8-12-2021 10:00 AM

Exploratory Search with Archetype-based Language Models

Brent D. Davis, *The University of Western Ontario*

Supervisor: Lizotte, Daniel J, *The University of Western Ontario*
Co-Supervisor: Sedig, Kamran, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

Abstract

This dissertation explores how machine learning, natural language processing and information retrieval may assist the exploratory search task. Exploratory search is a search where the ideal outcome of the search is unknown, and thus the ideal language to use in a retrieval query to match it is unavailable. Three algorithms represent the contribution of this work. Archetype-based Modeling and Search provides a way to use previously identified archetypal documents relevant to an archetype to form a notion of similarity and find related documents that match the defined archetype. This is beneficial for exploratory search as it can generalize beyond standard keyword matching. By training word embeddings to generate vector representations of all words in the archetypal document vocabulary, and then training an author representation which is a conglomeration of these word representations, a similarity metric can be constructed to use for searching. Unclassified author representations from new corpuses can then be directly classified by machine learning algorithms, compared, and ranked, allowing this technique to be search document collections. Archetype-based Information Retrieval provides a way to extract the keywords most associated with archetypal author representations. This allows integration with keyword-based information retrieval systems that use a probabilistic relevancy score to retrieve more pertinent results. Lastly, Archetype-based Temporal Language Adaptive Stratification makes use of the scoring of previous algorithms and adapts for transitions over time between archetypal states, such as depressive episodes. This algorithm is specialized to find these temporal transitions between archetypes (i.e., depressed and not depressed) and identify language associated with the transition. In concert with Public Health Ottawa, these techniques have been used to 1) identify the language online that is related to the opioid epidemic and the individuals suffering from addiction, and 2) estimate the number of individuals matching this archetype within the catchment area for Public Health Ottawa. These techniques have also been used to identify language associated with depression on social media, and a synthetic example of how to use this to look for transitions between depressive states is described in a partially synthetic case study.

Keywords

Machine Learning; ML; Information Retrieval; Unknown Vocabulary Problem; Exploratory Search; Natural Language Processing; Opioid Epidemic; Public Health Surveillance; Language Model; Author Representation; Document Representation; Representation Learning; Analytics; Data Analytics; Algorithms

Summary for Lay Audience

Advances in machine learning and natural language processing have changed the way that we can search data. When performing a search, such as an online web search using sites such as Google, we are required to input keywords to retrieve content that is relevant to us. This is an example of a look-up style search, where the relevant language is known and there is some idea of what the results should look like. On the other hand, there is exploratory search, which has less definite results and in most cases the relevant vocabulary is partially known at best.

One of the techniques in this dissertation, Archetype-based Modeling and Search, provides a way for a machine learning model to learn the relevant vocabulary from documents that were previously identified as being relevant. By using machine learning approximated notions of similarity, the system is able to find complex associations with words and an approximation of concepts to the relevant documents to find. However, this process is computationally demanding, and can be a bit of a 'black box' when trying to understand the decisions made by the machine learning system. The next technique, Archetype-based Information Retrieval, builds upon the first by extracting the keywords which best explain the decisions being made. We then show how these keywords can be used in a normal information retrieval system, which means that the task of forming a query has been changed from thinking of relevant words to identifying sets of documents which contain keywords thought to be relevant. The last technique, Archetype-based Temporal Language Adaptative Stratification, is a way to expand the previous two techniques to be better at identifying behaviours that change over time, and then analyzes the transition to see what language is associated with that change.

The first two techniques were demonstrated in coordination with Public Health Ottawa to examine the state of the opioid epidemic in their local catchment area as it was represented on social media. We estimate the number of individuals active on Reddit that match this profile and use this to estimate the population prevalence. The last technique was developed while working on analyzing depression on social media during the COVID-19 pandemic, and uses partially synthetic data to avoid the ethical complexities of analyzing and reporting on an individual's experience with depression.

Co-Authorship Statement

Chapters 1 is my original work in explaining the motivation, identifying the problem, framing the dissertation, and providing an overview between the different section. Chapter 2, which provides necessary background to understand the technical foundations of the algorithms described in later chapters, is my own original work explaining previously established background.

Chapter 3 has been published in MDPI's journal Big Data and Cognitive Computing, where my supervisors Drs. Lizotte & Sedig are co-authors. Co-authors were responsible for helping frame the problem, validating the experimental approach, and separating out the case study from the general algorithm.

Chapter 4 is developed in collaboration with Cameron McDermaid at Public Health. Drs. Lizotte & Sedig provided supervisory advice, editing, and direction for the work. I was responsible for all research and writing.

Chapter 5 represents my own research and writing, and my supervisors provided advice and direction.

Chapter 6 is my own writing summarizing and concluding the dissertation.

Acknowledgments

“If I have seen further, it is because I have stood on the shoulders of Giants”
- Isaac Newton

I dedicate this dissertation:

To Drs. Daniel Lizotte and Kamran Sedig, without whom this dissertation would not have been possible. Thank you both for helping me develop myself into who I have become today.

To the Computer Science Department at Western University, for fostering an environment of innovation. A special thank you to all of the administrative staff, retired and new, for all of your help.

To the Insight and Phi Labs, for showing me the incredible diversity of work being done in the department and at Western.

To Jason Baumgartner for making Pushshift.io, the source of the Reddit social media data used in this work; I have no doubt that your contributions to open data and social media research will go on to spawn a thousand such dissertations as this.

To Tanner Bohn, for exasperatedly asking me why I just would not use a classifier on the neural representations I was working with.

To my parents, Harold and Marie, for encouraging this path above all the others that seemed possible at the end of my undergraduate degree.

To Nicole, my beloved wife, for enduring all that comes with a degree of this nature. For the companionship, the love, the compassion, and the motivation to always be better.

Epigraph

“You are always searching for answers to your questions.

That is because you believe they mean something to you.

As long as you keep desiring answers, your life will remain a meaningful one.

You are constantly renewing yourself by thinking and feeling things.”

***From Legend of Mana, a Square Enix Game*.**

Table of Contents

List of Tables (where applicable)

List of Figures (where applicable)

1 Introduction

Searching for any particular thing comes with conditions and ontological baggage. These include: a thing to be searched for; a way of defining both the thing to be found and its relation to all other things that exist in the search space; and a way to navigate the collection of things until the thing is found or the options are exhausted. When we want to find something, we internally express a question such as “Where did I leave that sandwich?”. This mental question is answered by looking around until either the sandwich is found, or worse yet, the empty plate it used to reside upon is. While the root task is relatively simple to describe, “Look until you find the thing”, the task can quickly become more complex and abstract. Often, we want to find something from a specific range of time, or with a specific set of qualities that set it apart from other things.

Sometimes we are unable to formulate exactly what it is that we are looking for. An anecdotal example encountered on Twitter that identifies this phenomenon is in **Figure 1**. This is the phenomenon of finding a word that is said to be at the ‘tip of the tongue’ [1]. Here, Goodwin & Goodwin describe the interactive process that two individuals can go through where they do not know the words they are looking for. They describe how this is a shared cognitive task and go over how prompts from the first individual to the second can lead to new proposed words, even if the solution only exists intangibly in the first individual’s head.

Given that this formulation problem is hard, it is natural to consider using tools to make the formulation easier. This is the central idea around using artificial intelligence to augment human intelligence [2]. Tools like this are beneficial because humans do not, however, always form the query to find what they are searching for in the best way possible; sometimes it is not even constructed in a functional way. The root ingredient of a keyword query is the words that compose it. While it may include Boolean operators such as “AND”, “OR”, “NOT” or “XOR” (exclusive or), these are operators upon the language of keywords. This dissertation comes in a step earlier, at the formation of the keywords themselves. The central idea is to score the keywords based on their relevancy, and different scoring techniques exist to do this. These include *term frequency – inverse document frequency (tf-idf)* and Best Match 25(BM25)[3]. Given a set of keywords and these scoring methods, a search engine will retrieve a ranked list of results based on a metric such as either of the aforementioned.

These keywords must come from somewhere, and they come in a typical scenario from the operator of an information retrieval system. The simplest case is when trying to retrieve all documents from a system that contain a keyword. In this example, the query is an exact match to the object to be retrieved; it is representative completely of the things to be found. Another example would be if I would like to find all the documents in a database that contain the phrase “Douglas Adams” exactly, and I retrieve all of the documents that contain the exact phrase “Douglas Adams”. This is an example of a look-up search, where the results are objective and clear – if the database and retrieval method is functioning properly, the results are deterministic and identical every time. This is contrasted against the example of an exploratory search [4]. An exploratory search is

Figure 1. Anecdotal evidence of the need for this system. This figure shows an example of someone finding out the pre-existing term for a concept they had within their mind, but were unaware of the common language used to communicate it.

open ended, does not have definitive results, and is far more subjective than a look-up search. Were the question to be asked, “Which documents in this system are written in a literary style similar to that employed by Douglas Adams?”, the results are far more subjective. This is a problem for exploring complex literary collections, including ones found posted on social media.

When the operator does not know the words to be used, they encounter the classical *unknown vocabulary problem* [5]. Various phenomena can lead to this problem, non-exhaustively including: momentary slips of the mind about a topic; a long time between trying to recall specific phenomenon; or the overall complexity of a topic. In both a look up and exploratory search, missing keywords result in fewer relevant results being retrieved. To solve this problem for a lookup search, a detailed study of the subject matter often will reveal the vocabulary that was previously absent. In the case of an exploratory search, the same case is unavailable – the results of the search are the documents that would be used to find the relevant vocabulary, and they are difficult to retrieve with information retrieval techniques precisely because the vocabulary that would help identify them from a collection of random documents is unknown.

The algorithms developed in this dissertation help mitigate the unknown vocabulary problem, particularly in exploratory search settings. One of the most closely related techniques to the algorithms described here is query expansion. Query expansion is the idea of taking a root keyword-based query, and using prior knowledge to find additional terms that carry the same meaning or will retrieve relevant documents. The exact start of when this technique entered academic literature is difficult to narrow down, but gained significant traction with the idea of using concepts to expand the query [6]. Query expansion itself is far from a solved problem, and has seen various iterations, which we revisit in more detail in the background of this work. The core idea to focus on is using prior information about the semantic relevancy of other words to the original query words to expand the digital net we cast out when performing a search.

The task of capturing semantic similarity in digital representations of words is not new, but has seen a resurgence with the idea of using distributed vectors to capture different senses of word meanings, and the resulting technique, *word2vec*, has made waves in natural language processing, and artificial intelligence as a whole [7]. While word2vec has some similarities to classical techniques such as Bag-Of-Words, it uses a neural network to better learn the association between words, instead of using a contextless frequency measurement. By doing so, it and related algorithms opened the door to associations between words being captured in new ways. Previously, to try and expand a word by semantic meaning required the usage of an ontology that had been made by a person *a priori*. As ontologies can require significant effort to create, this time commitment opened doors for techniques that could imitate an ontology in cases where an expert or expert curated ontology was unavailable.

One particularly interesting technique came from the idea of locally trained query expansion, which uses the same documents which are of interest for a particular task as input to a machine learning algorithm [8]. Owing to their local training, they possess unique vocabulary that may be unknown even to experts in the field, and by using an algorithmic or machine learning technique to learn associations between new vocabulary and keywords that may be familiar to experts, this technique is able to discover vocabulary that would have previously been unknown. In this way, query expansion can facilitate an exploratory search, by finding new words with which to search document collections.

This dissertation goes one step further by removing the task of finding even root keywords for expansion, and changes the formulation to indicating document collections of interest. The central contribution of this work is establishing an algorithmic approach for using natural language processing, machine learning and information retrieval to automate the task of query formation in distinguishing two sets of documents when the relevant vocabulary is unknown. This is a further development upon the idea of using query expansion to assist in query formation. By using the information contained in pre-trained word embeddings, we measure a specific metric of linguistic similarity between the two datasets, and build a classifier that can separate representations of the documents. This classifier is then used to identify the weight of individual words as they apply to differences between the two sets of documents. The weights of these words are helpful for fine-tuning an information retrieval task. We also explore how a relevancy heuristic

can be used in combination with the identified words to search for relevant documents without keyboards needing to be formed. With these computational aids, we reframe the task of query formation to identifying sets of documents rather than knowing the relevant vocabulary from the documents.

1.1 Structure of this dissertation

The remainder of this work is structured into five additional chapters. In the Chapter 2, we provide background on the techniques and theory used in the dissertation. This chapter also give examples of related work for some of the modular parts of the search techniques that the remainder of the dissertation establishes. It also contains a summary of machine learning methodology; how representations of words can be generated with those machine learning methods; the basics of information retrieval systems; the task of query expansion, including some history of the technique and some modern approaches; how language is modeled in NLP and some specific adaptions and problems that exist when using it on social media; and lastly on the idea of using social media analytics for public health.

In the Chapter 3, we detail the first modular algorithm for exploratory search with a classification method used to explore an unlabeled collection of documents for similar documents. We identify applications for this in query formation, and show how it can be used to assist the task of query expansion. Further, we highlight the aspects of it that are relevant for fine-tuning an information retrieval task to be suited to an end user's needs. This chapter details the modular components and how they can be substituted with other techniques to increase the versatility of the algorithm. This technique is demonstrated on finding individuals who match a profile of opioid abuse on the social media site Reddit.

In Chapter 4, we explore how we can use the weights learned by the classifier of our choice, and the learned representations of words themselves in our vocabulary using techniques like word2vec, can be associated to identify the individual words in the vocabulary which have the strongest association. This helps to illumine the black box classification approach found in the third chapter. By using a similarity metric between the word representations and a separating hyperplane between author representations, it is possible to get the algorithm to 'explain' its decision in terms of keywords. This is the central part to how we take our NLP and ML approaches and integrate them into an IR environment. Further, we show how we can take this approach and use it to estimate the prevalence of a geographic area by using language learned from opioid forums on Reddit to gauge the amount of opioid related dialogue of authors on Reddit affiliated with Ottawa.

In Chapter 5, we explore other ways we can apply the framework of looking for the differences in neural representations. We discuss the way that this framework can be applied to a time series analysis. From this, we can model evolving changes in dialogue over time. As the method extracts the largest difference between aggregations of word embeddings, this approach particularly favours the extraction of new content, as word embeddings attempt to keep semantically similar content near each other. This does not mean that new vocabulary around the same topic cannot be discovered, only that it will

be less significant on average than entirely new vocabulary. This last section involves a method which is specialized in understanding an individual's transition to and from archetypal states. With respect to privacy and the modern ethical environment around social media, we use partially synthetic data to demonstrate this technique's utility and effectiveness. The technique builds a model to detect depressive content, and synthetic documents are composed to show the algorithm's ability to extract pertinent keywords between the depressive and non-depressive states.

Chapter 6 provides a general discussion and conclusion; we discuss the unifying threads throughout these works, identify promising avenues for future work, and identify present limitations that could be overcome by advances in the fields involved in this dissertation. A summary of the algorithms in Chapters 3, 4 and 5 is presented, and their individual and collective contributions discussed. This dissertation concludes on some suggested considerations the use of the provided algorithms for future applications.

1.2 Introduction Bibliography

[1] M. H. Goodwin and C. Goodwin, "Gesture and coparticipation in the activity of searching for a word." *Semiotica*, vol. 62, no. 1-2, 1986

[2] S. Carter and M. Nielsen, "Using Artificial Intelligence to Augment Human Intelligence," *Distill*, vol. 2, no. 12, p. e9, Dec. 2017.

[3] R. (Ricardo) Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM Press, 1999.

[4] G. Marchionini, "Exploratory search," *Commun. ACM*, vol. 49, no. 4, p. 41, Apr. 2006.

[5] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987.

[6] Y. Qiu and H. P. Frei, "Concept based query expansion," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Infofmation Retrieval*, 1993, pp. 160–169.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in neural information processing systems* 2013.

[8] F. Diaz, B. Mitra, and N. Craswell, "Query Expansion with Locally-Trained Word Embeddings," *arXiv* May 2016.

2 Background

This dissertation develops new techniques in information retrieval and natural language processing. A discussion of how related techniques work is necessary to explain the contribution of the methods described in this dissertation. Further, the techniques in this dissertation are built upon a substantial amount of machine learning and representation learning methodology. A summary is provided of this methodology. An introduction to information retrieval is provided, as the algorithms that are provided in Chapters 3, 4 and 5 facilitate information retrieval. The other areas described in this background are query expansion, language modeling in NLP & social media, and social media analytics for public health.

2.1 Summary of Machine Learning Methodology

This dissertation's contributions are built upon several machine learning techniques. An overview of the relevant machine learning paradigms, techniques used in the dissertation, and relevant theory is provided. As the dissertation heavily makes use of word and word collection representations, it is given its own section to better describe its particulars.

2.1.1 Supervised & Unsupervised Learning

There are three main branches of machine learning: supervised, unsupervised, and reinforcement learning [1]. The referenced textbook is applicable as reference for the entire section. This work makes use of supervised and unsupervised learning, and future work could benefit from the integration of reinforcement learning techniques. The most used technique in this dissertation is supervised learning.

