instances for which the learner is least certain how to label [109]. Within the uncertainty sampling strategy there are three primary measures that evaluate how uncertain the learner is about each instance: least confidence [40], margin sampling [107], and entropy [113]. The equations for these sampling techniques are as follows:

$$x_{LC}^{*} = \underset{x}{\operatorname{argmax}}\, 1 - P_{\theta}\left(\hat{y} \mid x\right) \tag{4.1}$$

$$x_{M}^{*} = \underset{x}{\operatorname{argmin}}\, P_{\theta}\left(\hat{y_1} \mid x\right) - P_{\theta}\left(\hat{y_2} \mid x\right) \tag{4.2}$$

$$x_{H}^{*} = \underset{x}{\operatorname{argmax}} - \sum_{i} P_{\theta}\left(y_i \mid x\right) \log P_{\theta}\left(y_i \mid x\right) \tag{4.3}$$

Ein-Dor et al. explore using AL in a realistic case - one in which data is both imbalanced and scarce [47]. These are two characteristics that are quite common in real world problems, and characterize the problem we are researching as well. Works such as [4] and [47] explore how AL can work for highly imbalanced datasets. Typically, high data imbalance presents a challenge for both deep learning and AL, as the low prior of the minority class often leads to low recall of that class. Aggarwal et al. present a method for combating data imbalance by using a large model pretrained on a source domain to start with, and then modifying the query strategies, which they refer to as acquisition functions, to focus on not just sampling for instances with high uncertainty, but also ones that most represented the overall dataset [4]. This research is in the computer vision domain, rather than NLP domain, however the framework can be applied to the imbalanced AL task we research here.

4.3 Methods

In this section, we describe our methodology for employing AL to an NLP problem using a real, highly imbalanced dataset.

4.3.1 Dataset

Similar to the dataset described in Chapter 3, our data is provided by 3ie and is derived from manually labeled documents from 3ie's Development Evidence Portal (DEP) [1], an expansive repository of impact evaluations and systematic reviews in international development across a wide range of sectors. For the classification model, we utilize a dataset of 68,539 documents from 3ie's DEP. This dataset is highly imbalanced, with the positive class, or class 1, (indicating an included document) in the minority. An imbalanced dataset represents a realistic corpus of literature that experts would filter through for relevance in an EGM. For each document we have the following information: title, year of publication, abstract, and inclusion decision.

For our imbalanced dataset, the positive class has a prior distribution of just 7.7%. Past research has measured data imbalance by observing the ratio $\frac{\sigma}{\mu}$ where σ is the standard deviation and μ is the mean of instances per class in a dataset [4]. By this metric, our dataset's imbalance ratio is $ir = \frac{28{,}988}{34{,}269} = 0.846$. This is more imbalanced than the other highly imbalanced datasets noted in [4], which had imbalance ratios ranging from $0.739 - 0.793$. Other studies have measured data imbalance in terms of the positive class having a prior distribution of 15% or less [47]. Using the prior distribution percentage, we compare our dataset to other common imbalanced NLP datasets in Table 4.1. As this table indicates, our dataset is highly imbalanced, with a lower prior for our positive class than all of the example imbalanced datasets used in [47].

We perform a 70%/15%/15% train/test/validation split of our data. We use only the training data as possible candidates for labelling and training. Therefore, our unlabeled dataset, $\mathcal{U}$, is the training dataset. We have the ground truth labels for all of $\mathcal{U}$, and use these as the oracle which provides the true labels to the queried instances in our AL approaches.

No.	Dataset	Size	Positive Class	Prior
1	TREC	5,952	location	15%
2	ISEAR	7,666	fear	14%
3	Wiki attack	21,000	general	12%
4	AG's news-imb	17,538	world	10%
5	Polarity-imb	5,923	positive	10%
6	Subjectivity-imb	5,556	subjective	10%
7	Ours	68,539	include	7.7%

Table 4.1: Common imbalanced datasets in NLP and their size, positive class, and the prior distribution of the positive class. Our dataset has a notably low prior distribution, indicating that it is very imbalanced.

4.3.2 Implementation Strategy

We compare four different data selection scenarios, which differ in how their initial seed $\mathcal{L}$ is selected, and how additional data from $\mathcal{U}$ is selected to be added to the training set $\mathcal{L}$. The four scenarios are as follows.

1. Balanced Initial + Least Confidence Selection
2. Balanced Initial + Random Selection
3. Random Initial + Least Confidence Selection
4. Random Initial + Random Selection

We chose these four scenarios in order to observe the effects of the initial training set on model performance, and the effects of the additional training data selection strategy on model performance. We compare two initial scenarios: a balanced initial training set and a random, and likely highly imbalanced, initial set. Our initial selection techniques are influenced by [47] which considers different scenarios for creating initial seeds for imbalanced datasets. They found that a randomly selected initial seed led to unstable BERT runs, and consequently ran experiments in which they started with either (1) a truly balanced $\mathcal{L}$, or (2) a hopefully balanced $\mathcal{L}$ derived from using a key-word based query which aims to find instances of the positive minority class. To note, [47] start with a smaller $\mathcal{L}$ than we start with, which likely impacts the instability of a BERT model trained on a random initial seed.

Generating a balanced initial dataset will require labeling more initial training points than generating a random initial dataset. Therefore, we aim to measure the effect of such an initial set to see if that effort is justified. Similarly, we experiment with a random data selection technique and a least confidence sampling technique, which is a form of AL. Our goal with these experiments is to measure the effect of AL on our model performance. Once again, we aim to determine which data selection technique results in the best model performance with the smallest size of required training data.

For each of the scenarios, we begin by selecting an initial seed of labeled data, $\mathcal{L}$, from $\mathcal{U}$ and providing their associated labels. For the two Balanced Initial scenarios, we choose 1,000 instances total: 500 from class 0, the negative class indicating excluded documents, and 500 from class 1, the positive class indicating included documents. For the Random Initial scenario we randomly select 1,000 instances from $\mathcal{U}$ to become our initial seed, $\mathcal{L}$.

With our initial labeled set $\mathcal{L}$, we train the classification model. We start with a pretrained BERT for Sequence Classification model [43], that we fine-tune using $\mathcal{L}$. This method is similar to Aggarwal, Popescu, and Hudelot's work in which they performed AL on imbalanced image datasets. They utilized the pretrained ResNet-18 model [57] as a base model and fine-tuned it to for classification tasks for each of their four imbalanced datasets [4]. We similarly start with a large pretrained model, but within the natural language processing domain, and go on to fine-tune it for our specific classification task.

After fine-tuning our model, we turn to one of our two methods for selecting additional data: Least Confidence Selection or Random Selection. For Least Confidence Selection we evaluate the model on both the test data, to examine model performance, and the unseen training data, $\mathcal{U}$, to choose new training data. We use the least confidence variant of uncertainty sampling shown in Equation 4.1 to select new instances from $\mathcal{U}$ to label. This method selects the instances that the model most believes it could have mislabeled [110], and therefore those ground truth labels should be very informative. We select 500 instances at a time, add those to the labeled training set $\mathcal{L}$, and then retrain the learner. For Random Selection we randomly select 500 instances from the unseen training data and add them to the labeled training set $\mathcal{L}$.