Supervised learning is one of the most common forms of machine learning, and is distinguished by the fact that the data involved has labels indicating the class to which each data point belongs. So, for a given data point, we at minimum know the class label that it could belong to. For an example relevant to this work, we could have a collection of documents that are labelled 'happy', and another collection that are labelled 'sad'. This is a case of binary classification, which this work focuses on, but it is important to note that there can be multiple classes that may or may not have a sensible ordering amongst themselves. In addition to these labels, each data point in question will also have associated features.

Features are data which describe each data point. They are meant to be useful for determining the class of each data point. To continue our text collection example, we could count the number of times that the word "joy" occurred as one of our features. It will contain different values for different documents and has an intuitive association with our label of 'happy'. It is, however, not foolproof either – a sad document could be about someone who is repeatedly complaining about how they cannot find joy. While a human reader could distinguish this with relative ease upon examination, attempting to get an automated algorithm, or supervised learning technique, to do so is comparatively harder. However, this is needed when the amount of data outpaces the ability for manual review to keep up with.

To try and find these patterns without human oversight, supervised learning learns to map the features of a data point to a predicted label using a provided set of labeled data. To assess the performance of the learned model, the data may be partitioned into a training set and a test set. In cases where the available data is small, a test set is omitted and techniques like cross validation are used [2]. Cross validation, briefly, involves a way of taking a single set of data and splitting it into *k* partitions, where *k* is the number of randomized groups that will be created. In both a train/test split, and in cross validation, the labelled data from the test set (or the labelled portion of the *kth* partition being used) is hidden from the supervised learning method. The method then uses the labels for the training set, and the features, to develop a classification procedure that it can use to label unseen data points based on the ones that the method has already seen. A measure of the accuracy of the training set on itself can be taken to see if the method is able to find any kind of pattern, but it is important to note that training set only testing is prone to finding local patterns and does not generalize to larger datasets well. This is the purpose of the test set – the values are given to the method, and labels produced, which are then compared with the real labels from the test set. The same procedure is repeated for the *k* partitions in cross validation as well. For binary classification problems with classes labeled "positive" and "negative," this produces the following four types of results: true positives (when the classifier indicates 'positive' and the data point really is positive), true negatives (when the classifier indicates 'negative' and the data point really is negative), false positive (when the classifier indicates 'positive' but the data point is in fact negative), and false negatives (when the classifier indicates 'negative' but the data point is in fact positive.)

From here, several other metrics can be derived. The most familiar of these is accuracy, which is simplify the number of predictions that were correct divided by the total number of predictions. However, there are several other metrics which can be used to better evaluate the performance of a trained classifier. The combination of the four metrics just listed can be organized into what is called a confusion matrix, and then this can be further refined into some commonly used metrics. We will list a number of these and the equation which defines them. Precision is defined as:

$$Precision = \frac{t_p}{t_p + f_p}$$

Where t_p is true positive and f_n is false negative. Similarly, t_n will be true negative, and f_p will be false positive. Recall is sometimes called sensitivity, hit rate, or true positive rate and is defined as:

$$Recall = \frac{t_p}{t_p + f_n}$$

There is also the idea of false positive rate (FPR) which is defined as:

$$FPR = \frac{f_p}{t_n + f_p}$$

Many classifiers can be adjusted to provide either better precision or better recall by changing a "threshold" at which they make a positive prediction. Together, Recall (or true positive rate, i.e., TPR) and FPR can be used to describe what is called the receiver operating characteristic (ROC). By plotting FPR and TPR on the x and y axis of a graph, respectively, for different thresholds, we generate the graph of the ROC curve. This helps to visualize the tradeoff between true positives and false positives, which can be adjusted after training a classifier. The curve is often summarized by the area underneath it, often called the Area Under the Curve (AUC). The AUC in this context represents the probability that a classifier will rank a randomly chosen positive data point as "more positive" than a randomly chosen negative data point. An illustrated example of the ROC curve can be seen in **Figure 2**.

Unsupervised learning differs from supervised learning in that there are no labels, so the entire idea of a training set, test set, or even a positive rate, has no meaning in this paradigm. Rather, unsupervised learning algorithms attempt to find structure among the provided data points using the feature values alone. The common thread between the two is the use of features in the data, which, again, cannot be monomorphic and must have diversity to be useful. The lack of any defined labels means that the notion of 'accuracy' is not appropriate. Regardless of the output of an unsupervised learning algorithm, there is no strict definition of how to assess the utility of the technique beyond its usefulness

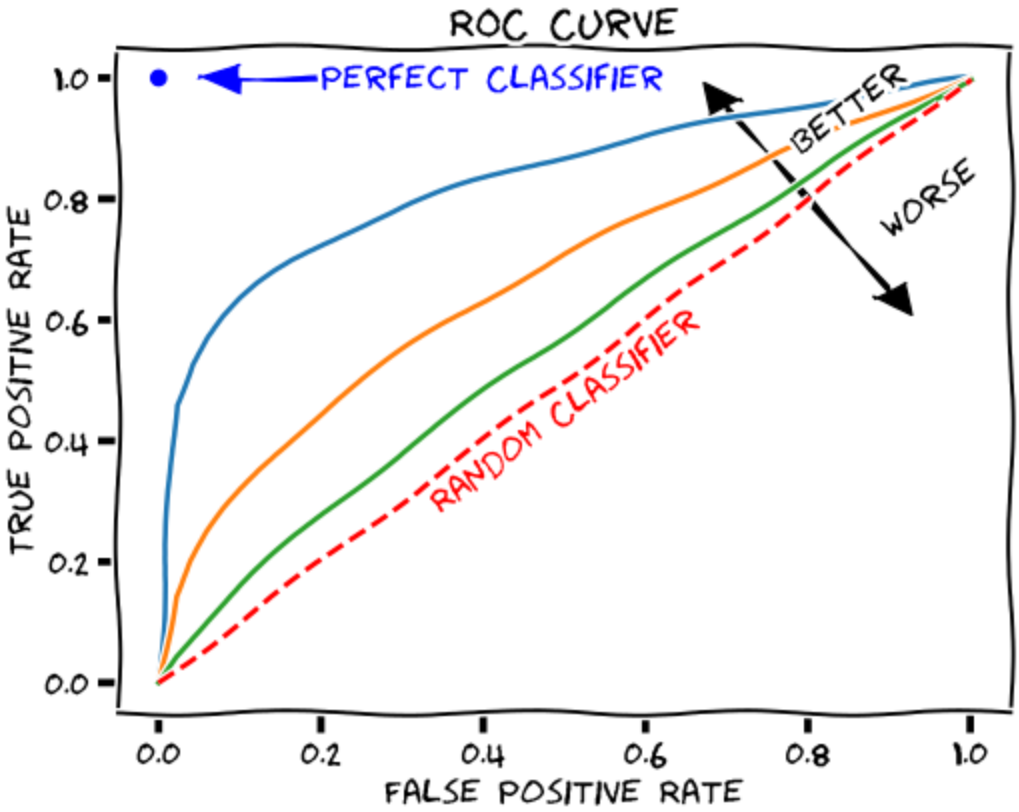


Figure 2 Example of a Receiver Operating Characteristic Curve with annotated explanations. Retrieved from Wikipedia, Creative Commons CCO 1.0 Universal Public Domain Dedication license.

for another task. This combines with the idea that unsupervised learning algorithms are often used as part of a collection of methods as a part of a whole. The effectiveness of the technique can then be assessed by the utility it offers in tasks, which themselves can be assessed for accuracy and utility. We describe one such example in the idea of *k*-means clustering.

One of the most common tasks we use unsupervised learning for is clustering. Given a set of *n* data points with *m* features, we want to know if we can find any structure that identifies subsets of *n* that have a difference which can be identified from the *m* features. Various means exist by which to form these clusters and find differences, but the interpretation of the differences must come *post-hoc*. With respect to the analysis in this dissertation, we deal with *Archetypes*, which are specialized groups within some whole. This dissertation uses unsupervised learning to aid in the task of representation learning for representations of documents, discussed later in this section.

Regarding the tasks in this dissertation, we would like to highlight that unsupervised learning can be helpful for identifying clusters that would not be identified simply by human intuition, but it also has no *a priori* reason to reproduce the labels and groupings that humans expect. With the right features, it is entirely plausible and happens that it can reproduce them, but this is not a reliable phenomenon and needs to be verified.

Most supervised and unsupervised learning methods rely on an optimization procedure. The most common ones, and the ones used in this dissertation, are detailed in the following section.

2.1.2 Loss Functions & Gradient Descent

Part of what allows machine learning algorithms to perform as well as they do in practice, is the choice of how we measure what is working well and what is not. This section is dedicated to what are called ‘loss functions’, which provide a numeric way of measuring just how well one of these algorithms is fitting the data provided to it. Further, there are regression losses and classification losses. Regression losses are the kind that would be used in linear regression, whereas classification is used by all the supervised learning methods used in this dissertation. This list of classification loss methods is non-exhaustive, and highlights the loss functions that are used in this dissertation. Interested readers that wish to go deeper can consult ‘Are Loss Functions All The Same?’, which we use as a reference for the following section [3]. We conclude with a discussion of gradient descent, which is an optimization procedure used to find the optimal values given these loss functions.

The first we will mention is the softmax function, as it features prominently in neural network applications, including ones used to generate representations of words. The topic is covered thoroughly in the textbook *Deep Learning* as it is used in many contexts [4]. For readers familiar with the logistic function, softmax provides a way to generalize it to multiple dimensions, which becomes important with distributed representations, high dimensional representations, and many neural network architectures. It takes as input a vector of real numbers, and normalizes it into a probability distribution where each component will be between 0 and 1. This makes it useful for selecting a single class out of many as the last layer of a neural network. As it is differentiable, it can also be combined with gradient descent, a popular optimization method described shortly.

The second loss function mentioned is one that is used in more methods in this work is the hinge loss objective function. Hinge loss objective functions are the most common for classification techniques [3]. Hinge loss is attractive as a loss function because it can be used for maximum margin classification. When trying to generate decision boundaries between classes, it is helpful to maximize the space between the two classes to try and find the most generic point that separates the classes. An example of how this can be beneficial over other possible lines that can separate the classes, see **Figure 3**.

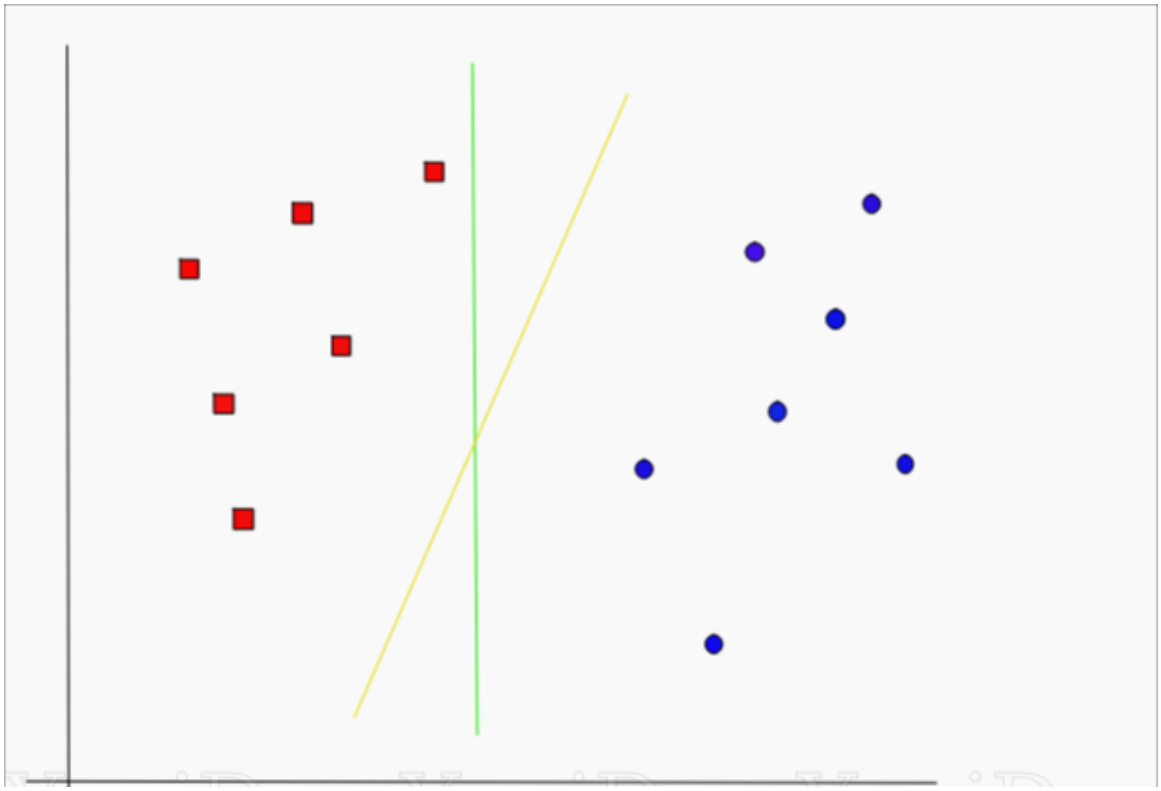


Figure 3. Illustration of the difference between a separating line in a classification system (green, vertical) and a maximum margin classification generated line (yellow, diagonal).

For the case of binary classification, the hinge loss formula can be represented as:

$$l(y) = \max(0,1 - t \cdot y)$$

Where l(y) is the loss, y is the output of the classifier's decision function, and t is either +1 or -1 depending on whether it is the positive or negative class. When both t and y are the same sign, i.e., predicting the same class, the loss function will be 0 at that point. When they are different, the loss will scale linearly and continue to penalize for increasingly misclassifying points.

Once a loss function is defined and a training set provided, training a classifier can be accomplished by optimizing the loss function. The last technique to discuss in this section is gradient descent. Gradient descent is used to find the local minimum of a differentiable function, and when the frameworks previously described are applied, allows for this optimization algorithm to find a minima [4]. If there is a multivariate function F(x) which is defined and differentiable, then we can measure the direction of the gradient to see which way the function is decreasing the fastest. By following this pattern repeatedly, the algorithm will continue in the direction of the minima until it finds it. One of the limits of this approach is that it can be prone to finding local minima and getting stuck there; there are various approaches to try and mitigate this, but it is important to highlight as a limitation of the technique that it does not guarantee to find the global minima, but only a local one. As it requires the function to be differentiable, it works with the softmax function, but does not work with hinge loss without modification or reformulation.

2.1.3 Support Vector Machine (SVM)

This dissertation makes frequent use of support vector machines (SVMs), which are a collection of models that fall under supervised learning [5]. Depending on how the problem is structured, SVMs can be used for classification or regression analysis. This dissertation uses them exclusively for classification. The idea behind SVMs is that they will simultaneously find a maximum margin separator while maximizing the accuracy of the resulting classifier.

In the case where the data is linearly separable, the SVM will find two parallel hyperplanes with extreme points on what is called a hard margin. The points on this hard margin are called the 'support vectors' from the name, and the maximum margin hyperplane is constructed between them.

When the data is not linearly separable, the SVM works with what is called a soft margin. Assuming we have training points in the form of:

$$(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)$$

Where the *x* values are a vector representing the features that will be used to predict the class, and the *y* values are the class labels. This allows the hyperplane to be represented as the set of points that satisfy:

$$\overrightarrow{w^T} \cdot \vec{x} - b = 0$$

It is this equation which is substituted for the value of 'y' in the hinge loss equation described previously, where t was the class label. From this combination, the formulation that the SVM wishes to minimize is as follows:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\vec{w}\cdot\vec{x}_i - b)\right)\right] + \lambda\|\vec{w}\|^2$$

Where λ is a tunable hyperparameter that represents the size of the 'soft' margin, which has an impact on how hinge loss is calculated. This can be adjusted during training. When this value is small, it becomes very similar to the normal hard margin classifier.

The next variation that is used in a few spots in this dissertation is to adjust the SVM to handle nonlinear classification. The system just described works very well in a linear space, but there are several situations where the decision boundary between classes could be expected to, or appears to, have nonlinearity. This need for nonlinearity is handled by using the 'kernel trick'. The kernel trick allows us to take the standard feature space and transform it, typically into higher dimensions, where a separating hyperplane will be able to find a more useful separation between the two classes [6]. This has the implication that, while the hyperplane is still a hyperplane in high dimensional space, it is likely nonlinear when brought back into the original untransformed space. This change allows the SVM to model a much wider variety of relationships, but, like many nonlinear classifiers, increases the risk of poor generalization and overfitting.

In many of the cases in this dissertation, we have enough data to justify the use of nonlinear kernels, and regularly find that the linear one performs best in a testing environment nonetheless. When discussing nonlinearity, it is presently more common to use neural networks, which we discuss next.

2.1.4 Neural Networks

Neural networks encompass a wide variety of methods under a unifying umbrella, and owing to their versatility, can be applied to a wide variety of tasks [7]. The core idea from neural networks is to take a set of features, like we have seen in SVMs, and to put them through a series of connected units that transform and pass the information between each other. These connected units are meant to mimic the connective structure of neurons in the brain, and by passing information from neuron to neuron through the network, come to a different representation of information by the end. When performing classification tasks, for example, there is often a softmax layer applied at the end such that each output neuron has a probability assigned, and the highest probability can be selected as the class prediction. It is easier to visualize a simple network of neurons than to describe; refer to **Figure 4** for a simple neural network.