We repeat the selection of new data until we reach a training size of 5,000 labeled instances. To accurately compare the four data selection scenarios described above, we hold the hyperparameters of the model the same for each of the scenarios. We use a learning rate of $2 \cdot 10^{-5}$ and keep the best model based on performance on the validation set.

4.4 Results

We follow the standard of evaluating the performance of a classifier model on an imbalanced dataset with the F1 score. We have both a baseline and a goal performance. The baseline, shown in 4-1, is Random Initial + Random Selection scenario, which shows how the model performs given random training data at different training sizes. The goal performance is the F1 score achieved when the model is trained on all training points in $\mathcal{U}$, which is 70% of the data or 47,977 training points. This goal F1 score is 0.81 and is achieved with a learning rate of $4 \cdot 10^{-5}$.

The results for our four data selection scenarios are shown in 4-1. For these experiments we are concerned with the relative performance of each scenario to the others, since we are keeping standard hyperparameters for all scenarios and not optimizing the hyperparameters of each model to achieve its highest F1 score. We observe that all three non-baseline scenarios outperform the Random Initial + Random Selection baseline. Additionally, we see that for the maximum number of training points, the top scenarios are the two with the Least Confidence Selection technique. These are the two scenarios which employ AL, and they reach an F1 score of 0.78 within a training size of 4,000, less than 10% of the training dataset.

These results align with our intuition that using the AL least confidence sampling strategy to select additional training points will improve model performance. This can inform model selection for AI assisted idea filtering. One unexpected and highly informative result is that the Random Initial + Least Confidence Selection scenario

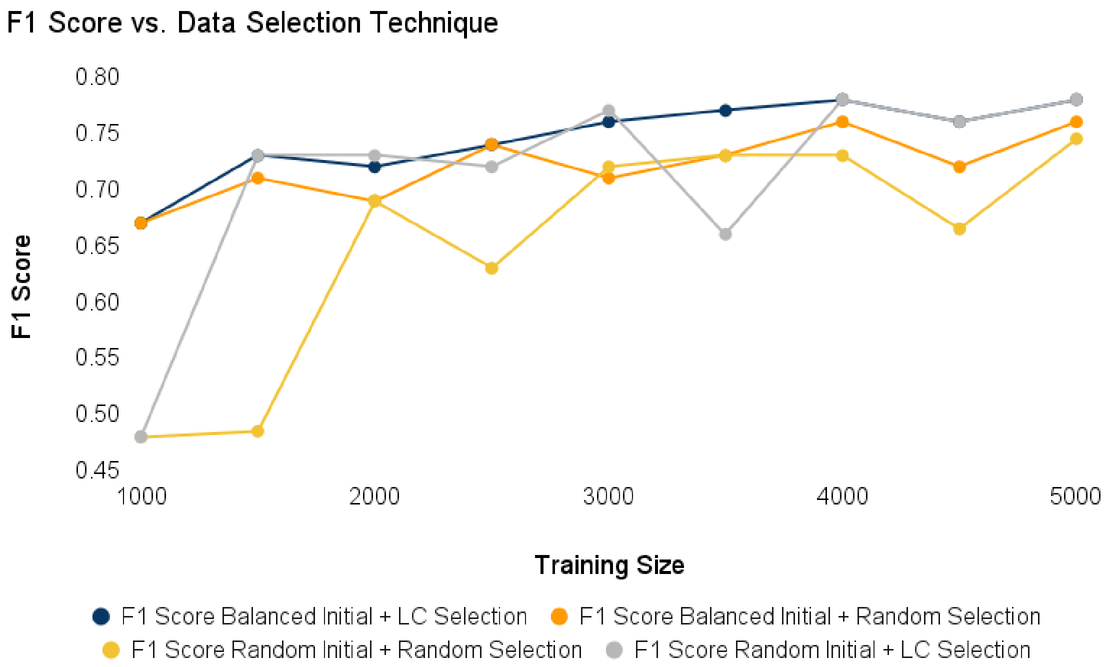

Figure 4-1: The F1 score vs. training data size for the four data selection techniques.

and Balanced Initial + Least Confidence Selection scenario perform equally well, as the two top performing models. Given the positive impact of a balanced training set on classification accuracy [129], we might anticipate that the scenarios with a balanced initial training set would outperform scenarios with a random initial training set. We find, however, that this is not the case for the two scenarios that use AL. This suggests that we do not need a balanced initial training set to reach top model performance. Since generating a balanced labeled training set takes more effort than generating a random labeled training set, this finding can lead to a decreased level of effort required from experts.

Furthermore, while both of the Random Initial scenarios start at a much lower F1 score than the Balanced Initial scenario, the Random Initial + Least Confidence Selection scenario reaches high F1 scores within one data selection iteration. We do observe that the Balanced Initial cases have a smoother positive trend between training size and F1 score, while the two Random Initial scenarios exhibit a less smooth positive trend, with outlying dips such as the Random Initial + Least Confidence

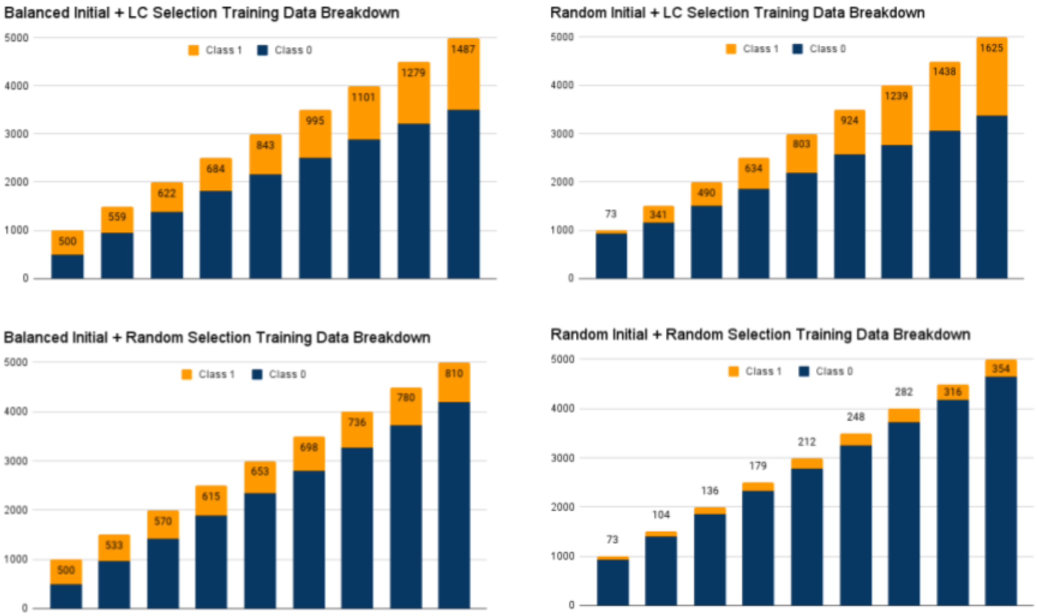

Figure 4-2: The class breakdown of training data from the initial training set of 1,000 instances to the final set of 5,000 instances.

Selection dip at 3,500 training points.

Both cases which employed an active learning strategy selected many more instances from the positive minority class than the two strategies which did not. These results can be seen in Figure 4-2, as well as in larger scale in Figures A-1, A-2, A-3, and A-4 included in the Appendix. Of the 4,000 additional instances selected after the initial seed was chosen, the Balanced Initial + Least Confidence Selection scenario selected 25% of instances from class 1. The Random Initial + Least Confidence Selection scenario started with an initial seed of 73 class 1 instances out of 1,000 total. This scenario selected 1,552 or 39% of additional training instances from class 1. The Random Initial + Random Selection scenario selected the fewest training instances from class 1. Only 7%, or 281, of the 4,000 additional training instances were class 1, resulting in a final training set that was *more* imbalanced than the full dataset.

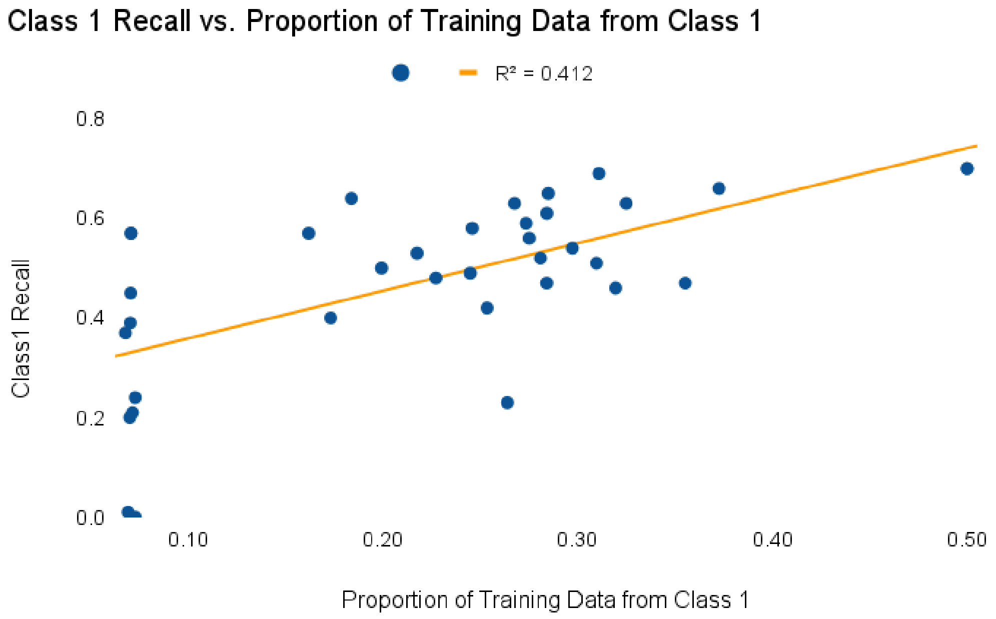

Figure 4-3: Class 1 Recall vs. Percent of Training Data that is Class 1. We observe a positive linear relationship between the percent of the training data that is from the positive minority class, class 1, and the recall of the positive class. The R^2 value is 0.412.

4.5 Discussion

In this section, we return to our overarching goal of reducing expert reviewers' level of effort while still accurately classifying documents for inclusion. We will use this section to discuss how our results tie into this goal.

A key result of this research is that both scenarios that employ an AL strategy have the highest F1 scores. By 4,000 training points, less than 10% of the total training data, both the Balanced Initial + Least Confidence Selection scenario and the Random Initial + Least Confidence Selection scenario achieve F1 scores of 0.78. We would like to emphasize that these results were generated without tuning hyperparameters for an optimal F1 score, since the hyperparameters were standardized to allow for comparison of scenarios. We hypothesize that with hyperparameter tuning, these two scenarios would reach greater F1 scores. Future work could explore the training size, scenario, and hyperparameters required to reach the global best F1 score.

The fact that we obtain quite similar performance with the Random Initial + Least Confidence Selection and Balanced Initial + Least Confidence Selection scenarios suggests to us that the former scenario would be the best in practice. This is because getting an initial balanced dataset required for the Balanced Initial scenario is not a trivial task.

While one may employ a similar keyword search technique as the one used in [47] with the aim of identifying instances from the positive minority class, expert reviewers would likely have to screen many more documents from the negative class before getting a desired number of documents from the positive class. Take, for instance, our balanced 500 negative class instances and 500 positive class instances scenario. We can find a worst case estimate of the number of documents an expert reviewer would have to screen to get 500 positive class instances using the prior distribution of 7.7% positive class. The reviewer would have to screen $500/0.077 = 6,494$ documents to identify 500 positive class documents to be used for the initial seed. Therefore, since we observe similarly high performance from the Random Initial + Least Confidence Selection scenario, this scenario may be the most beneficial in practice.

From Figure 4-2, we can see that the AL strategies result in a more balanced training dataset. The Random Initial + Least Confidence Selection scenario resulted in the most balanced training dataset (32.5% class 1 data), even moreso than the two scenarios which started with balanced initial training sets. Past research has shown that using balanced training data to train binary classification models for real-world applications results in the highest balanced accuracy (the average of True Positive Rate and True Negative Rate), Matthews correlation coefficient, and area under ROC curves [129]. These results hold regardless of the distribution of the test data, and suggest that a balanced training set will produce the best classification results even for imbalanced real-world applications like ours [129].

Figure 4-3 shows the impact that a more balanced training set might have. We observe a positive linear relationship between the proportion of training data that is from class 1, the positive class, and the class 1 recall. These results align with

[129], which suggests that more balanced training data produces the highest model performance.

In practice, an improved class 1 recall means our model would identify a greater number of the included documents from the entire literature corpus, as documents that should be included. Since the minority class is the class we are most concerned about identifying and including in an evidence synthesis product, class 1 recall is a high impact metric in this real-world scenario.

Given the nature of our experiments, we need to disaggregate the impact of total training size with the impact of proportion of class 1 in the training set on class 1 recall. In Figure A-5, included in the Appendix, we observe a very weak relationship between class 1 recall and training size, as indicated by an R^2 value of 0.02. This helps confirm that the positive linear relationship we observe between class 1 recall and the proportion of the training data from class 1 is, in fact, that relationship, and is not confounded by training data size.

4.6 Summary

In this chapter we address a common challenge found in automated idea filtering and in the engineering design field in general: difficulty in obtaining large labeled datasets. In the engineering design field there is a lack of large labeled datasets [93, 103], and generating such datasets using expert labels is expensive and time consuming [93, 39]. However, labeled training sets that allow for automated idea filtering with ML models can provide great benefit and effort-reduction for experts during the design process. Therefore, the research in this chapter explores ways to reduce the effort required to generate expert-labeled training data for NLP-based idea filtering. We find that data selection scenarios that incorporate active learning result in higher F1 scores, a more balanced training set, and fewer necessary labeled training instances. A model trained with 4,000 datapoints (less than 10% of the training dataset) selected via an active learning strategy reaches an F1 score of 0.78, compared to the baseline model with has an F1 score of 0.73 here. Furthermore, our research demonstrates that if a scenario incorporates active learning in its selection strategy, it reaches high performance regardless of its initial training set. In practice, this finding is useful in identifying the Random Initial + Least Confidence Selection scenario as that which produces the best model performance while requiring the lowest level of effort from expert reviewers, since it does not require a balanced initial training set. Lastly, the two data selection techniques that incorporate active learning result in training sets with 30% and 33% of minority class instances, while the two techniques without AL result in 7% and 16% minority class training sets. More balanced training sets are associated with higher classification accuracy [129], and our results show that class 1 recall increases with more balanced training sets. Overall, this research suggest that active learning is effective in decreasing expert level of effort for NLP-based idea filtering with highly imbalanced data. Future work should explore various active learning strategies, especially ones which specifically address the high level of class imbalance in our real-world dataset. Additionally, future work might aim to quantify the effort of labeling instances in batches, as required for AL, versus labeling instances all at once. This will help create a full picture of the effort-saving benefit of AL in real-world automated idea filtering applications.