The weights between the input layer to the hidden layer can be learned, and this is often done through the use of a method called backpropagation [8]. It is not completely removed from the idea of gradient descent; in fact, gradient descent is used as part of the backpropagation process. As can be seen in **Figure 4**, neural network architectures contain layers. Backpropagation works by computing the gradient of the loss function with respect to each weight using the chain rule, with each gradient being calculated per layer. It is called ‘backpropagation’ because it starts from the last layer and then proceeds backwards, right to left. In supervised learning, this allows for the network to learn and adapt from errors it generates in the training set until it minimizes the error. In unsupervised learning, this process is repeated until the metric being minimized or maximized is optimized.

With the popularity of deep learning, it is difficult to talk about neural networks without making reference to deep neural networks. Deep neural networks refer to neural networks that contain many layers, allowing them to have more functions and a larger potential information space to capture nonlinear relationships and associations. This idea behind neural networks being universal function approximators by the universal approximation theorem is that with enough complexity, enough depth, and enough data, they can approximate any function; a strong base from which to use them for complex problems [9]. Strictly speaking, this capability is associated with generic neural networks as well, but is emphasized more in the context of deep networks.

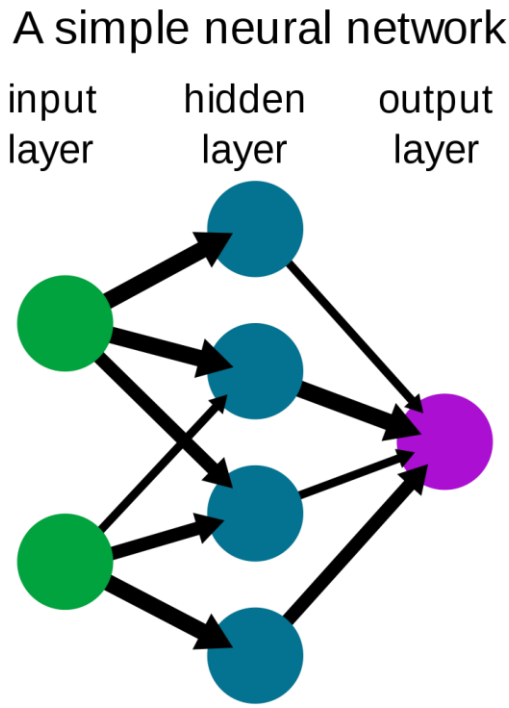


Figure 4. A simple neural network diagram from wikipedia. Attribution: By User:Wiso - from en:Image:Neural network example.svg, vectorialization of en:Image:Neural network example.png, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5084582

Typically, deep learning refers to a single network architecture which contains many layers. However, and we quote from the landmark paper *Deep Learning*, "Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level." [10]. This dissertation deals with a multi-step representation learning problem that goes from individual word representations to document or author representations. However, as the usage typically refers to a single network architecture with many layers, we do not refer to any part of this process as deep learning. Nonetheless, the task of generating author represention shares much in common with the idea of learning multiple stage representations. This idea of representation learning forms the last section of the machine learning methodology for us to consider.

2.1.5 Representation Learning

While we have touched on how both SVMs and neural networks can be used to classify data into predefined labels, there is another branch of the machine learning tree that is representation learning. Representation learning is involved in finding a way to quantify into numerical form the properties of objects, abstract or real, in a way that is helpful for other tasks. Unsupervised learning can often generate these representations, and this is the case with a classical technique, Principal Component Analysis (PCA) [11].

PCA is one of the most common techniques in machine learning, and is used to find the largest sources of variance in each set of numerical features in the form of principal components. This, immediately, is a form of representation learning. The information space that the features existed in has been transformed to a series of principle components. As some datasets have many features, or involve representations that are high dimensional vectors, this can reduce the number of uninformative features. This is the idea behind dimensionality reduction with PCA [12]. This has applications in other areas, such as visualization, as it allows for a lower dimensional approximation of high dimensional data which can be seen in 2 or 3D.

From the relatively simple PCA, there is also another way to try and condense data called *t-Stochastic Nearest Neighbour Embedding* (t-SNE) [13]. By computing probabilities that are proportional to the similarity of objects as defined by their features, the technique groups together similar points and moves dissimilar points further away, while mapping higher dimensional data onto a lower dimension. Further, it contains adaptations so that spots that have higher density of information are still afforded the information contained there. This technique is often used for visualization. An example of a 3-dimensional visualization of author representations using t-SNE is provided in **Appendix Figure 21.**

The two previous techniques are examples of way that representations of data can be learned from features. These new representations can make relationships that were previously obscured clearer. The task of representation learning is large and continues to find new applications. Literature surveys exist which cover a wide variety of representation learning and associated tasks [14]. Techniques that are involved in this include probabilistic models, autoencoders, manifold learning techniques and deep learning itself. One such representation learning task that appears prominently in this dissertation are word representations, which form the next section of this work.

2.2 Representations of Words & Word Collections

2.2.1 Representations of Words

To appreciate the information available for classification purposes for the techniques involved in this dissertation, it is necessary to understand the kinds of information that are available to be captured in representations of words. For example, one of the most natural ways that humans use to impose a structure on language is a taxonomy of 'nyms'. We have, for example: synonyms, such as joyous and happy; antonyms, such as happy and sad; and homonyms such as light and, which could refer to the physical phenomena of light or could refer to something not being particularly heavy. It can be seen here the difficulty of the task of capturing the meanings of these words in vector form; there is nothing about the alphabetical letter composition of the words happy and joy that would relate they signify the same concept, and there is a subjectively different sense of the same positive sensation conveyed by each word.

The abstract goal when creating a representation of a word is multifaceted and complex. It non-exhaustively includes capturing the concept which the word signifies, the potency with which it indicates it (this example made me *mad* contrasted with this example made me *furious*), the hierarchy of potencies – although adjectives and subjectivity make this difficult even between humans – and whether the word was used with sincerity, in jest, in sarcasm, or in a pun. We would like words that are synonyms to be more similar with each other than ones which are not, we would like antonyms to have some understandable relationship that separates them as opposites, and we would like the system to understand the complexity of words with multiple meanings like homonyms.

The way this modeling is done in many modern NLP efforts is with neural representations of words. The first that comes to mind for many people is the usage of Mikolov *et al.*,'s word2vec [15]. Before distributed representations of words were established, each word in consideration is given its own unique entry in a matrix consisting of the vocabulary on one axes and the numerical spot in the document on the other. Every word in the vocabulary is given a unique entry, and the entry is 1 or 0 for a given spot in a document corresponding to whether it is or is not that word in that spot. This has been called 'one-hot encoding', and is an implicit, albeit simple, representation learning. The unique column of word identity in the matrix of words & document location is the representation of the word in a very basic form. Independently, this representation does not offer much that allows us to model word properties, aside from their frequency.

Word2vec builds upon this foundation by introducing a notion of word similarity that is approximated using a machine learning approach. Notably for the time it was done, this was done through the of an artificial neural network. As manually curating words is time intensive, that this was able to be done in an unsupervised learning framework was notable. With this optimization procedure, a training corpus is provided which contains

sentences written by humans. The training data can be from various sources, but is often sourced from news articles, Wikipedia pages, or social media. From this source, word2vec uses one of two model architectures to produce what is called a distributed representation of a word. A distributed representation is different than the one-hot encoding we recently described in that it contains multiple components, or layers, or dimensions [16]. No single dimension contains all the information needed to construct the entire representation, hence the 'distributed' nature. By increasing the number of places that information can be contained, these distributed representations offer an extended information space wherein different kinds of information can be contained.

Once the idea of distributed representations took hold in the literature, two more methods in the same vein followed. Specifically, these include Global Vectors (GloVe) and FastText [17], [18]. GloVe is the technique used to create representations of words, and performs novel operations with matrix factorization to generate co-occurrence statistics that are better for using metrics like cosine or Euclidean measures. FastText bears many similarities to word2vec and was notable that it made use of subwords to better infer possible relationship with words that are said to be out-of-vocabulary (OOV).

The most recent advance in word embeddings is that of contextual word embeddings. Since we did not find that it made sense to use the contextual embeddings in the works applied here, but they are popular and show promising results, we briefly explain why we did not use them. Other practitioners may find they wish to explore their usage. Further, as a sign of the pace of this field in times contemporary to this dissertation, many of these techniques would not have been published yet. These techniques include, non-exhaustively, BERT, XL-NET, MEGATRON, and TURING-NLG [19], [20], [21], [22]. The common thread between this new wave of models involved deep learning, a new neural network architecture called the Transformer, and a relative explosion of the number of parameters and training size [23]. Where this work involved the creation of multiple pre-trained word embeddings with GloVe, training one of these models requires a massive amount of computation that is not easily accessible to most researchers. Further, given a large enough training set, there are recent works suggesting it may not be worth the effort to do so for many applications [24].

One of the advantages of contextual word embeddings is that words are given a specific representation based on the context they occur in, rather than having a single vector representation. Potentially ironically, it is this contextual locking which makes them ill-suited for the exploratory search task, where we are trying to ensure words have a single meaning by which to compare to others. As relatively new techniques, the author was unable to find any research which quantify the amount of diversity that contextual words can have with each other, and the exact amount of different context necessary before a new representation is formed. Further, even generating a complete enumeration of all the representations for different contexts of a word is not easily doable. These techniques may offer great benefits to these algorithms later, but we find them to be currently unsuitable.

2.2.2 Neural Representations of Authors or Documents

Building from the idea of representing words in vector form comes the thought of more complex representations of symbolic meaning that are of interest. A natural extension of words, which are meant to be independent symbols of meaning that aggregate into larger structures to represent thought, are sentences. This is where the idea of a thought vector, popularized by Dr. Geoffrey Hinton, and appearing in the literature as Skip-thought vectors, is grounded [25]. This is another unsupervised learning technique, like word2vec, although it is instead focused on capturing the semantic meaning of a sentence.

Typically, this is done through aggregating individual units of meaning in the form of words. One of the limitations of this approach is that when trying to jointly learn both word meaning and sentence meaning, there is a sharp increase in the complexity of the task. Further, this can limit the model's ability to be transferred to new environments without retraining the entire model. This notable because these methods continue to get more complex and involve more stages of learning; recent works have adapted sentence embeddings have improved task specific performance by transforming them between a learned BERT sentence embedding space and a standard gaussian latent space that is isotropic [26]. This concept of adapting a learned word aggregate representation to a task will be seen in a varied form in this dissertation.

From the idea of sentence embeddings, comes the idea of document embeddings. For the task in this dissertation, finding archetypal behaviours, we are not interested in sentences of interest but rather authors matching an archetypal profile. From words and sentence representation, there is a need to go one step deeper into author or document representations. The idea of using neural representations to aggregate an individual's language into a single high dimensional representation is found in usr2vec [27]. Interestingly, this approach followed an attempt to model one of the *simplest* in NLP, sarcasm [28]. Most representations and analyses with them would, for example, have difficulty identifying that the word *simplest* in the previous context was sarcastic. This learning is done by maximizing the conditional probability of:

$$P(C_j|u_j) \propto \sum_{S \in C_j} \sum_{w_j \in S} logP(w_i|u_j)$$

Where U is a set of authors, C_j is the collection of posts generated by a given author $u_j \in U$, and $S = \{w_1, \ldots, w_N\}$ be a document or other collection of words (ex: social media post) that come from a vocabulary V. However, this is not a quantity which can be directly estimated without severely computationally expensive procedures. To reduce this

computational burden, and because we are only interested in the values of the author, document, or user vectors and not the probabilities, we approximate the term $P(w_i|u_j)$. This is done using a Hinge-loss objective as follows:

$$L(w_i, u_j) = \sum_{\widetilde{w}_k \in \mathrm{V}} \max\,(0, 1 - w_i \cdot u_j + \widetilde{w}_k \cdot u_j)$$

Where $\widetilde{w}_k$ is a word embedding from the vocabulary of pre-trained word representations *V*, and specifically a negative sample; this is a word not occurring in the sample that was written by user u_j. This particular technique pre-dates word embedding, but is used to learn how to discriminate between positive and *pseudo*-negative samples such that the probability mass is shifted towards more plausible observations [29]. This complete procedure is what is called 'usr2vec' or 'USER2VEC'. The optimization is done using gradient descent to minimize the hinge loss objective.

From here, it is common to use deep learning to combine the user context with a specific sentence context when trying to perform applied downstream tasks, as in 'Content and User Embedding Convolutional Neural Network' [28]. The algorithms in this dissertation want to work with the author, or user, representation directly and do not combine them with intermediate or ancillary representations.

2.3 Information Retrieval Systems

Information retrieval systems exist to retrieve the result of a query from somewhere information is stored. The most common example of this that most people encounter in their day-to-day lives is using a search engine on the internet to find websites, images or other content that is related to the query they put into the search engine. The central idea behind these systems is providing retrieval of information of unknown location that a user queries the system for. This could be as simple as retrieving one file out of thousands, or as complex as finding all matches of a query in a system that is distributed across the globe, or even into space.

As information takes many forms, it follows there are many branches of information retrieval. This can include image retrieval, music retrieval, speech retrieval, video retrieval, and text retrieval [30]. This dissertation focuses on the task of retrieving text, which is typically done through search engines and related techniques.

Beyond the idea of finding exact matches comes the idea of ranking documents that are more relevant to the query higher than others which may match, but may not match the full query, or have a lower similarity score, or some other defined metric of similarity. The most common metric here is term frequency – inverse document frequency (tf-idf) [31]. The metric can be described as:

$$w_{i,j} = tf_{i,j} \times log \frac{N}{df_i}$$

Where $w_{i,j}$ is the tf-idf score for a given word i in document j, $tf_{i,j}$ is the number of occurrences of the word i in document j, df_i is the number of documents that contain the word i, and N is the total number of documents. This allows for a measure of the impact of a term adjusted by how frequently it occurs in a document. Common grammatical words like the will appear frequently and have low impact, but specialized words like verbosity or quintessence are more elusive; finding rare words like this suggests the results are more impactful, and the tf-idf metric helps with this.

A more recent advancement from tf-idf is BM25 (sometimes called Okami BM25) [32]. BM25 fits in the class of probabilistic relevance frameworks, which take advantage of statistical properties and heuristics to further tune the relevancy of search results. The metric itself is described by:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Where score(D,Q) is the score of a document D given a query Q that contains keywords $q_1,\ldots,q_n$. The terms $f(q_i,D)$ refer to the term frequency in Document D, |D| is the length of the document D as measured by the number of words, and avgdl is the average document length in the text collection. The terms k and b are free parameters, which are given a default value by the search engine that implements the metric. In this context, inverse document frequency (IDF) is calculated as:

$$IDF(q_i) = ln\left(\frac{N - n(q_i) + 0.5}{n(q_{i)} + 0.5} + 1\right)$$

Where N is the total number of documents, $n(q_i)$ is the number of documents containing the word i in the query. There are a few variations of this that exist, which can further vary between implementations. The information retrieval system used in this work is Lucene, which is an open source implementation [33]. This system is further integrated into a search engine called Elasticsearch, which allows for large scale usage of these metrics [34].

2.3.1 Query Expansion

The root idea behind query expansion is to take a pre-existing query, and increase its ability to find relevant information. This is something that can be done manually, and this is likely the simplest way. If the query is, for example, “cat”, a few moments reflection will usually bring about other words which are relevant. Words such as “kitten” or “feline” communicate a different sense than the word “cat”, but would capture more documents that included information about cats than just the original query. However, this task too suffers from the unknown vocabulary problem; if a suitable root word cannot be thought of, there is nothing to expand [35]. Still, it is one of multiple tools that demonstrate where artificial intelligence, and other techniques, for augmenting human intelligence [36].

A clever observation from here is that the words mentioned are synonyms of the original word, cat. From an object-oriented programming perspective, these are all instantiations of the concept of a cat. This is the idea behind one of the early works on query expansion, which dealt with what it calls concept based query expansion [37]. Concept based query expansion involves using a probabilistic model combined with a similarity thesaurus; many query expansion methods of the time were based on trying to find words that had similar spellings, rather than semantically similar meanings. By trying to find words with similar meanings the relevancies of the information retrieval results were improved.

Efforts like this have been aided by the construction of massive lexical sources of meaning, such as WordNet [38]. WordNet groups nouns, verb, adjectives and adverbs into sets of what it calls cognitive synonyms, or synsets, which each express a distinct concept. The links between these synsets can be defined as well, allowing for conceptual-semantic and lexical relations. For information retrieval, the presence of efforts like WordNet enabled the work of concept-based query expansion and related work. Largely through manually effort, there was a computational way to relate words to each other.

This idea of using information repositories to inform computational linguistic tasks would continue to see a rise in the field of query expansion. This included looking through the results of the original query to see if there were words which seemed particularly informative for finding more documents after the original query returned a set of results [39]. By finding words from the discovered documents and applying a method to assign relevancy scores to candidate expansion terms, this method can retrieve documents that would not have been found by doing a strict keyword match. This is one way where query expansion can start building a bridge between a look up search and an exploratory search, which as we have discussed, has unknown outcomes [40].