Chapter 5

Conclusion

This treatise examines three research scenarios that employ NLP-based idea filtering to lower the level of effort required of human experts and therefore accelerate the design process. There are several common challenges surrounding idea filtering. Firstly, manual idea filtering is often a bottleneck in the design process, as evaluating a plethora of ideas is both time and resource intensive for experts [25, 127, 89]. We aim to address this by automating idea filtering, which brings us to our second large challenge. In engineering and design, there is a lack of large labeled datasets [93, 103]. Furthermore, generating a labeled dataset for a specific application entails significant effort from expert reviewers, which increases with the size of required training data. We address these challenges through three research scenarios in order to understand the possibility and benefits of using NLP-based idea filtering to accelerate the design process.

In Chapter 2, we explore the efficacy of using machine learning to predict expert ratings of design characteristics like creativity, usefulness, and elegance. In particular, we explore whether we can predict expert creativity ratings from non-expert design survey results. Our results demonstrate that incorporating NLP in the prediction process improve the model's predictive power across design metrics, including CAT ratings. This study also found that the predictability of different design metrics varies, with Usefulness and Elegance having the best predictability, likely because these are the least subjective design metrics. These findings also serve as empirical evidence supporting the investigation of novice vs expert usage in creativity assessment. As the surveys can be completed by novices, utilizing the surveys more can decrease the level of effort required of experts. The preliminary success in using NLP to predict design metrics helps show the wide application of NLP, and also supports its usage in modeling subjective ratings like creativity. Being able to model and evaluate design metrics provides basis for automatically filtering design ideas based in these metrics. The potential of utilizing a small number of expert assessments to train a model to then predict future design idea assessments can decrease the level of effort of expert reviewers. Therefore, this work addresses two common challenges in idea filtering, that manual filtering by experts is expensive and time consuming, and that large labeled datasets are not frequently available and are expensive to generate.

In Chapter 3, we move to a different domain: accelerating evidence synthesis for the design of international development projects using NLP-based idea filtering. Our research explores the effectiveness of three NLP classification models on determining if a document should be included in an international development evidence synthesis. In this chapter we work with a much larger dataset of over 50,000 documents, which would require significan effort for human experts to filter through. We find that an iBERT model comprised of a fine-tuned BERT model with additional extracted information performs the best with an accuracy of 79%. We further explore the influence of the amount of training data on the accuracy. As our intuition suggests, accuracy improves with the amount of training data used. However, the rate of improvement slows down as the amount of training data increases, and after about 10,000 training datapoints the improvements are small. Our results can inform decisions between increased accuracy and the level of effort to manually create training data in real-world applications of this work. Our work is motivated by the potential of NLP to automate time-intensive aspects of evidence synthesis models like filtering through tens of thousands of text documents. By automatically and accurately filtering documents for inclusion in an evidence gap map, we can quickly provide decision-makers with the tools to design evidence-based policies.

Our work in Chapter 4 builds from Chapter 3, where we demonstrate the efficacy of using NLP to classify literature for inclusion in an evidence synthesis product. In Chapter 4 we address the idea filtering challenge that generating a labeled training dataset entails significant effort from expert reviewers, which increases with the size of required training data. In Chapter 3, our work relies on a labeled training set of documents that we use to fine tune a BERT-based model. As mentioned, generating this training set requires significant expert resources and time, especially in the realistic setting where the positive class is in the minority. Motivated by this, we aimed to decrease the level of effort required of expert reviewers to generate training data. We turned to the field of active learning (AL) to decrease the necessary training size while retaining or improving model performance. We compared classifier models with four different data selection scenarios using our highly imbalanced real-world dataset. We found that data selection techniques that incorporate active learning result in higher F1 scores, a more balanced training set, and fewer necessary labeled training instances. These results suggest that active learning is effective in decreasing expert level of effort for NLP-based idea filtering with highly imbalanced data.

To conclude, throughout these three projects we address the common idea filtering challenges we identified above. In Chapters 2, 3, and 4, we address the overarching bottleneck of manual idea filtering by presenting automated NLP-based models. In Chapters 2 and 4, we address the challenge of obtaining large expertly labeled datasets in the engineering design domain because of their scarcity and expense by maximizing learning from a smaller number of expertly labeled datapoints. Lastly, in Chapter 4, we address the challenge of the effort required to generate a sufficient training dataset by utilizing AL. Through working on these three research projects, we find that NLP can accelerate design processes in various domains by automating idea filtering and decreasing the level of effort required of human experts.

Appendix A

Figures

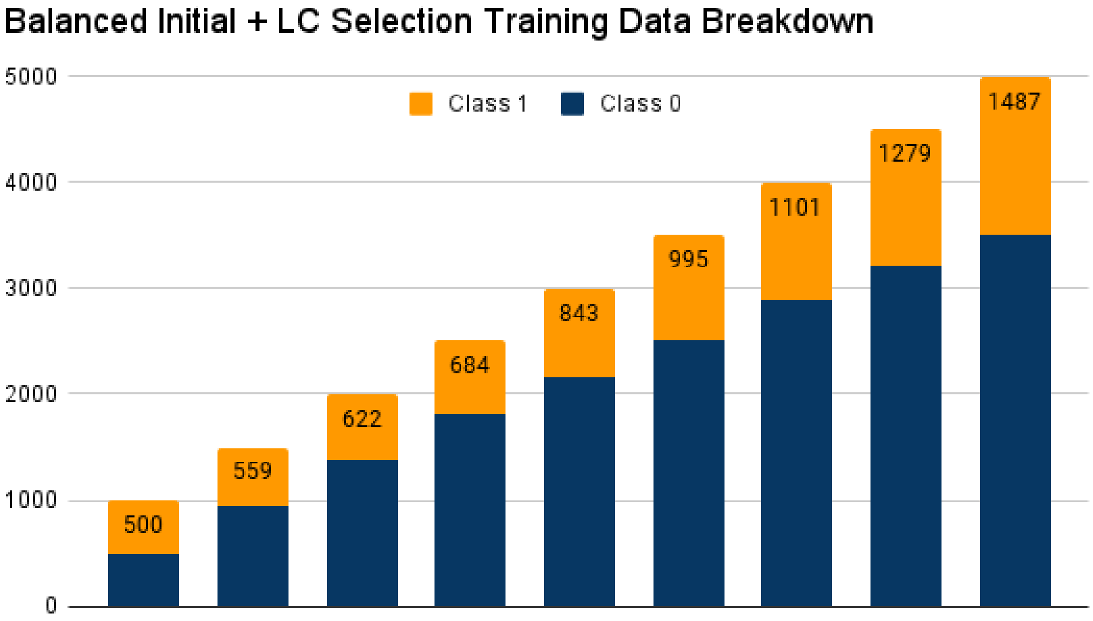

Figure A-1: Training data breakdown per iteration for the Balanced Initial + Least Confidence Selection Technique.

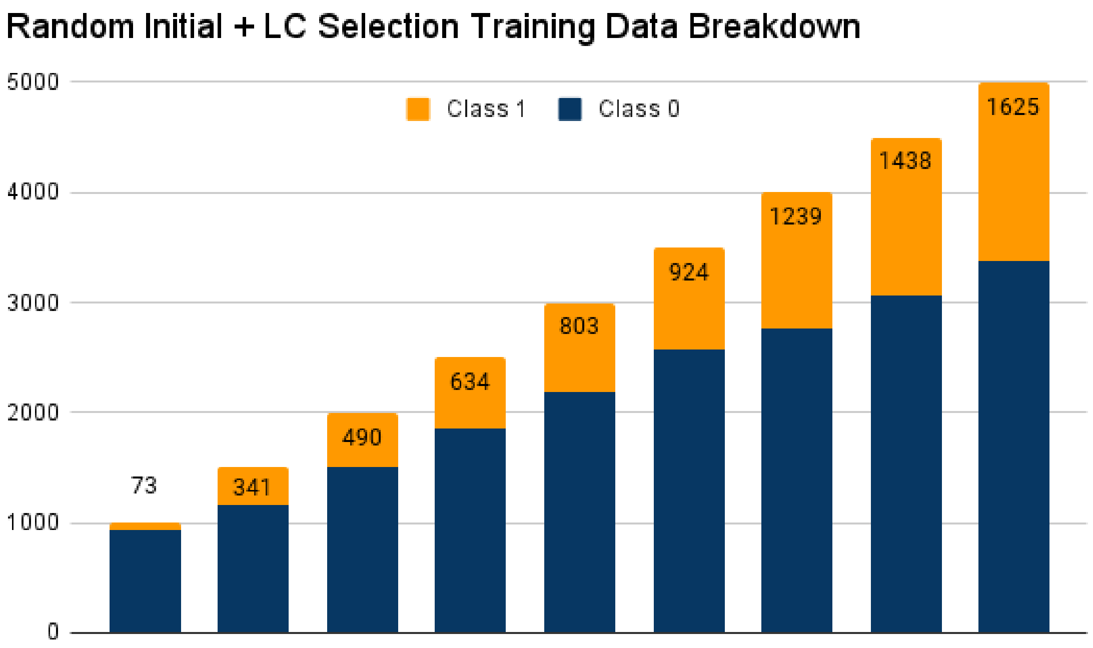

Figure A-2: Training data breakdown per iteration for the Random Initial + Least Confidence Selection Technique.

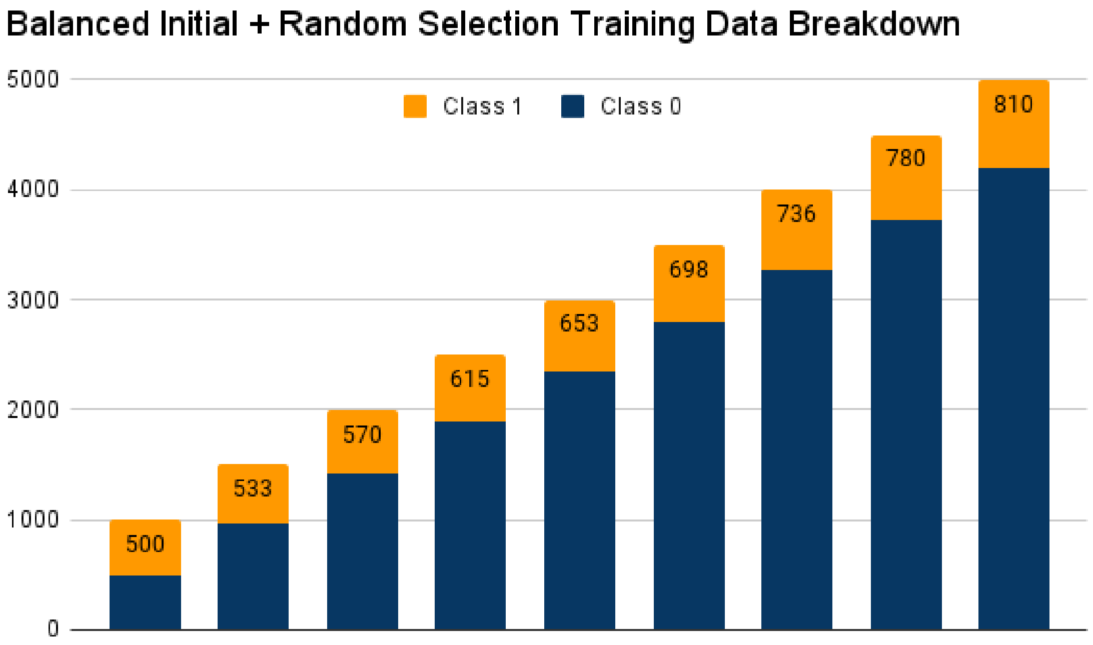

Figure A-3: Training data breakdown per iteration for the Balanced Initial + Random Selection Technique.

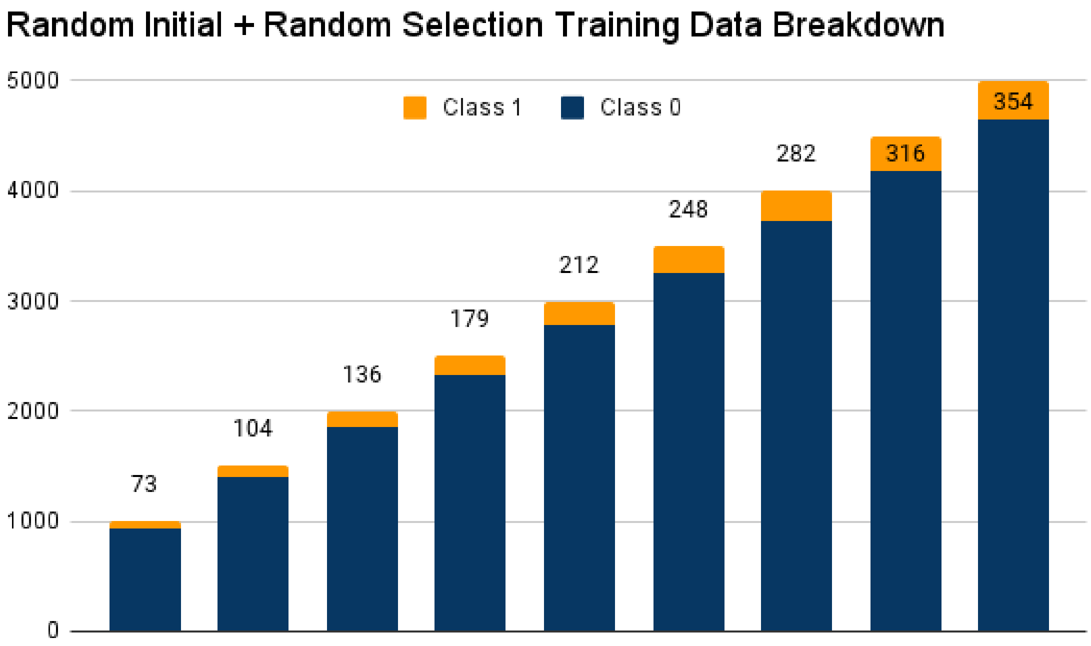

Figure A-4: Training data breakdown per iteration for the Random Initial + Random Selection Technique.