Further efforts with WordNet and query expansion brought the idea of adaptation to the local environment. Now ubiquitous and integrated into daily life, the world wide web was coming to greater prominence at the same time as the rise of many of these techniques. This saw the generation of techniques such as web query expansion by WordNet, which is notable for taking the context of the Web and integrating it with the formal synsets of WordNet [41]. This kind of query expansion used something called a collection-based term semantic network which used word co-occurrence in the collection. While this appears to not have garnered much attention at the time, this idea is remarkably like one found in later works that incorporate advances in NLP into query expansion. Indeed, it was around the same time a method for query expansion started to take into account term relationships inside language models to enhance the results of these query expansions [42]. This too was based around an idea of capturing some of the context of the words for a query expansion.

With the awareness that the context words were being used in mattered, there was a simultaneous rise in the ways that query expansion could be adjusted for the specific case it was being used for. This was the case with a query expansion technique that focuses on personalizing to the individual user that was searching the web with the technique [43]. This technique still took advantage of co-occurrence statistics and thesauri, but it included the context of the user's previous search history to try and better inform the expanded query results.

From these handfuls of specialized techniques, there came the idea of using ontologies that could be customized to purpose. These ontologies describe a set of formal names, definition of categories, properties and relationships between the concepts data and entities that exist within the ontology [44]. It is sensible that such a grouping, like the previously described WordNet, could offer utility when trying to find related words in the context of query expansion. At about the same time as the previously mentioned papers were coming out, there was a number of ontology based query expansion methods which have been summarized well [45].

When it was available, a pre-defined ontology helped better formulate information system queries. Query expansion in this context has been included in search and triage systems based around ontologies [46]. Ontologies in contexts like this one, which included searching PubMed using technical language, have been shown to be helpful when the associated language is only available to experts. Specific implementations of query expansion with medical language have been performed [47].

The previous examples of query expansion serve as reference for techniques that were established before the introduction of word embeddings. With the introduction of word2vec, there was a new way by which words could be related to each other [15]. As described in the first chapter of this work, word2vec generates the word representations intends to group similar words together and preserve other semantic relationships in vector space. Given that the task of query expansion is to find similar words, it was sensible for the query expansion research community to wonder if this could be used in a similar manner to an ontology to retrieve similar words. Beyond being sensible, it was done and saw success.

One of the earliest works to build word2vec into query expansion goes back to the idea of concept based query expansion, by using quantum entropy minimization to learn concept embeddings [48]. This technique had success in improving previous benchmarks, and opened the door to using word embeddings for query expansion in other contexts. Since the words representations are meant to be clustered near together in n dimensional space, query expansion techniques with pre-trained word embeddings performed a k-nearest neighbour search, where k is the number of words that are to be added to the query [49].

One of the limitations of this approach is that the relatedness of query relevant words may be quite dependent on the context that these words occur in. This motivated the idea of training a word embedding specific to task, and also simultaneously validated that both word2vec and GloVe were useful for the query expansion task [50].

There are also ways to use deep learning to try and evaluate which words in a query expansion are useful; this technique is appealing for the output of the algorithms in this dissertation, but the method of training them would require large amounts of labelled data that would need to be annotated by specialized domain experts [51]. Other recent methods in query expansion involve combining classical lexical resources and word embeddings, with a focus on retrieving sentences rather than keywords [52]. There has also been success in using specialized query expansion for retrieval of posts from microblogs such as Twitter, which supports the usage of similar methods like those in this dissertation on social media content [53].

2.4 Data Sources for Language Modeling: Social Media

The idea of studying the language something is used with to try and understand more abstract meaning is a familiar concept; taken very abstractly, this could be considered reading. One of the changes that has happened relatively recently with the increased prominence of social media is the explosion in availability of data generated by individuals. Previously, large text collections would come from business related purposes, writers, or legal documents. These studies, while useful for evaluating techniques on a shared dataset, are not specific to all of the tasks approached in this dissertation. Further, for the study of human behaviour, social media presents a new avenue of information; it comes with challenges of big data scale sizes and causal inference, but offers an amount of specialized information that is nigh impossible to match [54].

These efforts have been helped by the availability of large scale open source repositories such as PushShift [55]. This repository hosts a variety of posts collected from social media platforms. This dissertation uses it to retrieve a multitude of specialized data from Reddit, but it also contains data from Twitter, Telegram, and other sources. There are few, if any, equivalents which can be thought of to a platform which contains so much data authored by such a wide variety of people. The only contenders are the social media platforms themselves, and they typically do not offer data openly and freely like PushShift does. This new availability and the data intensity of machine learning methods provided motivation to focus on social media rather than traditional document collections.

However, this amount of data is not without its costs. Beyond the practicalities of being able to download, store, and filter this data, there is also the task of retrieving relevant information from it. While we have an armament of information retrieval techniques available to handle it, retrieving high quality information from social media is still documented to be a hard problem [56].

This has not prevented others from going and using Twitter to collect data for research [57]. Various ways of collecting and processing data exist, and corresponding methods can vary. One of the unique facets of social media is how with the right approach it can be used to generate and collect primary data [58]. There is likely little opposition to social media’s ability to act as data generator in the modern context, but analyzing the data comes with its own challenges.

First, when using NLP, there is typically some kind of pre-training or initializing step done when using word embeddings or machine learning techniques. This is typically sensible to do, as the single training of a large language model with current NLP techniques can be time, energy (in reference to electricity and effort) and cost. The problem emerges, then, about what to do with what is called non-canonical language [59]. This can include data that is generated by underrepresented demographics, modern

vernacular, and slang, memetics, and all of the other kinds of text that socially media is resplendent in. All of these increase the difficulty in discovering topics, collecting data with information retrieval, and preparing the data for future analysis [60].

The non-canonical language problem, and even the canonical language problem, is further complicated by the way language is used. Word relatedness is not a static thing, and while the methods described in section 2.2 approximate this relation, they cannot do so in perpetuity [61]. There have been approaches to model this, but it is a difficult task even when dealing with a known and defined vocabulary [62]. To summarize so far: the full vocabulary that will be encountered in an applied setting will be unknown; the relationships between this vocabulary will change over time, so even historical data will be limited in its efficacy; and information retrieval, along with query expansion, rely upon these to be effective.

This is amongst the most important problems that the techniques in here try to mitigate. Before the mitigations are discussed, it is important to stress why addressing these is imperative for any kind of word in an applied setting. Research in this setting has shown that the language we use is a proxy for our identity in a digital setting; this is sensible, as the only way someone online can be understood is through the use of their language, excluding images, videos (containing language), and emoticons [63].

Continued research into the language that we use online to express ourselves has revealed that it contains associations towards our gender, age, personality, and various other kinds of signifiers for identity that most people do not remain conscious of, nor conscious of expressing [64]. As one author puts it, these individuals are essentially putting autobiographies of themselves for public consumption out on social media, including a detailed composition of how their identity is changing over time with respect to the ways they write and engage in online dialogue [65].

If this was not a complex enough concept already, it is further complicated because some social media is anonymous, while others are not. It is near universally accepted that people behave differently when anonymous and when they are not. Depending on the part of Twitter, anonymity may be the standard or rare. So-called ‘Academic Twitter’, for the most part, uses real names and identity with which to communicate. Reddit, conversely, has anonymity as the standard. It very rarely comes as a surprise when it is mentioned that this anonymity can have a large impact on the ways that people conduct themselves online [66]. However, this can have its benefits. There are topics that are not easily discussed when it must be shared with a permanent identity. This dissertation, for example, was applied in a setting to try and estimate the rate of illicit opioid drug usage in Ottawa, Canada in concert with the Ottawa Public Health. Very few people come forward and discuss their drug habits, particularly if they have an otherwise functional life and role in society.

Briefly, the algorithms in this dissertation attempt to address these with a few mitigating efforts. Algorithms can be trained on a set time window, allowing them to have temporally informed word associations from the underlying word and author representations. The vocabulary learning is learned directly from the users who self-

identify with a particular behaviour or identity, i.e., opioid usage. Relatedly, there is an extensive literature detailing health and public health applications of social media.

2.5 Application: Assisting Public Health Decision Making

Social media presents several opportunities for assisting public health decision making, but these opportunities are made difficult by complexities of social media. Some of these difficulties, too, are not unique to social media. The first that is impactful for studying public health relevant topics is the difficulty encountered in surveying some populations [67]. Whether by traditional surveys, or by reaching out on social media, there are populations which are reluctant to answer the questions that would help public health stakeholders better understand the difficulties these populations encounter. Often, this is because of a stigma that is associated with the population in question. For example, users of opioid injection drugs are often targets of stigma and discrimination, and this makes members of this population reluctant to come forward, digitally or otherwise, to share their experiences [68], [69].

This difficulty in getting engagement from these populations has motivated a different approach, which is passive surveillance on social media of accounts that display these behaviours in some way. There are several works in this area. For example, multiple efforts have been made to monitor food poisoning outbreaks based on reports of food poisoning [70]–[74]. The common thread here is using some method to decide whether a tweet involves food poisoning or not, and then use some kind of geographic association – either from the tweet or inferred using a method that can estimate a user's location – and use it to look for outbreaks earlier than they would be reported to a health unit [75]. As an aside, this was the original motivation for the public health application of the methods in this dissertation; changing priorities caused it to be retargeted onto the opioid epidemic, but there is no reason the method could not be adapted to this domain at another time.

With regards to opioid drug surveillance on social media, there have been previous efforts to show that this is worthwhile for public health stakeholders to be involved in. While not linked to opioids, a study of e-cigarettes on social media uncovered new vocabulary and products that were unknown to public health stakeholders [76]. Foundational efforts here include finding evidence of codeine abuse on Instagram [77]. Further, efforts have been made to study the easily findable opioid drug content on Reddit to better understand the opioid epidemic through the lens of social media [78].

A common thread through these studies is that they take the approach of information provider, and that is the role that these algorithms take as well. The journey from discovery to implementation is much longer than that of discovery, and requires careful consideration by experts in the domain. The contribution of this dissertation is in discovery, rather than public health policy. The algorithms described in the following chapters offer new ways to discover content on social media surrounding the opioid epidemic, and other public health efforts.

2.6 Background Bibliography

[1] "Introduction to Machine Learning - Ethem Alpaydin - Google Books." [Online]. Available: https://books.google.ca/books?hl=en&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&ots=F3SX8Y6pBh&sig=ODi1OqDQIpgPIZ6ighG_-wolZus&redir_esc=y#v=onepage&q&f=false. [Accessed: 13-Apr-2021].

[2] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat. Surv.*, vol. 4, no. none, pp. 40–79, Jan. 2010.

[3] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri, "Are Loss Functions All the Same?," *Neural Comput.*, vol. 16, no. 5, pp. 1063–1076, May 2004.

[4] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*, vol. 1. MIT press Massachusetts, USA:, 2017.

[5] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks Editor," Kluwer Academic Publishers, 1995.

[6] M. Hofmann, "Support Vector Machines-Kernels and the Kernel Trick An elaboration for the Hauptseminar 'Reading Club: Support Vector Machines,'" 2006.

[7] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[8] D. E. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin, "Backpropagation: The basic theory," *Backpropagation Theory, Archit. Appl.*, pp. 1–34, 1995.

[9] F. Scarselli and A. Chung Tsoi, "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results," *Neural Networks*, vol. 11, no. 1, pp. 15–37, Jan. 1998.

[10] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, 27-May-2015.

[11] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987.

[12] M. Partridge, "Fast dimensionality reduction and simple PCA," *Intell. Data Anal.*, vol. 2, no. 3, pp. 203–214, Jan. 1998.

[13] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, Nov 2008.

[14] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *J. Financ.*

Data Sci., vol. 2, no. 4, pp. 265–278, Dec. 2016.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in neural information processing systems* 2013.

[16] A. Preece, D. Braines, F. Cerutti, and T. Pham, "Explainable AI for Intelligence Augmentation in Multi-Domain Operations," Oct. 2019.

[17] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," 2014.

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Jul. 2016.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," Jun. 2019.

[21] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," *arXiv*, Sep. 2019.

[22] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: Memory optimizations toward training trillion parameter models," in *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2020, vol. 2020-November.

[23] N. A. Smith, "Contextual Word Representations: A Contextual Introduction," *arXiv*, Feb. 2019.

[24] S. Arora, A. May, J. Zhang, and C. Ré, "Contextual Embeddings: When Are They Worth It?" *arXiv*, May 2020.

[25] R. Kiros *et al.*, "Skip-Thought Vectors*,"Advances in neural information processing systems*, 2015.

[26] B. Li *et al.*, "On the Sentence Embeddings from Pre-trained Language Models." *arXiv*, Nov 2020.

[27] S. AMIR, G. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace, "Quantifying Mental Health from Social Media with Neural User Embeddings," *Machine Learning for Healthcare Conference*, Nov 2017.

[28] S. AMIR, B. C. Wallace, H. Lyu, P. Carvalho, and M. J. Silva, "Modelling Context with User Embeddings for Sarcasm Detection in Social Media," *arXiv* Jul

2016.

[29] N. A. Smith and J. Eisner, “Contrastive Estimation: Training Log-Linear Models on Unlabeled Data *,” *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Jun 2005.

[30] R. (Ricardo) Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM Press, 1999.

[31] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries.” *Proceedings of the first instructional conference on machine learning,* vol.242, no. 1, Dec 2003.

[32] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Found. Trends® Inf. Retr.*, vol. 3, no. 4, pp. 333–389, Dec. 2010.

[33] A. Trotman, C. L. A. Clarke, I. Ounis, J. S. Culpepper, M.-A. Cartright, and S. Geva, *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval*. 2012.

[34] C. Gormley and Z. Tong, *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. “ O’Reilly Media, Inc.,” 2015.

[35] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, “The vocabulary problem in human-system communication,” *Commun. ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987.

[36] D. C. Engelbart, “Augmenting human intellect: A conceptual framework,” *Menlo Park. CA*, 1962.

[37] Y. Qiu and H. P. Frei, “Concept based query expansion,” in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Infofmation Retrieval,* 1993, pp. 160–169.

[38] G. A. Miller and G. A., “WordNet: a lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[39] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, “An information-theoretic approach to automatic query expansion,” *ACM Trans. Inf. Syst.*, vol. 19, no. 1, pp. 1–27, Jan. 2001.

[40] G. Marchionini, “Exploratory search,” *Commun. ACM*, vol. 49, no. 4, p. 41, Apr. 2006.

[41] G. Zhiguo, W. C. Chan, and H. U. Leong, “Web query expansion by WordNet,” in *Lecture Notes in Computer Science*, 2005, vol. 3588, pp. 166–175.

[42] J. Bai, D. Song, P. Bruza, J. Y. Nie, and G. Cao, “Query expansion using term relationships in language models for information retrieval,” in *International*

Conference on Information and Knowledge Management, Proceedings, 2005, pp. 688–695.

[43] P. A. Chirita, C. S. Firan, and W. Nejdl, "Personalized query expansion for the web," *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR'07*, pp. 7–14, 2007.

[44] B. Smith and C. Welty, "Ontology: Towards a New Synthesis." *Formal Ontology in Information Systems*. vol. 10, no. 3, Oct 2017.

[45] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Inf. Process. Manag.*, vol. 43, no. 4, pp. 866–886, Jul. 2007.

[46] J. Demelo, P. Parsons, and K. Sedig, "Ontology-Driven Search and Triage: Design of a Web-Based Visual Interface for MEDLINE," *JMIR Med. Informatics*, vol. 5, no. 1, p. e4, Feb. 2017.

[47] P. Srinivasan, "Query expansion and medline," *Inf. Process. Manag.*, vol. 32, no. 4, pp. 431–443, Jul. 1996.

[48] A. Sordoni, Y. Bengio, and J.-Y. Nie, "Learning Concept Embeddings for Query Expansion by Quantum Entropy Minimization," Jun. 2014.

[49] S. Kuzi, A. Shtok, and O. Kurland, "Query expansion using word embeddings," in *International Conference on Information and Knowledge Management, Proceedings*, 2016, vol. 24-28-October-2016, pp. 1929–1932.

[50] F. Diaz, B. Mitra, and N. Craswell, "Query Expansion with Locally-Trained Word Embeddings," *arXiv* May 2016.

[51] A. Imani, A. Vakili, A. Montazer, and A. Shakery, "Deep neural networks for query expansion using word embeddings," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11438 LNCS, pp. 203–210.

[52] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, "Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering," *Inf. Sci. (Ny).*, vol. 514, pp. 88–105, Apr. 2020.

[53] Y. Wang, H. Huang, and C. Feng, "Query Expansion with Local Conceptual Word Embeddings in Microblog Retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1737–1749, Apr. 2021.

[54] J. Grimmer, "We are all social scientists now: How big data, machine learning, and causal inference work together," in *PS - Political Science and Politics*, 2014, vol. 48, no. 1, pp. 80–83.

[55] J. M. Baumgartner, "Pushshift API." 2018.

[56] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding high-quality content in social media,” in *Proceedings of the international conference on Web search and web data mining - WSDM ’08*, 2008, p. 183.

[57] S. Yoon, N. Elhadad, and S. Bakken, “A practical approach for content mining of Tweets.,” *Am. J. Prev. Med.*, vol. 45, no. 1, pp. 122–9, Jul. 2013.