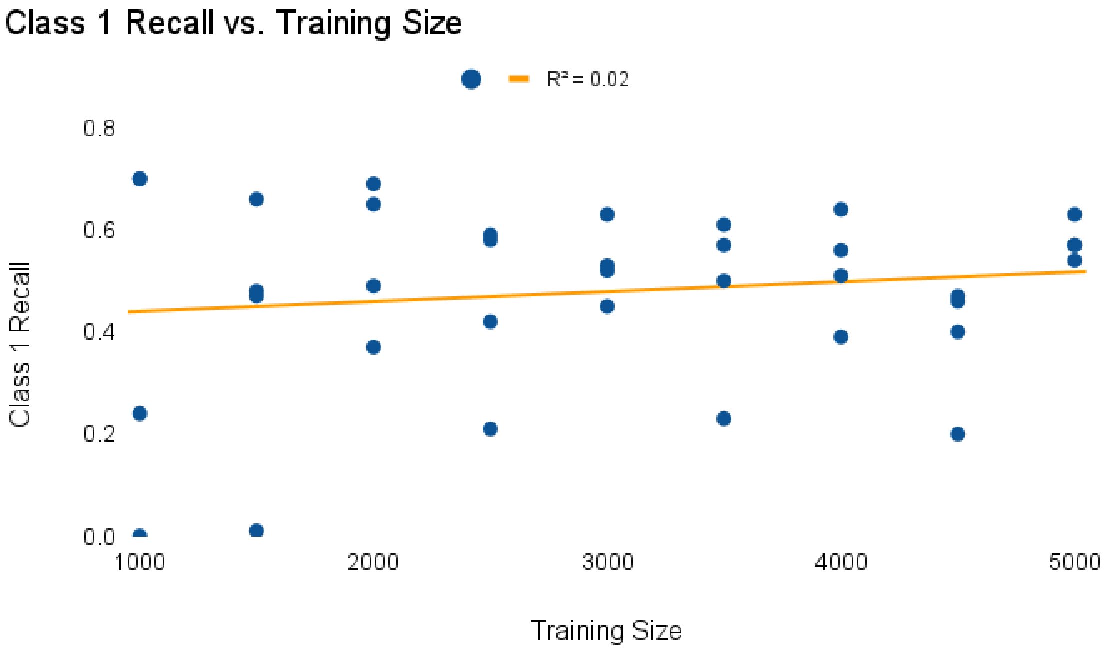

Figure A-5: Class 1 recall vs. Total Training Size. We observe a very weak relationship between class 1 recall and training size, as indicated by an R^2 value of 0.02. This helps confirm that the positive linear relationship we observe between class 1 recall and the proportion of the training data from class 1 is in fact that relationship, and is not confounded by training data size.

Bibliography

[1] 3ie. Development evidence portal. Available at https://developmentevidence.3ieimpact.org (2021/08/04).

[2] 3ie. Evidence gap maps. Available at https://www.3ieimpact.org/evidence-hub/evidence-gap-maps (2021/08/04).

[3] 3ie. Evidence mapping. Available at https://www.3ieimpact.org/evidence-hub/evidence-gap-maps (2021/08/04).

[4] Umang Aggarwal, Adrian Popescu, and Celine Hudelot. Minority class oriented active learning for imbalanced datasets. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, jan 2021.

[5] Faez Ahmed and Mark Fuge. Capturing winning ideas in online design communities. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1675–1687, 2017.

[6] Faez Ahmed, Sharath Kumar Ramachandran, Mark Fuge, Sam Hunter, and Scarlett Miller. Measuring and optimizing design variety using herfindahl index. In *ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection, 2019.

[7] Faez Ahmed, Sharath Kumar Ramachandran, Mark Fuge, Samuel Hunter, and Scarlett Miller. Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel. *Journal of Mechanical Design*, 141(2), 2019.

[8] Ahmed Aldahdooh, Enrico Masala, Glenn Van Wallendael, Peter Lambert, and Marcus Barkowsky. Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in vqa datasets using objective measures. *Signal Processing: Image Communication*, 74:32–41, 2019.

[9] Leyla Alipour, Mohsen Faizi, Asghar M Moradi, and Gholamreza Akrami. The impact of designers' goals on design-by-analogy. *Design Studies*, 51:1–24, 2017.

[10] Teresa M Amabile. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5):997, 1982.

[11] Teresa M Amabile. Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology*, 19(2):146–156, 1983.

[12] Teresa M Amabile. A model of creativity and innovation in organizations. *Research in organizational behavior*, 10(1):123–167, 1988.

[13] Teresa M Amabile. *Creativity in context: Update to the social psychology of creativity*. Routledge, 2018.

[14] T. M. Ambile. *Creativity in Context*. Westview Press, Boulder, Colorado, 1996.

[15] Dana Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 04 1998.

[16] Dana Angluin. Queries revisited. In Naoki Abe, Roni Khardon, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 12–31, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[17] Olufunmilola Atilola, Megan Tomko, and Julie S Linsey. The effects of representation on idea generation and design fixation: A study comparing sketches and function trees. *Design Studies*, 42:110–136, 2016.

[18] Les Atlas, David Cohn, and Richard Ladner. Training connectionist networks with queries and selective sampling. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[19] Michael J Bommarito II au2, Daniel Martin Katz, and Eric M Detterman. Lexnlp: Natural language processing and information extraction for legal and regulatory texts, 2018.

[20] John Baer. The importance of domain-specific expertise in creativity. *Roeper Review*, 37(3):165–178, 2015.

[21] John Baer and James C Kaufman. Assessing creativity with the consensual assessment technique. In *The Palgrave Handbook of Social Creativity Research*, pages 27–37. Springer, 2019.

[22] John Baer, James C Kaufman, and Claudia A Gentile. Extension of the consensual assessment technique to nonparallel creative products. *Creativity research journal*, 16(1):113–117, 2004.

[23] Philipp Barth and Georg Stadtmann. Creativity assessment over time: Examining the reliability of cat ratings. *The Journal of Creative Behavior*, 2020.

[24] Mark Batey and Adrian Furnham. Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, social, and general psychology monographs*, 132(4):355–429, 2006.

[25] Andy Berndt. Project 10ˆ100, Sep 2008.

[26] Susan P Besemer. Creative product analysis matrix: testing the model structure and a comparison among products–three novel chairs. *Creativity Research Journal*, 11(4):333–346, 1998.

[27] Susan P Besemer and Karen O'Quin. Confirming the three-factor creative product analysis matrix model in an american sample. *Creativity Research Journal*, 12(4):287–296, 1999.

[28] Taweh Beysolow II. *What Is Natural Language Processing?*, pages 1–12. Apress, Berkeley, CA, 2018.

[29] Rob B. Briner and David Denyer. Systematic review and evidence synthesis as a practice and scholarship tool. In *Handbook of evidence-based management: Companies, classrooms and research*, pages 112–129. New York University Press, 2012.

[30] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[31] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[32] Joel Chan, Steven P Dow, and Christian D Schunn. Do the best design ideas (really) come from conceptually distant sources of inspiration? In *Engineering a Better Future*, pages 111–139. Springer, Cham, 2018.

[33] Peiyao Cheng, Ruth Mugge, and Jan PL Schoormans. A new strategy to reduce design fixation: Presenting partial photographs to designers. *Design Studies*, 35(4):374–391, 2014.

[34] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[35] Bo T Christensen and Linden J Ball. Dimensions of creative evaluation: Distinct design and reasoning strategies for aesthetic, functional and originality judgments. *Design Studies*, 45:116–136, 2016.

[36] Mu-Huan Chung, Mark Chignell, Lu Wang, Alexandra Jovicic, and Abhay Raman. Interactive machine learning for data exfiltration detection: Active learning with human expertise. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 280–287, 10 2020.

[37] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 05 1994.

[38] David Cohn, Zoubin Ghahramani, and Michael Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[39] Genevieve M Cseh and Karl K Jeffries. A scattered cat: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):159, 2019.

[40] Aron Culotta and McCallum Andrew. Reducing labeling effort for structured prediction tasks. In *AAAI*, page 746–751, 2005.

[41] Nasrin Dehbozorgi, Mary Lou Maher, and Mohsen Dorodchi. Sentiment analysis on conversations in collaborative active learning as an early predictor of performance. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2020.

[42] Xinyang Deng, Qi Liu, Yong Deng, and Sankaran Mahadevan. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, 340:250–261, 2016.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[44] Christl A. Donnelly, Ian Boyd, Philip Campbell, Claire Craig, Patrick Vallance, Mark Walport, Christopher J. M. Whitty, Emma Woods, and Chris Wormald. Four principles to make evidence synthesis more useful for policy. *Nature*, 558:361–364, 2018.

[45] Kristen M. Edwards, Aoran Peng, Scarlett R. Miller, and Faez Ahmed. If a Picture is Worth 1000 Words, Is a Word Worth 1000 Features for Design Metric Estimation? *Journal of Mechanical Design*, 144(4), 12 2021. 041402.

[46] Kristen M. Edwards, Aoran Peng, Scarlett R. Miller, and Faez Ahmed. If a picture is worth 1000 words, is a word worth 1000 features for design metric estimation? In *Proceedings of the IDETC-CIE 2021 Conference*, volume Volume 6: 33rd International Conference on Design Theory and Methodology (DTM) of *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 08 2021.

[47] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for bert: An empirical study. In *EMNLP*, 2020.

[48] Eric Francis Eshun and KG de Graft-Johnson. Learner perceptions of assessment of creative products in communication design. *Art, Design & Communication in Higher Education*, 10(1):89–102, 2012.

[49] National Science Foundation. Publications output: U.s. trends and international comparisons. `https://ncses.nsf.gov/pubs/nsb20206/`. Accessed: 14-January-2022.

[50] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data, 2017.

[51] Francesco Galati. Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings in assessing creativity. *Creativity Research Journal*, 27(1):24–30, 2015.

[52] Anindya Ghose and Panagiotis G Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, pages 303–310, 2007.

[53] Anindya Ghose and Panagiotis G Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*, 23(10):1498–1512, 2010.

[54] Christopher A Gosnell and Scarlett R Miller. But is it creative? delineating the impact of expertise and concept ratings on creative concept selection. *Journal of Mechanical Design*, 138(2), 2016.

[55] Joshua T Gyory, Kenneth Kotovsky, and Jonathan Cagan. A topic modeling approach to study the impact of manager interventions on design team cognition. In *ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection, 2020.

[56] John T. Hancick and Taghi M. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7, 2020.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

[59] Brian E. Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R. Shah, Stephanie Holmgren, Katherine E. Pelch, Vickie Walker, Andrew A. Rooney, Malcolm Macleod, Ruchir R. Shah, and Kristina Thayer. Swift-review: a text-mining workbench for systematic review. *Systematic Reviews*, 5, 2016.

[60] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230, 2008.

[61] Baer John and S. McKool Sharon. *Assessing Creativity Using the Consensual Assessment Technique*, pages 65–77. IGI Global, Hershey, PA, USA, 2009.

[62] Tyler A Johnson, Avery Cheeley, Benjamin W Caldwell, and Matthew G Green. Comparison and extension of novelty metrics for problem-solving tasks. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 50190, page V007T06A012. American Society of Mechanical Engineers, 2016.

[63] Junegak Joung and Harrison M Kim. Importance-performance analysis of product attributes using explainable deep neural network from online reviews. In *ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection, 2020.

[64] James C Kaufman, John Baer, and Jason C Cole. Expertise, domains, and the consensual assessment technique. *The Journal of creative behavior*, 43(4):223–233, 2011.

[65] James C Kaufman, John Baer, Jason C Cole, and Janel D Sexton. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20(2):171–178, 2008.

[66] James C Kaufman, John Baer, David H Cropley, Roni Reiter-Palmon, and Sarah Sinnett. Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, 7(4):332, 2013.

[67] James C Kaufman, Claudia A Gentile, and John Baer. Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly*, 49(3):260–265, 2005.

[68] James C Kaufman, Jonathan A Plucker, and John Baer. *Essentials of creativity assessment*, volume 53. John Wiley & Sons, 2008.

[69] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 3–12, Berlin, Heidelberg, 1994. Springer-Verlag.

[70] Runhua Li, Yonghua Zhu, and Zhiguo Wu. A new algorithm to the automated assessment of the chinese subjective answer. In *2013 International Conference on Information Technology and Applications*, pages 228–231. IEEE, 2013.

[71] Elizabeth D Liddy. Natural language processing. *Encyclopedia of Library and Information Science*, 2001.

[72] Lassi A Liikkanen, Matti M Hämäläinen, Anders Häggman, Tua Björklund, and Mikko P Koskinen. Quantitative evaluation of the effectiveness of idea generation in the wild. In *International Conference on Human Centered Design*, pages 120–129. Springer, 2011.

[73] Julie S Linsey, Matthew G Green, Jeremy T Murphy, Kristin L Wood, and Art B Markman. "collaborating to success": An experimental study of group idea generation techniques. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 4742, pages 277–290, 2005.

[74] Julie Stahmer Linsey. *Design-by-analogy and representation in innovative engineering concept generation*. PhD thesis, University of Texas at Austin, 2007.

[75] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Modeling and predicting the helpfulness of online reviews. In *2008 Eighth IEEE international conference on data mining*, pages 443–452. IEEE, 2008.

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

[77] Haiying Long and Weiguo Pang. Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, 15:13–25, 2015.

[78] Panagiotis Louridas. Design as bricolage: anthropology meets design thinking. *Design Studies*, 20(6):517–535, 1999.

[79] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700, 2010.

[80] Alex Maritz and Jerome Donovan. Entrepreneurship and innovation: Setting an agenda for greater discipline contextualisation. *Education & training (London)*, 57(1):74–87, 2015.

[81] Ian J. Marshall and Byron C. Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8, 2019.