[58] W. Clyne, S. Pezaro, K. Deeny, and R. Kneafsey, “Using Social Media to Generate and Collect Primary Data: The #ShowsWorkplaceCompassion Twitter Research Campaign.,” *JMIR public Heal. Surveill.*, vol. 4, no. 2, p. e41, Apr. 2018.

[59] B. Plank, “What to do about non-standard (or non-canonical) language in NLP,” *arXiv* Aug 2016.

[60] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics – Challenges in topic discovery, data collection, and data preparation,” *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, Apr. 2018.

[61] G. D. Rosin, E. Adar, and K. Radinsky, “Learning Word Relatedness over Time,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* Jul. 2017.

[62] E. Sagi, S. Kaufmann, and B. Clark, “Semantic Density Analysis: Comparing word meaning across time and phonetic space,” *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* Mar 2009.

[63] R. Vessey, “Zappavigna, M. (2012). Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. London: Bloomsbury,” Springer, Cham, 2015, pp. 295–299.

[64] H. A. Schwartz *et al.*, “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,” *PLoS One*, vol. 8, no. 9, p. e73791, Sep. 2013.

[65] A. Morrison, “Social, Media, Life Writing,” in *Research Methodologies for Auto/biography Studies*, New York, NY: Routledge, 2019.: Routledge, 2019, pp. 41–48.

[66] D. Correa, L. A. Silva, M. Mondal, F. Benevenuto, and K. P. Gummadi, “The Many Shades of Anonymity: Characterizing Anonymous Social Media Content,” *Ninth Int. AAAI Conf. Web Soc. Media*, Apr. 2015.

[67] R. Tourangeau, B. Edwards, and T. P. Johnson, *Hard-to-survey populations*. Cambridge University Press, 2014.

[68] J. D. Livingston, T. Milne, M. L. Fang, and E. Amari, “The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review,” *Addiction*, vol. 107, no. 1, pp. 39–50, Jan. 2012.

[69] S. E. Wakeman, “Using Science to Battle Stigma in Addressing the Opioid Epidemic: Opioid Agonist Therapy Saves Lives,” *The American journal of medicine*, vol. 129, no. 5, May 2016.

[70] C. Harrison *et al.*, “Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City, 2012–2013,” *Morb. Mortal. Wkly. Rep. CDC*, vol. 63, no. 20, pp. 441–445, 2014.

[71] B. M. Kuehn, “Agencies use social media to track foodborne illness.,” *JAMA*, vol. 312, no. 2, pp. 117–8, Jul. 2014.

[72] E. O. Nsoesie, S. A. Kluberg, and J. S. Brownstein, “Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports.,” *Prev. Med. (Baltim).*, vol. 67, pp. 264–9, Oct. 2014.

[73] J. K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, and J. Bhatt, “Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014,” *Morb. Mortal. Wkly. Rep. CDC*, vol. 63, no. 32, pp. 681–685, 2014.

[74] B. Chapman, B. Raymond, and D. Powell, “Potential of social media as a tool to combat foodborne illness.,” *Perspect. Public Health*, vol. 134, no. 4, pp. 225–30, Jul. 2014.

[75] T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis, “Multiview Deep Learning for Predicting Twitter Users’ Location,”, *arXiv*, Dec. 2017.

[76] J.-P. Allem, E. Ferrara, S. P. Uppu, T. B. Cruz, and J. B. Unger, “E-Cigarette Surveillance With Social Media Data: Social Bots, Emerging Topics, and Trends.,” *JMIR public Heal. Surveill.*, vol. 3, no. 4, p. e98, Dec. 2017.

[77] R. Cherian, M. Westbrook, D. Ramo, and U. Sarkar, “Representations of Codeine Misuse on Instagram: Content Analysis.,” *JMIR public Heal. Surveill.*, vol. 4, no. 1, p. e22, Mar. 2018.

[78] S. Pandrekar *et al.*, “Social Media Based Analysis of Opioid Epidemic Using Reddit,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2018, pp. 867–876, 2018.

3 Archetype-based Modeling and Search

A version of this paper, titled *Archetype-based Modeling and Search of Social Media*, was accepted after peer review and published by MDPI in their journal Big Data and Cognitive Computing. It was published in Issue 3 of Volume 3 in 2019, with authors Brent Douglas Davis, Kamran Sedig, and Daniel J. Lizotte

This paper presents a new algorithm that allows for the search of a document collection by creating representations of documents from representations of words and using vector measures of similarity. This allows documents to be queried without the usage of keywords, a central theme which later chapters continue to build on.

3.1 Introduction

Searching for information within different collections of documents is important in many domains and contexts. Diverse people need to find documents that are of interest to them, including social scientists, health experts, and legal experts, to name but a few. Digital text-based data is being created at such a rate that it is well-nigh impossible to keep abreast of new content. This is particularly true of text being generated through social media. An increasing number of people conduct discussions and post content on social media. Words, phrases, and their nuances in social media are ever-evolving and can quickly seem unfamiliar to those who are not actively engaged in such environments. Searching for information within this ever-growing corpus has become important, but poses new challenges. A study of the difficulties encountered by researchers cites the size of the data available, the rate at which new data is generated, and the complications of unstructured textual data as common problems [1].

A recent study of social media activity in Canada found that 94% of online adults had signed up for at least one social media site [2]. Social media users are increasingly representative of general populations in many countries and can offer an avenue for searching for data representative of these populations. Searching for, studying, monitoring, and understanding relevant discourse on social media is an emerging frontier for population-level informatics in different areas—an example being public health [3]. Understanding populations based on content they create online, monitoring at-risk groups based on online activity, searching for prototypical individuals represented through textual models, and reaching out to people on social media have been tied to measurable real-world change. For instance, there has been increased interest in using social media for activities such as pharmacovigilance—the prevention and intervention of adverse reactions to drugs [4]. One of the complications of performing pharmacovigilance in social media is that stigmatized populations engaging in illicit substance use face risks ranging from social repercussions to imprisonment.

One of the features that many social media sites offer is anonymity. Members of stigmatized communities who use these sites find support through anonymous discussion of their activities and problems. Communities for stigmatized topics can be found on a variety of social media, but are often more active on anonymous platforms. Anonymity allows for the existence of communities that deal with topics which can be regarded as illegal in some countries, such as discussion of opioid drugs. There is a multiplicity of such communities on social media. The combination of anonymity, along with the aforementioned shifting and evolving of linguistic vocabulary, makes it very challenging to devise techniques that facilitate searching for text-based documents that represent at-risk individuals and populations.

Reddit, a popular social media site, is composed of anonymous, ever-evolving, linguistically difficult-to-analyze online communities of users who share a common interest. Since the site’s creation in 2005, Reddit has rapidly increased in size; users have generated over a billion posts and comments in a single month as recently as 2018. These posts contain unstructured text which is valuable to researchers wanting to understand

different phenomena on social media. Online communities exist on Reddit as "subreddits," where users can post material and interact with one another. These subreddits have a label that describes each community's focus. Examples include /r/Happy, /r/CasualConversation, and /r/pics, where online participants discuss happiness, facilitate casual conversation, and share photography, respectively. For researchers studying stigmatized topics, there is population-level data in communities such as /r/SuicideWatch, /r/Depression, and /r/Anxiety. Content from these online communities have been used to prototype systems that detect comments containing suicidal ideation [5]. Studying these communities (e.g., Suicide Watch) provides insight into the language surrounding stigmatized topics.

The scale of Reddit, however, makes manual review of all relevant content and users impossible. Data from Reddit possesses all "Six Vs" of Big Data [6]: Value, Volume, Velocity, Validity, Veracity, and Variability. The data on Reddit is continually generated by a mixture of real-world users and automated bot accounts, possesses specialized information on a multitude of topics, and contains shifting language [7] from users of varying educational backgrounds and demographics. Anonymity makes it difficult to tell whether a user's posts are genuine, further complicating analysis. Still, these difficulties are worth mitigating due to the rarity and value of the data that can be retrieved from Reddit, such as dialogue from online communities focused on topics relevant to research.

Given the existence of many communities on social media that use their own linguistic jargon, we are interested in modeling, understanding, and searching for social media users who employ population-specific, or population-enriched, language. By necessity, this modeling and searching involves big data. To this end, in this paper we present a new technique for the search of big data in social media for discovering users based on population-specific vocabulary. We call this new technique 'Archetype-Based Modeling and Search' (henceforth referred to as ABMS).

We demonstrate this technique's effectiveness firstly by modeling the vocabulary and discourse of an existing online community—namely, the subreddit /r/Opiates—and secondly, by searching for additional individuals who demonstrate an affinity or similarity to those in /r/Opiates to better understand the discourse around opiates in other online communities. The subreddit /r/Opiates contains a case study of the vocabulary used to discuss opioids and of the current discourse among members of that community. By using a combination of natural language processing and machine learning, we both model the within-community discourse and use it to identify other discourse elsewhere on Reddit that is most similar to the within-community discourse. This, in turn, supports a richer understanding of opioid discussion on social media. Our main contributions are as follows:

- We establish ABMS for retrieving documents of interest in domains where the language of discourse is not well-understood. This is accomplished by using tailored representations of the language of discourse and by searching using archetypes rather than keywords.
- We provide a concrete example of how ABMS can be applied to big data in social media in order to retrieve authors of interest.
- We explain how the ABMS technique may be extended by incorporating emerging AI methodologies, and we discuss its generalizability to additional domains.

3.2 Background

Search for information can be decomposed into two types: Lookup and exploratory [8]. Lookup searches involve the retrieval of documents identified by known keywords, such as finding all the posts that contain the word ‘oxycodone’. Often, lookup searches have an identifiable, concrete result, because the goal is specific recall of prior information. Exploratory searches, on the other hand, are open-ended, have more imprecise results, and require considerable time and effort to extract meaningful information from said results. A relevant exploratory search example would be finding all Reddit posts related to opioids.

Our ABMS technique supports exploratory search activities by using archetypes rather than keywords to identify and retrieve useful information. We borrow the word archetype from its definition [9] as “The original pattern or model from which copies are made; a prototype.” This usage is distinct from Archetypal Analysis [10], which represents multivariate data as a convex set of extreme points. Archetypal Analysis has seen use in representation learning [11] and data mining [12], but is distinct from our usage of the term archetype. By calling the technique ABMS, we intend to convey the idea that we are searching for documents with high similarity or affinity to specified collections of prototypical forms.

To capture semantic information within documents and, in turn, assess similarity of a document to the archetypes, ABMS uses word embeddings, such as those produced by GloVe [13] or word2vec [14], as a foundation to capture semantic similarity between words. Such embeddings map words to points in a vector space. This is done such that the points are nearby if the corresponding words are semantically similar, where this similarity is learned from a language corpus. An advantage of using embeddings in the manner we do is that the search technique can be tailored to specific domains by using word embeddings that are trained using domain-specific language. For example, ABMS can use word embeddings from multiple languages and from different sources specialized in understanding domain vocabulary. Existing efforts to create repositories for word embeddings [15] support the performance and generality of ABMS by providing a menu of “pre-made” embeddings useful in a variety of domains. Based on its words [16], these word embeddings are then used with a machine learning technique that learns a vector representation for each document in the collection. Having defined vector-based

representations for each of the documents, we may develop a machine learning classification model [17] in order to capture the patterns that distinguish the archetypes from "control" documents that are known not to be of interest. This model can then be applied to other documents to identify those that are most likely to be of interest.

The archetype-based approach and specialized representations used by ABMS are particularly useful in the context of social media because of their properties of discourse, which contains language that is enriched in usage by a population, including vocabulary, slang, and memetics that are unique to it. Language analysis among social media users shows that different groups use different vocabularies to communicate [18]. These linguistic differences reflect traits of the individuals in these groups. However, there is an ongoing evolution of memes and slang in social media, causing vocabulary and meanings to change [7].

For instance, the nature of pharmaceutical drug chatter on social media has been documented to have changed over time [19]. The variability of words, the evolution of their meanings, and the difficulty of maintaining knowledge of their specifics is referred to as the 'unknown vocabulary problem' [20]. Keyword search techniques are hampered by the unknown vocabulary problem because they rely on users to have a priori and in-depth knowledge of the discourse within communities and how that discourse evolves. Making sense of, and adapting to, the evolving language of social media in order to use keyword-based search approaches requires considerable time and effort on the part of the user that is often not practical to expend.

Although these challenges make keyword search impractical, it is often possible to identify some users of interest manually to serve as archetypes for ABMS—for example, by using knowledge about the topics discussed in different subreddits. Because ABMS uses domain-specialized representations for documents, these challenging aspects of language are captured and used to inform search. We now describe ABMS in general terms and present a case study of modeling and searching social media for opioid dialogue.

3.3 Archetype-Based Modeling and Search

ABMS is an exploratory modeling and search technique that solicits documents, rather than keywords, from a user in order to retrieve additional documents of interest. These solicited documents consist of archetypes, which are themselves documents of interest, and controls, which are documents not of interest. ABMS is especially useful when the specific vocabulary and patterns of discourse that distinguish archetypes from controls are unknown to the user. The key technology underlying ABMS is the development of representations for words and documents that capture these unknowns in a way that enables the retrieval of additional relevant documents. These representations map from the space of words or documents to the space of vectors with a fixed dimension, allowing for the learning of models that distinguish archetypes from controls. Once constructed, these models may be applied to any document, including those that are

neither archetypes nor controls, thereby enabling retrieval of new relevant documents. The main steps of ABMS are as follows:

1. Develop or identify a word representation for the vocabulary of the archetypes.
2. Create a document representation for each archetype and control using the word representation.
3. Construct a classification model that distinguishes archetype document representations from control document representations.
4. Apply the model to the document representations in the search corpus to identify new documents that have the strongest evidence for being similar to the archetypes.

Each of these steps contributes to the ability of the ABMS technique to facilitate exploratory search in challenging settings. Step 1 captures the specialized vocabulary of the documents under study so that end users are not required to have a deep knowledge thereof. Step 2 implicitly captures structure within the documents beyond the presence/absence of keywords, e.g., it captures word co-occurrence information, so that 'cat' and 'feline' occur in more similar contexts than 'cat' and 'dog'. Step 3 creates a way of assessing the affinity of a document to the archetypes as opposed to the controls, and Step 4 extends this assessment to a new set of documents.

We describe an algorithm for our ABMS technique below.

Algorithm 1 Archetype-Based Modeling and Search

- Input *D*, a set of documents, which contains the following subsets: *A*, a subset of archetypes (documents of interest), *C* a subset of controls (documents not of interest), and *U*, an unlabeled subset of *D*, which will be searched to identify additional documents of interest.
- Use the words in all documents in *A* and develop a word representation W that maps words to vectors.
- Use W and develop a document representation, V, that maps all documents in *D* to vectors.
- Train a classifier to distinguish V(*A*) from V(*C*). The classifier must be able to rank inputs according to their likelihood of belonging to *A* versus *C*.
- Apply the classifier to V(*U*) and rank the unknown documents.

Return as top-ranked documents those most likely to be of interest.

- Algorithmic Complexity: ABMS scales dependent on five variables: (1) The number of unique words in the vocabulary; (2) the number of words in each document; (3) the number of documents to be learned; (4) the combined number of archetypes and controls; and (5) the representation size for the word and document representation. The number of documents and the number of words in each document scale linearly for the number of representations to be learned. The larger the vocabulary size, the more computation is required to form each representation. For many representation techniques, this scaling will not be linear. Classifying the archetypes against the controls will scale dependent on the classifier that is chosen. The representation size is a parameter that scales to increase representation learning time and classifying time.

Figure 5 depicts the ABMS algorithm above, showing the extraction of the archetypes and controls from the document representation corpus and finishing with a ranked list of search results. The user identifies a collection of documents that are of interest (archetypes) and a collection of documents which are not of interest (controls).

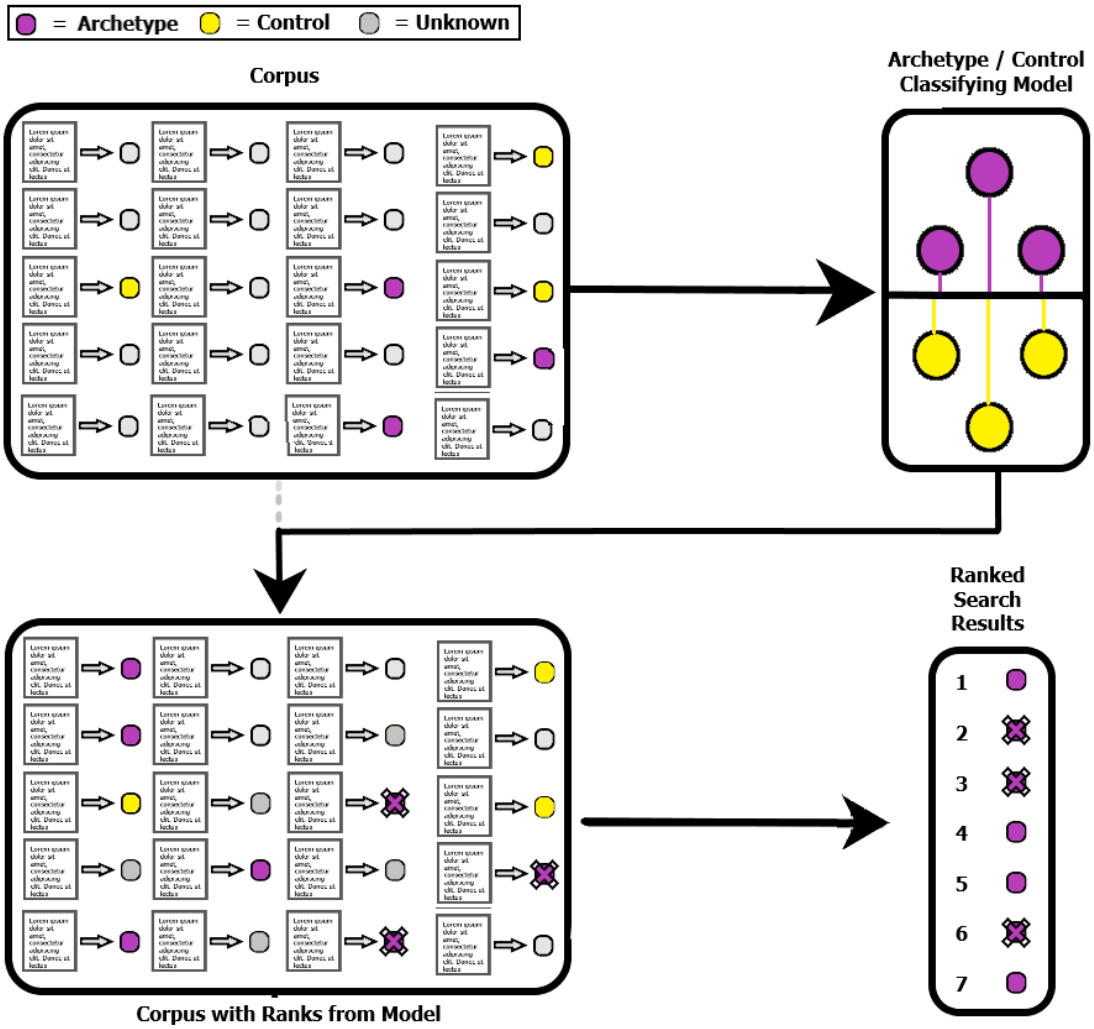


Figure 5. This overview shows the process behind an Archetype-Based Modeling and Search performed on a collection of documents. A word representation is learned, and then using the word representation each document is assigned a vector-based representation. A classification model is learned to distinguish representations of archetypes from representations of controls. The remaining unlabeled documents are then ranked by the model. The resulting scores are sorted and form a ranking for the search results.

ABMS as a technique is independent of the specific methods chosen for developing the word representations, document representations, and classification model. In the following section, we present a case study that illustrates how ABMS can be used with representation learning techniques to retrieve documents from social media.

3.4 Case Study: Opioid Dialogue on Social Media

We now present a case study that aims to retrieve a collection of social media authors whose discourse suggests an affinity to authors who discuss opioid use. In this case study, each "document" consists of collected posts from a particular Reddit author. Archetypes consist of the entire posting history of every Reddit author who has posted in /r/Opiates in 2017, and controls consist of posts by reddit authors in /r/CasualConversation in 2017, which is a moderated subreddit that forbids controversial topics like illicit drug use. For our search corpus, we use the posting histories of authors who posted in /r/Ottawa in 2017. Using these archetypes, controls, and search corpus, we performed the four steps of ABMS as follows.

3.4.1 Case Study Methods

Step 1: Develop or identify a word representation for the vocabulary of the archetypes.

To create our model, we construct a word embedding tailored to our community of interest that includes jargon and specialized language. A word embedding is created given a corpus of documents and a specified dimension *d*. The resulting embedding maps each vocabulary word to a *d*-dimensional vector, where pairs of semantically similar words are placed closer together in vector space than pairs of less similar words [17]. To construct our embedding for the case study, we collected all posts from 2015 through 2018 from the Reddit archives at pushshift.io made on /r/Opiates, and extracted the text. We lower-cased the words but did not perform any other common modifications such as stemming or lemmatizing. This was done to preserve words and phrases that may possess unique domain meanings. We then applied the text2vec [21] R package to these documents to train a new word embedding using the GloVe method.

Step 2: Create a document representation for each archetype and control using the word representation.

To construct a representation for each archetype and control, we first retrieved authors who, in 2017, posted in /r/Opiates (archetypes) or in /r/CasualConversation (controls). Authors who posted in both were considered archetypes. For each archetype, we created a document consisting of the concatenation of their entire posting history on Reddit. For controls, we restricted to posts within /r/CasualConversation. We excluded any authors who had post counts of more than 1500 to filter out accounts that were likely automated, and we enforced a minimum of 25 words for an author to be included. Posts were overwhelmingly in English; other languages were not explicitly excluded. We cleaned the text of any non-alphanumeric characters. There may be Reddit data that are missing from our sample [22], but ABMS will still search the data we have available.

We constructed a model of each archetype and each control in the dataset which maps their concatenated posts to a *d*-dimensional vector. To construct the vectors, we used a neural network architecture originally established for understanding sarcasm in a body of text by Amir et al. [23]. This architecture has also been applied to author representations that were later used to detect post-traumatic stress disorder (PTSD) and depression [16]. The method works as follows: For each author, we draw pseudo-random words from the vocabulary in the pre-trained word embedding, and we train the neural network to distinguish these artificial negative example words from the true words used by the author. The resulting network forms that author's representation.

This approach to constructing models for archetypes and controls requires a word representation like the one constructed in Step 1 in order to train the neural network models, because the neural networks take vectors as inputs. To investigate the effect of using different word embeddings on the resulting author representations, we constructed two such models, one based on our new embedding tailored to /r/Opiates, and one based on the existing Stanford Twitter GloVe embedding [13]. The GloVe twitter model has a vocabulary of 1.2 million, while our /r/Opiates embedding has a vocabulary of 77,036. We capped the vocabulary within the Twitter embedding at the 20,000 most common words in our data, and we used the full vocabulary for the /r/Opiates embedding to ensure that all jargon was captured. Models derived from the Twitter embedding and the /r/Opiates embedding have dimensionality 100 (chosen by the Stanford team) and 300 (typical default setting for GloVe), respectively.

Step 3: Construct a classification model that distinguishes archetype document representations from control document representations.

Having created the models for our archetypes and controls, we then trained classification models to distinguish them. We trained support vector machine (SVM) classifiers using e1071 in R [24] using both linear and radial basis function (RBF) kernels. We configured all SVMs to output the decision value for each prediction. We trained and tuned our SVMs with e1071, using costs of 0.1, 1, and 10 for both models and gammas of 0.5, 1, and 2 for the RBF.

Step 4: Apply the model to the document representations in the search corpus to identify new documents that have the strongest evidence for being similar to the archetypes.

For this case study, we searched through authors who posted in /r/Ottawa, and retrieved those with the highest affinity to the archetypes. To do so, we first used the same document representation approach from Steps 1 and 2 to produce representations for all of the /r/Ottawa authors. Next, we took the resulting classifier from Step 3, and we applied it to all of the authors, obtaining a 'decision value' for each one. SVMs output a decision value for any input vector; it gives the orthogonal distance from that vector to the separating hyperplane. Positive decision values indicate the SVM assigns a positive label to the vector, and negative decision values indicate a negative label. The magnitude of the decision value is a measure of the confidence of the SVM in its decision. We used the decision values output by the SVM for each /r/Ottawa author to rank them in terms of affinity to the archetypes—i.e., affinity to authors who post in /r/Opiates.

3.4.2 Case Study Results

In the following, we present the results of our case study. We begin by assessing the ability of our classifying model to distinguish archetypes from controls. We then consider the impact of the representations contained in our word embeddings on the properties of the resulting models. Finally, we describe the results of using the classifier to retrieve additional authors from /r/Ottawa who appear to have a high affinity to authors in /r/Opiates. Throughout, we identify lessons learned that may be helpful for future applications of ABMS.

3.4.2.1 Modeling and Classification

In order for ABMS to successfully retrieve new documents similar to the archetypes, the classifier learned in Step 3 must be able to distinguish archetypes from controls. We used the area under the Receiver Operating Characteristic (ROC) curve (AUC), as well as precision and recall rates, to evaluate the success with which our SVM classifiers were able to accomplish this. To estimate these performance indicators, we split our examples into a train/test set containing approximately 30,000 cases for training and 6000 cases for testing. The training set is composed of 13,000 /r/CasualConversation authors and 17,000 /r/Opiates authors. The testing set is composed of 2000 /r/CasualConversation authors and 4000 /r/Opiates authors. Results are shown in **Table 1**. While radial kernel SVMs can effectively separate the training data, indicated by high training set (resubstitution) AUC, it appears they drastically overfit, indicated by much lower test set AUC. This is particularly evident in the case of using the RBF kernel with the /r/Opiates embedding. Linear SVMs offer better explanation, lower computation cost, and better generalization performance, making them an obvious choice for the remainder of the case study.

It is notable and perhaps surprising that using the /r/Opiates embedding, which is tailored to our task, only caused a slight improvement in classification performance. This suggests that relevant language and useful word vectors for them exist in both embeddings. In the absence of datasets large enough to train a specialized embedding, we suggest that large pre-trained models from social media text may be able to perform reasonably when used with ABMS, which is a benefit if learning new embeddings is not feasible. However, this still induces a risk of missing some specialized vocabulary and can result in non-intuitive behavior, as we illustrate below. The relatively reduced recall rates may be explained by our choice of archetypes. An author who posted once in /r/Opiates and never participated again would still be captured and labeled as originating from /r/Opiates, even though their discourse on Reddit may be predominately unrelated to archetypal discussion. We therefore suggest that it may be important in some applications to use a more stringent definitions for archetypes.

Table 1. Support vector machine (SVM) performance on classifying authors as originating from the subreddits /r/Opiates or /r/CasualConversation. Models were trained on 30,000 authors and tested on 6000 authors, giving a margin of error of ±0.013 for these estimates at the 95% confidence level.

Embedding Source	Metric Used	Linear Kernel Training	Linear Kernel Test	Radial Kernel Training	Radial Kernel Test
Pre-trained	AUC	0.780	0.780	0.883	0.722
Twitter	Precision	0.985	0.908	0.991	0.968
Embedding	Recall	0.795	0.879	0.834	0.882
	AUC	0.805	0.783	0.888	0.624
/r/Opiates	Precision	0.985	0.978	0.967	0.931
	Recall	0.821	0.895	0.878	0.876

Although, using the two embeddings, classification performance is similar, the actual classification models themselves are quite different in terms of the author attributes they use to make decisions, as we will show. This in turn means that the authors with the strongest affinity with the archetypes, as measured by the decision values, are quite different from model-to-model. We illustrate this by first examining the vocabulary on which the two models rely most heavily, and then by examining which authors are assigned the highest affinity to the archetype class.

By taking the decision direction—the vector orthogonal to the separating hyperplane found by the SVM—and comparing it to the original word vectors from the embedding that was used to construct the document representations and classification model, we can see which words have a vector that is most similar to the decision direction. These words are the ones that exert the most influence on the decision value in the positive direction, meaning that their presence in a document increases that document's affinity with the archetypes. We visualize the 200 words most aligned to the decision direction for both the Twitter and /r/Opiates embeddings in **Figures 6 and 7** respectively**.**

The difference in content between the Twitter embedding and the /r/Opiates embedding can be seen in the diversity of vocabulary in the two figures. While there are words related to /r/Opiates participation in the top words from the Twitter embedding—fentanyl, painkillers, codeine, hydrocodone—many words are nonsensical. This is not the case in the words sourced from the /r/Opiates embedding, which are much more frequently linked to the vocabulary from the topic in which we are interested. Notably, many match our conceptions of words that an author from /r/Opiates might use, suggesting the model is effective in capturing aspects of population-specific vocabulary that are linked to the online community of origin. The clarity offered here is a compelling reason to use locally trained embeddings where possible.

Figure 6. A word cloud generated from the most aligned words to the decision direction separating author vectors from /r/Opiates and /r/CasualConversation. This was generated using the restricted Stanford Twitter GloVe embedding with a 20 thousand size vocabulary.

Figure 7. A word cloud generated from the most aligned words to the decision direction separating author vectors from /r/Opiates and /r/CasualConversation. This was generated using the GloVe embedding based on /r/Opiates authors, with a 78 thousand size vocabulary.

The vocabulary associated with the decision direction also offers insights into the discourse of /r/Opiates authors. While ‘opioid’ makes an appearance in the top 200, numerous other drugs are present. A slang form of fentanyl, ‘fent’, can be observed. Cocaine makes an appearance, as does impulsivity. ‘Slinging’ makes an appearance, which can be a term used to describe the selling of drugs. Some words defy these intuitive explanations, such as ‘bookcase’ and ‘germaphobe’, but their discovery here allows keyword-base retrieval of these posts so that they can be examined further to determine why these words appear to be important.

We now consider how these observed differences in the models translate into differences in which authors are assigned highest affinity. The decision values for each author assigned by the linear SVMs are shown in **Figure 8** for the Twitter embedding and **Figure 9** for the /r/Opiates embedding. By comparing the peaks in **Figures 8 and 9**, we can see that the location of the peaks in decision value, which correspond to the authors with strongest affinity, vary depending on which embedding is used. This is important because it implies that the choice of embedding could have a significant impact on which documents are retrieved by the search process.

Examining authors that have extreme decision values can be useful for understanding misclassifications as well. For example, the two most extreme decision values from the /r/Opiates embedding correspond to authors who frequently post about politics as well as drug use, and the most extremely misclassified /r/CasualConversation author in the test set has the username (partially obscured) ‘XXXX-XX-MAGA’. The ‘maga’ portion is in reference to a political slogan and their posts match this. If political discourse among archetypes is highly prevalent, the classifier may use this as a ‘cue’ to classify them. This in turn may lead to false positives among authors who also post political discussions, but who should not be considered archetypes. This again illustrates the need for stringent archetype definition. In any case, examining the most aligned vocabulary provides insights into what kinds of words and dialogue are being used to classify authors and can provide clues for refining the models.

3.4.2.2 Exploratory Search Results

Both classification models were used to produce a ranking of /r/Ottawa authors in terms of their affinity to the archetypes. This ranking provides a starting point for exploratory search, in which a user would review highly-ranked documents to gain insights into not only what documents in the search corpus have strong affinity to the archetypes, but also what discourse appears to drive that affinity. We investigate the top five documents (authors) from the /r/Ottawa set, as defined by the Stanford Twitter word embedding and by our new Opiates word embedding, and discuss our findings. We have not reproduced the retrieved documents here because of privacy concerns.

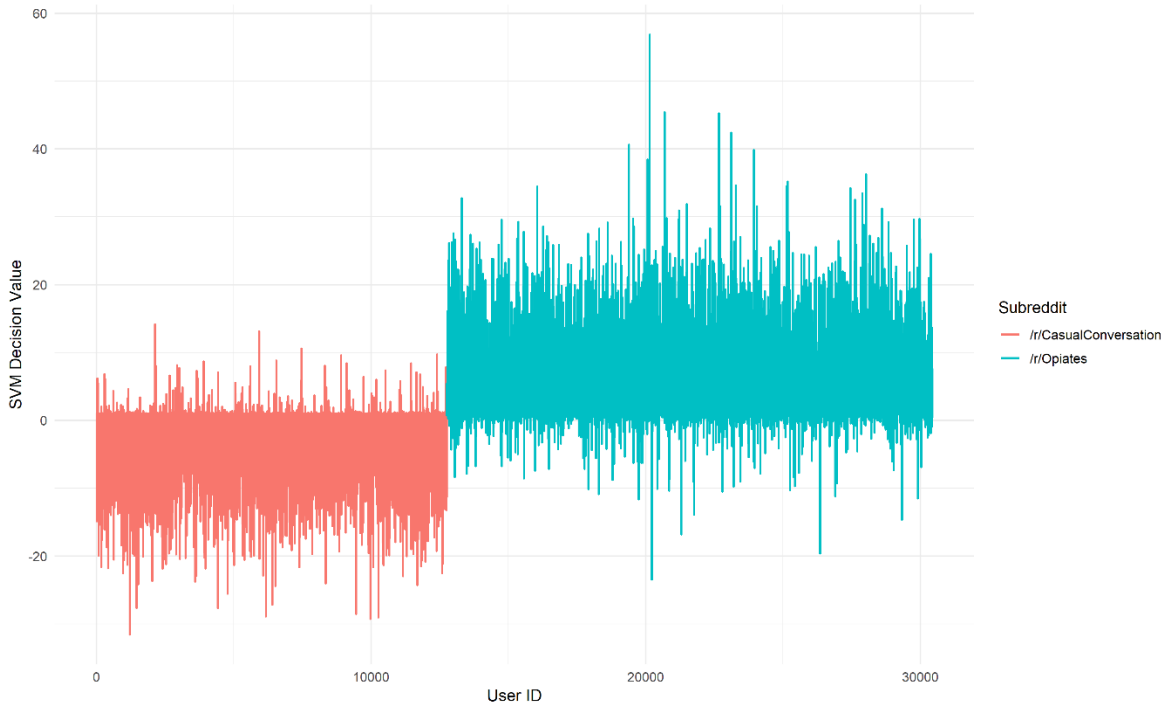


Figure 8. Decision values produced by a Linear SVM for author vectors trained with the Twitter GloVe embedding, colored by subreddit of origin. A large magnitude of decision values corresponds to a large distance from the decision boundary of the SVM. The x-axis is an arbitrary user ID assigned to an author vector from /r/CasualConversation (left) or /r/Opiates (right).