[82] Stéphane Meystre and Peter J Haug. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of biomedical informatics*, 39(6):589–599, 2006.

[83] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[84] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[85] Scarlett R Miller, Samuel T Hunter, Elizabeth Starkey, Sharath Ramachandran, Faez Ahmed, and Mark Fuge. How should we measure creativity in engineering design? a comparison between social science and engineering approaches. *Journal of Mechanical Design*, 143(3), 2021.

[86] Karim Moustafa, Saturnino Luz, and Luca Longo. Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In *International symposium on human mental workload: Models and applications*, pages 30–50. Springer, 2017.

[87] Michael D Mumford and Sigrid B Gustafson. Creativity syndrome: Integration, application, and innovation. *Psychological bulletin*, 103(1):27, 1988.

[88] Brent A Nelson, Jamal O Wilson, David Rosen, and Jeannette Yen. Refined metrics for measuring ideation effectiveness. *Design Studies*, 30(6):737–743, 2009.

[89] Kathryn Oliver, Simon Innvar, Theo Lorenc, Jenny Woodman, and James Thomas. A systematic review of barriers to and facilitators of the use of evidence by policymakers. *BMC Health Services Research*, 14:2, 2014. 00000.

[90] Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. Technical report, Swedish Institute of Computer Science, 2009.

[91] Sarah K Oman, Irem Y Tumer, Kris Wood, and Carolyn Seepersad. A comparison of creativity and innovation metrics and sample validation through in-class design projects. *Research in Engineering Design*, 24(1):65–92, 2013.

[92] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):1, 2015. 00000.

[93] Soya Park, April Wang, Ban Kawas, Q. Vera Liao, David Piorkowski, and Marina Danilevsky. Facilitating knowledge sharing from domain experts to data scientists for building nlp models, 2021.

[94] Jef Peeters, Paul-Armand Verhaegen, Dennis Vandevenne, and JR Duflou. Refined metrics for measuring novelty in ideation. *IDMME Virtual Concept Research in Interaction Design, Oct*, pages 20–22, 2010.

[95] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[96] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[97] Jaron Porciello, Maryia Ivanina, Maidul Islam, Stefan Einarson, and Haym Hirsh. Accelerating evidence-informed decision-making for the sustainable development goals using machine learning. *Nature Machine Intelligence*, 2:559–565, 2020.

[98] Kedar Potdar, Taher Pardawala, and Chinmay Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175:7–9, 10 2017.

[99] Natural Language Processing and Information Extraction Program at the University of Minnesota Institute for Health Informatics. Biomedicus. `https://nlpie.github.io/biomedicus/`, 2019.

[100] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[101] Sharath Kumar Ramachandran. *Investigating the Accuracy of Creativity Metrics Used in Engineering Design.* PhD thesis, Pennsylvania State University, 2019.

[102] Matthew R Redmond, Michael D Mumford, and Richard Teach. Putting creativity to work: Effects of leader behavior on subordinate creativity. *Organizational behavior and human decision processes*, 55(1):120–151, 1993.

[103] Lyle Regenwetter, Brent Curry, and Faez Ahmed. Biked: A dataset for computational bicycle design with machine learning benchmarks. *Journal of Mechanical Design*, 144(3), 2022.

[104] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning.* Springer Science & Business Media, 2011.

[105] Prabir Sarkar and Amaresh Chakrabarti. Assessing design creativity. *Design studies*, 32(4):348–383, 2011.

[106] Prabir Sarkar and Amaresh Chakrabarti. Ideas generated in conceptual design and their effects on creativity. *Research in Engineering Design*, 25(3):185–201, 2014.

[107] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *IDA*, 2001.

[108] Scikit-learn. *Ensemble Methods.* `https://scikit-learn.org/stable/modules/ensemble.html`.

[109] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[110] Burr Settles, Mark W. Craven, and Lewis A. Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

[111] Jami J Shah, Santosh V Kulkarni, and Noe Vargas-Hernandez. Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments. *J. Mech. Des.*, 122(4):377–384, 2000.

[112] Jami J Shah, Steve M Smith, and Noe Vargas-Hernandez. Metrics for measuring ideation effectiveness. *Design studies*, 24(2):111–134, 2003.

[113] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948.

[114] Sonit Singh. Natural language processing for information extraction, 2018.

[115] Wouter Sluis-Thiescheffer, Tilde Bekker, Berry Eggen, Arnold Vermeeren, and Huib De Ridder. Measuring and comparing novelty for design solutions generated by young children through different design methods. *Design Studies*, 43:48–73, 2016.

[116] Birte Snilstveit, Jennifer Stevenson, Ian Shemilt, Mike Clarke, Emmanuel Jimenez, and James Thomas. Timely, Efficient,and Living Systematic Reviews: Opportunities in International Development. Inception paper, Centre of Excellence for Development Impact and Learning, London, March 2018.

[117] Birte Snilstveit, Martina Vojtkova, Ami Bhavsar, Jennifer Stevenson, and Marie Gaarder. Evidence & gap maps: A tool for promoting evidence informed policy and strategic research agendas. *Journal of Clinical Epidemiology*, 79:120–129, 2016.

[118] Jamie Snyder. Visual representation of information as communicative practice. *Journal of the Association for Information Science and Technology*, 65(11):2233–2247, 2014.

[119] Binyang. Song, Scarlett. Miller, and Faez Ahmed. Hey, AI! Can You See What I See? Multimodal Transfer Learning-Based Deisgn Metrics Prediction for Sketches with Text Descriptions. In *Proceedings of the ASME IDETC/CIE Conference*, 2022.

[120] C. O. S. Sorzano, J. Vargas, and A. Pascual Montano. A survey of dimensionality reduction techniques, 2014.

[121] Elizabeth M Starkey, Samuel T Hunter, and Scarlett R Miller. Are creativity and self-efficacy at odds? an exploration in variations of product dissection in engineering education. *Journal of Mechanical Design*, 141(1), 2019.

[122] Robert J Sternberg. *Handbook of creativity.* Cambridge University Press, 1999.

[123] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017.

[124] Per Sundström and Annika Zika-Viktorsson. Innovation through explorative thinking in product development projects. In *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm*, 2003.

[125] Christine A Toh and Scarlett R Miller. Creativity in design teams: the influence of personality traits and risk attitudes on creative concept selection. *Research in Engineering Design*, 27(1):73–89, 2016.

[126] Gokhan Tur, Dilek Hakkani-Tur, and Robert E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45:171–186, 2005.

[127] Lorraine Twohill. $10 million for project 10ˆ100 winners, Sep 2010.

[128] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[129] Qiong Wei and Roland L. Dunbrack Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, 8(7), 2013.

[130] Adam B. Wilcox and George Hripcsak. The role of domain knowledge in automating medical text report classification. *Journal of the American Medical Informatics Association*, 10(4):330–338, 2003.

[131] Daya Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *J. Information Science*, 36:306–323, 05 2010.

[132] Yun Xu and Royston Goodacre. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2:249–262, 10 2018.

[133] Maria C Yang. Observations on concept generation and sketching in engineering design. *Research in Engineering Design*, 20(1):1–11, 2009.

[134] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[135] Xiaojin Zhu. *Semi-Supervised Learning with Graphs.* PhD thesis, Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA, 2005.