The top five documents derived from the Stanford Twitter embedding were challenging to interpret. The main themes of the first document were collectible card games and philosophy; it could be that the singular mention of "rehabilitation" led to its high score. The second document primarily discussed hockey with two mentions of "weed." The third was a bot (automated) account that may have been identified because its posts always included numbers, a feature also common in /r/Opiates posts because of reference to dosages. The fourth contained discussion on drunk driving, mentioning: Being drunk, medical marijuana, video games, politics, and philosophy. The fifth discussed sports and politics and had one post consisting of just the word 'drugs.'

Figure 9. Decision values produced by a Linear SVM for author vectors trained with the /r/Opiates GloVe embedding, colored by subreddit of origin. A large magnitude of decision values corresponds to a large distance from the decision boundary of the SVM. The x-axis is an arbitrary user ID assigned to an author vector from /r/CasualConversation (left) or /r/Opiates (right).

The top five documents derived from the new /r/Opiates embedding were somewhat easier to interpret. The first consisted entirely of discussion around buying, selling, and cryptocurrency, though not about drugs, although it is important to note that cryptocurrency is widely-used in the illicit drug market [25]. The second primarily discussed firearms control and law enforcement, as well as drugs and politics. The third primarily discussed cannabis, going as far as discussing different forms of the drug and the equipment used to consume it. The fourth described the author's ongoing use of both methamphetamine and heroin; this was the author who, from our point of view, most closely resembled an archetype of /r/Opiates and who appeared most in need of support. The fifth primarily discussed law enforcement and video games.

Our takeaway from this initial exploratory search is that the models chosen to construct the archetype and control representations can have a large impact on the resulting exploratory search. Our second takeaway is that "ancillary topics"—for example, commerce and law enforcement—can be important indicators of affinity to our archetypes. Further refinement of archetype and control definitions has the potential to reveal further insights about this, or any other, community.

3.4.3 Discussion

We have demonstrated how the ABMS technique enables exploratory searches of big data sources when relevant vocabulary is not well understood. Our case study provides insights into some of the practical aspects of deploying ABMS. In this section, we discuss some other characteristics of ABMS and considerations surrounding its use.

3.4.3.1 Scalability of ABMS

One of the characteristics of the ABMS technique is its scalability. ABMS relies on having a sufficiently large amount of data to capture vocabulary and provide archetypes; hence, it is particularly suited to analysis of big data from social media. On the other hand, the ABMS technique presented here uses a deep neural network to produce each document representation, which is computationally very costly if there are many documents.

Fortunately, this process is embarrassingly parallel; each document's representation can be computed simultaneously. Hence, distributed computing resources can be readily used to create the representations in a reasonable time, and the neural network training process can make use of graphical processing unit (GPU) resources to further speed up computation. In our case study, we used Compute Canada systems to train a total of 47,000 deep neural networks to create the necessary author models, which took approximately 18 CPU years at 2.6 GHz. If we assume the average author on Reddit posts 300 times, the approximately 5 billion posts on Reddit yield a set of 10 million authors and require a computation time of almost 6500 CPU years at 2.6 GHz.

Once the document representations have been computed, they can be re-classified using a different set of archetypes and controls, with the only computational cost being that of training the new classifier. This increases the reusability and utility of the models for other tasks, and makes iterative refinement of the archetype and control sets feasible. For example, while our case study illustrated exploratory search for authors similar to /r/Opiates authors, it is easy to extend this technique to include /r/Drugs authors alongside them to broaden the search.

3.4.3.2 Generalizability of ABMS

Another characteristic of the ABMS technique is its generalizability—that is, it can be used across different social media domains, as long as there is user content in written form. For example, ABMS could be applied to Twitter data as follows. First, all the Twitter activity that contains a hashtag of interest could be collected and used to make a word embedding. This embedding could then be used to develop author representations based on each author's tweet history. Archetypes and controls could be defined in

different ways—for example, archetypes could be all authors who used the hashtag, and controls could be a subset of those who did not, or who used another hashtag. Alternatively, to identify users with an affinity to a given geography, archetypes could be Twitter users who have a particular geotag included in their tweet metadata, and controls could be a subset who did not, or who have a different geotag.

ABMS could also be applied to Facebook. Facebook authors provide many possible labels from their activity on the site that could identify subgroups that use specialized language, and that could be used to define archetype/control status. For example, authors organize themselves into groups, like pages based on their interest, and even take personality quizzes, which are potentially shared with their friends. The textual activity within these groups could be collected to build a word embedding for representation learning. Authors that are in a group can be considered archetypes for performing ABMS. Facebook authors can provide more detailed geographic information on their profiles if they so choose, including both hometown and current residence, and some authors provide information such as cell phone numbers that contain area codes which help with identifying geography. A characteristic of ABMS is that it can use any of this information to define archetypes and controls and, in turn, to enable exploratory search.

3.4.3.3 Transferability of ABMS Models Across Social Media Platforms

Another characteristic of the ABMS technique is the transferability of the models it creates. The models learned by ABMS from one social media platform can be transferred to other social media platforms to search for relevant authors. This has the advantage of using labels that are available from one social media platform—such as the subreddits from Reddit or the hashtags from Twitter—to define archetypes and controls and learn a model which can then be applied to search for authors in other social media sites where such labels are not available.

When transferring models from a source platform to a destination platform, it is important to use the same representations for words and authors in the source and the target. If author representations in the target platform are not developed using the same techniques as for the source—that is, using the same word embedding and representation training procedure—they will not work properly with the learned model. The potentially new vocabulary used in other social media can present complications for transferring models. Vocabulary irrelevant to the user's desired search results can still occur disproportionately in one group over another.

If this should happen, however, adjusting some of the labels and adding more examples to the control set can help tune ABMS to search more effectively. One of the ways that the locally-trained embeddings are helpful is that, when using techniques like usr2vec, any new words not in the local word embedding's vocabulary are discarded. While this lowers the number of total words that can be used to construct the document representation, it can increase the specificity of the search.

3.4.3.4 Adaptability of ABMS to Different Tasks

Another characteristic of the ABMS technique is its adaptability to different tasks. Our case study focused on the task of identifying authors who use language associated with discussion of opioids. However, strongly associated authors sometimes discussed not opioids, but rather topics that are also discussed by authors who discuss opioids. While one user may find it interesting that discussion of opioids is associated with discussion of hobbies and politics, another user whose task is to find opioid vocabulary and quantify the amount of it might be less interested in these aspects of discourse. To accommodate this different task, one could apply clustering techniques to the most highly ranked documents to identify topics or subgroups that use vocabulary associated with archetypes but that may not exclusively use the vocabulary we are interested in.

For example, in our case study, clustering might reveal groups of authors or posts who have high affinity to archetypes but who post primarily about cryptocurrency, politics, video games, and non-opioid drugs. By removing these authors, we may be better able to identify vocabulary of interest. This opens the possibility for further analysis beyond the initial ranking and is suggested as a possible extension to the results of our exploratory search. By visualizing the words which are most closely aligned to the author vectors can help the user decide whether to remove authors who use task-irrelevant vocabulary. This can be done by reducing the set of archetypes, or potentially even moving archetypes to the control group. Iterating in this manner facilitates users tuning ABMS to be most sensitive to the vocabulary of their choosing.

We anticipate that better word embeddings and the fidelity they provide in capturing semantics will provide new opportunities to apply ABMS to different tasks. New techniques for producing word embeddings, such as BERT [26] and XLNet [27], continue to produce increasingly impressive benchmark scores on natural language tasks. As these models improve representations of words in vector form, it is our expectation that the effectiveness of searches using this technique will also continue to improve. As advancements are made in word and document representation learning, the ability to perform increasingly specialized search tasks will become possible.

3.4.3.5 Ethics of ABMS and Related Techniques

As online surveillance evolves, we will have to confront the question of to what extent members of the populace should and can be monitored. Word embeddings have already been applied to monitor influenza activity [28], but not to the extent of identifying individuals who have been sick. While we apply our technique to a social media site that enables anonymity, this anonymity can be compromised by other advances in deep learning. Prominently, there has been success in identifying user location from Twitter—without the use of geotags [29].

As our ability to infer information about people based on their online postings continues to improve, the role that these inferences take in society will have to be addressed. There is a multitude of other information being gathered beyond the text generated on social media. Big data's integration into the Internet of Things [30], means that the number of sensors that collect personalized data, including biometrics, is ever increasing while sending private information [31]. The security of big data from social networks and new techniques to support it [32] will only become more important and foundational as techniques like ABMS increase the potential damage that can be done by data leaks.

Here, we have presented another way that potentially identifiable information can be found, even from de-identified social media data. There is no simple answer to how to handle the available data and the implications that can be found in it, but we demonstrate that the possibility for it to be abused is real. For example, one can imagine a law enforcement system which identifies users from social media and then pulls historical biometrics from fitness wearables data to look for patterns matching those of individuals on various illicit drugs. Hence the social, governmental, and ethical implications of the use of ABMS and related social media search techniques require careful consideration.

3.4.3.6 Biases in Word Embeddings and Their Effects on ABMS

Word embeddings have their own non-technical problems. An examination of the relationships found in word embeddings has shown that they can contain sexist and other biases which can be difficult to remove [33]. As our model is ultimately using the vocabulary of these subreddits to model their associations, there are important implications to its sensitivity. Any differences in language between communities—for example, the presence of Spanish or French text in /r/Opiates but not in /r/CasualConversation—are going to be detected by the model if it gives it a reasonable boost in classification performance.

If a model derived from such data were used naively without adjusting for this kind of phenomenon, the model could become biased toward labeling any users engaging in French or Spanish dialogue as originating from /r/Opiates. This can be seen to some extent in the extreme decision values we noted in our results, where political leaning was given some weighting in the decision between classes, and extreme results were pulled to extremes so far from the decision boundary that their other dialogue made little contribution.

Judicious selection, or careful training, of word embeddings can help these problems but are unlikely to remove them entirely. Researchers looking to apply word embeddings and ABMS in sensitive or clinical environments are advised to thoroughly evaluate them for biases. The task considered here, searching, is more likely to demonstrate these biases in search results than it is to act on them in a way that is directly harmful. Should ABMS be adapted to other applications, steps must be taken to ensure that the word embeddings and archetypal documents/authors do not propagate these inherent biases.

3.5 Conclusions and Future Work

We present Archetype-Based Modeling and Search, a technique for exploratory search of large corpora when keywords are not well-understood because of highly complex, evolving discourse with population-specific vocabulary. We demonstrate from our case study that this technique can model complex behaviors present in the author's representations. By performing ABMS, we can rank new authors based on their similarities to these archetypal authors. We anticipate that the ABMS technique will find increased use as the volume and relevance of social media data increases further, providing an ever-widening window into text-rich human activity.

Since its inception, social media has grown at an incredible rate. ABMS has numerous applications for analyzing the content of social media, which is too complex in its discourse and too large for manual review. The text produced through social media continues to provide insights into the populations that generate them. As such data continues to become available, researchers are going to be able to perform analyses on human behavior at scales never previously achieved. Techniques such as ABMS are an effort towards being able to sort through data at population scale in an intelligent, assisted manner.

We intend to apply this technique to other tasks beyond identifying authors who talk about opioids. We suspect this technique offers opportunities to assess the size of populations on social media without having to compromise their anonymity or solicit their engagement. We have begun analyzing one such population, opioid abuse sufferers, here. By identifying authors with a condition such as an opioid addiction, or a mental health disorder, there are opportunities to research social co-occurrences, investigate comorbidities within populations, and understand social challenges and needs.

Avenues for improving ABMS originate from the sourcing of labels for the archetypes and controls, and from the chosen classification model. Our model is generated with an SVM; however, there may be models that are better suited for measuring the differences between archetypes and controls, depending on the task that the ABMS results are supporting. The classification is highly dependent on the representations learned. We have used an existing machine learning architecture to form an author representation from word representations, but other representation techniques may be able to better represent relevant qualities for the classifier to use to distinguish archetypes from controls.

3.6 ABMS Bibliography

1. Stieglitz, S.; Mirbabaie, M.; Ross, B.; Neuberger, C. Social media analytics–Challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manag.* **2018**, *39*, 156–168.
2. Gruzd, A.; Jacobson, J.; Mai, P.; Dubois, E. The State of Social Media in Canada 2017. *SSRN Electron. J.* **2018**. doi:10.5683/SP/AL8Z6R.
3. Wang, Y.C.; DeSalvo, K. Timely, Granular, and Actionable: Informatics in the Public Health 3.0 Era. *Am. J. Public Health* **2018**, *108*, 930–934.
4. Sarker, A.; Ginn, R.; Nikfarjam, A.; O'Connor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; Gonzalez, G. Utilizing social media data for pharmacovigilance: A review. *J. Biomed. Inform.* **2015**, *54*, 202–212.
5. Aladağ, A.E.; Muderrisoglu, S.; Akbas, N.B.; Zahmacioglu, O.; Bingol, H.O. Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *J. Med. Internet Res.* **2018**, *20*, e215.
6. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.Z. Big Data for Health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1193–1208.
7. Rosin, G.D.; Adar, E.; Radinsky, K. Learning Word Relatedness over Time. *arXiv* **2017**, arXiv:1707.08081.
8. Marchionini, G. Exploratory search: From finding to understanding. *Commun. ACM* **2006**, *49*, 41–46.
9. "archetype, n."|OED Online; Oxford University Press: Oxford, UK, 2019. Accessed on 14 June 2019
10. Cutler, A.; Breiman, L. Archetypal Analysis. *Technometrics* **1994**, *36*, 338–347.
11. Chen, Y.; Mairal, J.; Harchaoui, Z. Fast and Robust Archetypal Analysis for Representation Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
12. Mørup, M.; Hansen, L.K. Archetypal analysis for machine learning and data mining. *Neurocomputing* **2012**, *80*, 54–63.
13. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing, Doha, Qatar, 25–29 October 2014.
14. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*; Mit Press: Cambridge, MA, USA, 2013.
15. Fares, M.; Kutuzov, A.; Oepen, S.; Velldal, E. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, Gothenburg, Sweden, 22-24 May 2017.
16. AMIR, S.; Coppersmith, G.; Carvalho, P.; Silva, M.J.; Wallace, B.C. Quantifying Mental Health from Social Media with Neural User Embeddings. *arXiv* **2017**, arXiv:1705.00335.
17. Camacho-Collados, J.; Pilehvar, M.T. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *J. Artif. Intell. Res.* **2018**, *63*, 743–788.

18. Vessey, R.; Zappavigna, M. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*; Springer: London, UK, 2015.
19. Wiley, M.T.; Jin, C.; Hristidis, V.; Esterling, K.M. Pharmaceutical drugs chatter on Online Social Networks. *J. Biomed. Inform.* **2014**, *49*, 245–254.
20. Furnas, G.W.; Landauer, T.K.; Gomez, L.M.; Dumais, S.T. The vocabulary problem in human-system communication. *Commun. ACM* **1987**, *30*, 964–971.
21. Selivanov, D.; Wang, Q. text2vec: Modern Text Mining Framework for R. Computer Software Manual(R Package Version 0.4. 0). Available online: https://CRAN. R-project. org/package= text2vec (accessed on 14 June 2019).
22. Gaffney, D.; Matias, J.N. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLoS ONE* **2018**, *13*, e0200162.
23. AMIR, S.; Wallace, B.C.; Lyu, H.; Carvalho, P.; Silva, M.J. Modelling Context with User Embeddings for Sarcasm Detection in Social Media. *arXiv* **2016**, arXiv:1607.00976.
24. Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Maintainer, A.W. *The e1071 Package*; Misc Functions of Department of Statistics: Vienna, Austria, 2005.
25. Foley, S.; Karlsen, J.R.; Putniņš, T.J. Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies? *Rev. Financ. Stud.* **2019**, *32*, 1798–1853.
26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
27. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* 2019, arXiv:1906.08237.
28. Dai, X.; Bikdash, M.; Meyer, B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon*; IEEE: Piscataway, NJ, USA, 2017.
29. Do, T.H.; Nguyen, D.M.; Tsiligianni, E.; Cornelis, B.; Deligiannis, N. Multiview Deep Learning for Predicting Twitter Users' Location. *arXiv* **2017**, arXiv:1712.08091.
30. Ge, M.; Bangui, H.; Buhnova, B. Big Data for Internet of Things: A Survey. *Future Gener. Comput. Syst.* **2018**, *87*, 601–614.
31. Rui, Z.; Yan, Z. A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification. *IEEE Access* **2019**, *7*, 5994–6009.
32. Tariq, N.; Asim, M.; Al-Obeidat, F.; Zubair Farooqi, M.; Baker, T.; Hammoudeh, M.; Ghafir, I. The Security of Big Data in Fog-Enabled IoT Applications Including Blockchain: A Survey. *Sensors* **2019**, *19*, 1788.
33. Gonen, H.; Goldberg, Y. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv* **2019**, arXiv:1903.03862.

4 Archetype-based Information Retrieval

4.1 Introduction

Information retrieval requires the formation of a query. Without something to search for, the entire process of retrieving information is moot. When we are performing information retrieval for previously known information, we can adjust our query to match some part of a document, file, or other object we are retrieving. Searches of this nature have definite results. If a file system were missing a record that we previously knew had ‘cat’ in the title, and we were searching for all records with ‘cat’ in the title, we would know that the file system had lost information, or the retrieval process was faulty.

This contrasts against searches with indefinite results, or ones that can only have definite results by having an expert review all documents. By introducing some semantics, such as saying that we now wish to retrieve all the documents that are about cats, a host of new problems are introduced. For example, a document that exclusively refers to cats as felines will not be found by the query ‘cat’. To mitigate problems with query formation where vocabulary is unknown, we propose the usage of a novel information retrieval process called *Archetype-based Information Retrieval* (AIR).

The purpose of AIR is to enhance the capabilities of an exploratory search by reducing the cognitive load, time, and effort needed to mitigate the unknown vocabulary problem. This is done by using advances in natural language processing and multiple sub-domains of machine learning with the end goal of augmenting human intelligence. In a traditional search setting with unknown results, the information seeker has a complex cognitive task – not only are they attempting to recall all words which could be associative by their own understanding, they are also attempting to formulate all words that all other persons could use to describe the focus of their search. From this standpoint, we consider the task that public health stakeholders encounter when trying to assess complex populations.

Public health stakeholders (PHS) are tasked with a difficult problem: trying to estimate the needs and number of at-risk people in the communities they oversee. These at-risk people often have a stigma that is related to their at-risk status. These negative beliefs and attitudes make assessing their needs with traditional means such as surveys, difficult [1]. Stigma, which leads to negative attitudes and beliefs about those stigmatized, exist in examples such as illicit drug usage. Opioid drug users, then, are an at-risk group of people facing stigma. We refer to this group of people collectively as an at-risk subpopulation. To assess these needs without surveys, we turn to other sources. Previous research has shown that discussion of opioid drug usage appears online on the social media site Reddit [2], [3]. While there is data available in these sites, it requires significant processing to support the need of PHSes. Hence, we identify some of the challenges and tasks PHSes face when using social media and how to support them.

Social media presents challenges for analysis when looking to support PHSes [4]. First, PHSes are frequently tasked with a specific geographic area over which they have jurisdiction. This presents a problem with social media – only a subset of social media is relevant for PHSes. Activity on social media sites such as Reddit is not tied to a geographic location, and authors of content on the website can interact with other authors across the globe. Thus, much information of the information that is easily findable by searching these social media sites has no obvious geographic information. While there may be unprecedented amounts of information directly from the people making up the at-risk subpopulation in the entirety of Reddit, it is not simple to narrow our lens of analysis to the needs of PHSes.

Take the modern case of the opioid epidemic, which is bringing the needs of this at-risk subpopulation to the fore [5], [6]. PHSes face a difficult task in assessing this, as use of opioid drugs when not prescribed by a doctor is illegal (in Canada) and stigmatized. Social media provides a lens into these behaviours, however. As such, there have been analysis done on the opioid epidemic as it is observed on Reddit [3] Content from Reddit is archived and supported for research by Pushshift, making this data reasonably available [2]. Pushshift is an open data effort that facilitates research of many topics, including public health. Information of this kind – textual documents directly authored by the people in the at-risk subpopulation – was previously difficulty to access. This has started new kinds of analyses of public health topics. However, this analysis is independent of geography. To understand the different kinds of analysis possible, it is necessary to discuss the structure of Reddit.

Reddit is divided into, appropriately named, subreddits. There is a multiplicity of subreddits. They have a name which is related to the kind of discussion meant to go in them. Examples include /r/CasualConversation, for conversation about topics that are not considered to be intense, /r/Opiates, for discussion of opioid drugs and their use, and /r/Ottawa, for discussion of things going on in the city of Ottawa. The dialogue in each subreddit will be enriched in different kinds of information. For example, /r/Opiates is enriched in dialogue about the use and addiction to opioid drugs. Also, /r/Ottawa provides a way to estimate the authors from Ottawa. The knowledge gap that occurs is trying to assess the activity of authors from Ottawa that is about opioids but connecting the two knowledge bases is not a simple task for PHSes.

One approach is to get the user lists from each subreddit and to combine them together. This produces a list of some authors on Reddit that are active in both, but this does not capture all the activity that is occurring. This is because many authors on /r/Opiates use accounts that are solely dedicated to their discussion of opioids or other illicit topics. Similarly, just because has never posted in /r/Opiates, it does not mean that they do not use opioid drugs or engage in discussion about it in other parts of Reddit. It follows that we want to use the information contained /r/Opiates to identify the relevant parts – here, vocabulary – that allow us to find relevant content in that generated by authors who have a tie to Ottawa, such as being active in /r/Ottawa.

With the use of a novel machine learning, natural language processing and information retrieval search tool for social media, we demonstrate a way to anonymously estimate the size of a digital stigmatized at-risk community with regards to a physical place. This is done by estimating opioid usage by authors with a tie to Ottawa on social media using a scoring system that is designed to focus on the vocabulary of personal experiences.

We highlight our contributions as the following:

- We modify an existing machine-learning assisted search technique to automatically form keywords suitable for queries with information retrieval by identifying archetypal vocabulary from a collection of documents; we call this technique *Archetype-based Information Retrieval* (AIR)

- We show how to use AIR to estimate the size of a digital at-risk subpopulation as a proportion of authors of interest from social media; we call this AIR-Count.

- We show how to use AIR for public health search tasks on social media

4.2 Background

The use of social media for public health surveillance is made possible by massive amounts of data that have not previously been observable. It has taken on a few forms with regards to social media. One usage of social media was to gather primary data – for example, social media data was solicited from twitter authors who contributed to a hashtag as part of a research campaign [7]. Other usages involve observing public content which contains drug information: Personal drug correspondence has been extracted and analyzed from Twitter [8]. Researchers have investigated drug abuse behaviours on Instagram by examining images labeled with hashtags related to codeine abuse [9].

Studies in this area are beginning to confront new kinds of challenges in their analyses. A study of over 6 million tweets containing e-cigarette related hashtags had to handle filtering out advertising bots which contained their own health messaging [10]. One opportunity identified by the e-cigarette study was that new devices and products could be identified in the vocabulary of these tweets. However, owing to the classical unknown vocabulary problem, new words and products had to be manually identified [11]. Our work demonstrates how to capture new vocabulary and trends intelligently without requiring review of all possible vocabulary by PHSes.

Finding relevant content on social media is difficult [12]. The relatedness of words has a temporal aspect, as meanings and slang change over time [13], [14]. All these problems compound the difficulty of searching social media. Specifically, it

increases the difficulty of exploratory searches. Exploratory searches are for content that is unknown to the searcher – such as posts describing personal experience with opioid drugs [15], [16]. In order to address these challenges, we expand upon the use of an exploratory search technique called Archetype-based Modeling and Search (ABMS) [17].

This technique uses a repository of posts authored by Archetypal authors, such as opioid users from /r/Opiates, and compares them against a sample of authors who do not have this trait, and learns to identify the vocabulary which is most informative for Archetype membership. One of the advantages of using this technique in this space is that it can use temporally labeled data to identify modern slang, as pharmaceutical drug chatter has been shown to shift over time as well [18]. Identifying this vocabulary is important, as our language has been shown to be integral to how we create affiliation online [19].

In the original ABMS by Davis *et al.*, (2019), every author in the search space must have a neural representation generated for them. The scale of social media can make this challenging, and for the uses identified here, we make modifications such that AIR does not require neural representations for content it searches. Further, the results of the neural representation searches can be opaque to explanation by PHSes. AIR naturally identifies the language that it is using to identify potential archetypal authors with, allowing for better involvement by decision makers; this is a modern challenge in automated decision-making [20].

4.3 Methods

We begin by summarizing our whole methodology and then expand into specifics below. We present modifications to increase usability and interpretability of ABMS. With AIR, we produce a ranked list of words that are most representative of the archetypal class, in this case an online community related to public health interests, and facilitate keyword-based search. These words can be shown to PHSes before they perform their search of relevant social media. This approach combines all the advantages of ABMS with human in the loop oversight, as PHSes can customize the keywords before performing a search. Note, that while possible, logistical constraints prevented PHSes customizing the query. Further, this has greater utility for searching large collections of text without intensive computation on each author, as ABMS requires all authors being classified to have a vector representation learned by a deep neural network.

Once we have sourced a set of keywords, it is possible to integrate information retrieval techniques to analyze and retrieve relevant posts made by /r/Ottawa users. We use an open source information retrieval search and analytics engine called Elasticsearch [21]. Elasticsearch can use a probabilistic relevance framework called Best Matching 25 (BM25) to score documents [22]**;** we use this scoring. This allows for ABMS to form a query based on the most informative words for distinguishing opioid discourse. With this query, we can then retrieve relevant documents from /r/Ottawa social media. Detailed explanation of each step of our process follows**,** including an algorithm for the process**.**

Algorithm 2: Archetype-Based Information Retrieval

- (1) Input *D*, a set of documents, which contains the following subsets: *A*, a subset of archetypes (documents of interest), *C* a subset of controls (documents not of interest), *U*, an unknown corpus from which to extract relevant documents and *q*, the number of words to use in the query on *U*.
- (2) Use the words in all documents in *A* and develop a word representation *W* that maps words to vectors, with these words $\{ w_1, w_2, \ldots, w_i \} \in W$
- (3) Use *W* and develop a document representation, *V*, that maps all documents in *D* to vectors.
- (4) Train a classifier to distinguish V(*A*) from V(*C*). The classifier must be able to rank inputs according to their likelihood of belonging to *A* versus *C*.
- (5) Classifier outputs hyperplane *H* in the form of an *q* dimensional vector of the same dimensionality as all $w \in W$.
- (6) Apply a similarity metric between all values $w \in W$ & *H*
- (7) Retrieve *q* most similar *w*
- (8) Index all documents in *U* with an information retrieval system
- (9) Run a query with keywords $w_{1..q}$ using similarity metric *sim* (Ex: BM25).
- (Return the ranked set of results as output.
- Algorithmic Complexity: AIR scales dependent on five variables: (1) The number of unique words in the vocabulary; (2) the number of words in each document; (3) the number of documents to be learned; (4) the combined number of archetypes and controls; (5) the representation size for the word and document representation; (6) the similarity metric chosen to determine similarity between the decision boundary and the vocabulary words (i.e., cosine or dot product); (7) the choice of indexing method and implementation; (8) the query similarity metric chosen.
- The number of documents and the number of words in each document scale linearly for the number of representations to be learned. The larger the vocabulary size, the more computation is required to form each representation. For many representation techniques, this scaling will not be linear. Classifying the archetypes against the controls will scale dependent on the classifier that is chosen. The representation size is a parameter that scales to increase representation learning time and classifying time. Cosine and dot product scale linearly with input. Indexing implementations and search result times vary.

Algorithm 3: Archetype-Based Information Retrieval Count (AIR-Count)

- Input: Algorithm 2 (AIR), AIR requirements, Corpus of interest *U* consisting of documents consisting of all posts made by authors $a \in A$, *a priori* accuracy assessment of AIR relevancy *ACC* as a percentage, score threshold *S* as percentage of max score returned by AIR.
- Run AIR as described in Algorithm 2 on unknown set *U*
- From AIR results, return all documents scoring above *S*
- From *S*, multiply by *ACC*, *S x ACC* to output an estimate of authors matching the Archetypal representation adjusted by manual accuracy estimate.
- Algorithmic Complexity: Linear; all new operations are linear, and this algorithm is a trivial extension to Algorithm 2; the complexity is dependent on the implementation of Algorithm 2.

To model the specialized vocabulary and discourse that we are interested in, it is necessary to carefully cultivate relevant examples from social media. By using the collection of Reddit comments and submissions hosted by Pushshift we sourced all posts made in /r/Opiates from 2015 to 2018 [2]. Further, we sourced all posts made across all of Reddit by /r/Opiates participants in 2017. Next, we collected all posts made to /r/CasualConversation in 2017 to serve as contrasting examples to the archetypes sourced from /r/Opiates. Lastly, to demonstrate the utility of our technique in surveilling geographically relevant social media, we also pulled all posts made in 2017 by participants of /r/Ottawa.

This resulted in retrieving 1,280,681 posts from /r/CasualConversation, 10,479,255 posts made by /r/Opiates participants in 2017 and 6,371,674 posts made by /r/Ottawa participants in 2017 on Reddit. We retrieved 2,502,867 posts made directly to /r/Opiates from the start of 2015 to the end of 2018 to generate a word embedding with.

Data files were retrieved as Javascript Object Notation (JSON) and contain metadata about the posts, such as the time posted, the number of positive and negative votes on the post, identifiers to unique post ids that a comment could be responding to, amongst others. We extract authors' usernames and their posts' text for our analysis.

We remove all authors who have 'bot' somewhere in their username, as well as posts by 'Automoderator' as a basic method for removing automated accounts. Further, we remove any authors that posted more than 1500 times during the year 2017. The selected number of 1500 is arbitrary but aimed at removing accounts that were automated in some way. Hyperlinks and non-alphanumeric characters are removed from all posts by pattern matching. Then, all posts with less than 4 words and authors with less than 5 posts each are removed.

Once we have our collection of posts made by authors with their respective subreddit labels, we perform an ABMS [17] on the archetypal authors from /r/Opiates contrasted against the dialogue used by /r/CasualConversation authors. We present a brief summary of the technique and the choice of components used here.

Next, we need a vector representation of our words that captures semantic similarity. We generated a word embedding using the GloVe method [23] to measure statistical co-occurrences between words, grouping similar words together in vector space. The embedding was derived from the 2015 to 2018 /r/Opiates posts, capturing just over 77,000 unique words used during that time frame.

The vector representation of these words is used as input to a deep neural network which creates an aggregated representation of an author based on the words in their posts [24]. These vector representations are informed by the word co-occurrences in the /r/Opiates word embedding. Aggregate vectors are generated to represent each author in /r/Opiates and in /r/CasualConversation; generating vectors for /r/Ottawa authors allows other kinds of analysis but is not necessary for the work presented here.

The resulting author representations exist in high dimensional space and contain complex information about the relationships between author vocabulary. To isolate the relevant relationships, we use a support vector machine (SVM) [25] to classify author vectors as being more akin to /r/Opiates or /r/CasualConversation authors. This SVM produces a decision direction vector which points orthogonal to a high dimensional hyperplane that maximally separates the archetypes from the controls. The accuracy of this SVM on two different embeddings can be seen in **Table 1** in Chapter 3.

Using ABMS, we would stop here and use the SVM to directly classify authors and rank them by their distance from this hyperplane. However, we found that this method can be prone to biases. Thus, we reworked our design to allow a human in the loop to evaluate both the vocabulary used for searching and for reviewing the posts most indicative of /r/Opiates affinity. Hence, we measure the similarity of the *words* in our vocabulary to the hyperplane (since they are also embedded in the same vector space) and we collect the top 200 such words to use as a search query using BM25 on the documents from /r/Ottawa participants [22].

We use Elasticsearch, a state-of-the-art keyword-based information retrieval system to help solicit relevant documents from the /r/Ottawa participant corpus using the resulting words [21]. To do so, we first take all the posts made by /r/Ottawa participants and index them with this system. After indexing, we use the BM25 bag of words ranking function to score all the posts made by /r/Ottawa authors in 2017. BM25 is a variation on the popular *term frequency – inverse document frequency* technique, which increases the importance of words that occur in relatively few documents in the corpus and decreases the importance of words which appear ubiquitously such as ‘the’. From this ranking, we filter to the top ¾ of scores from the /r/Ottawa posts. We consider this set of documents to be of enriched relevance. Then, we count the number of unique authors with opioid related dialogue from /r/Ottawa participants.

This filtering allows for professionals to review the posts and determine if they appear to be from at-risk individuals or if they were identified because they contain dialogue that relates to opioids in incidental ways. We include a sample of authors with one or more high scoring documents with information that may be relevant in the appendix.

4.4 Results

We present the top ten results from using AIR to solicit words that are representative of our population of interest – the participants of /r/Opiates. Additionally, we perform an established search assistance technique called query expansion [26] on the word ‘opioids’ and form a query with the 200 most similar words as measured by cosine similarity. This is done to compare the effectiveness of AIR to a query expansion search, and to measure any overlap in results. This query expansion is done using the word embedding trained on /r/Opiates that is used in ABMS. We present the top ten results of each search in **Appendix Table 1**.

One immediate observation is that the two lists are not completely distinct, even in the top 10. This is particularly interesting for ABMS, as it is never directly provided with keywords, only subreddit labels in the form of ‘1’ or ‘0’ for binary classification. One of the posts describing a person’s experience is found in both, but it is considered more relevant in the AIR query than in the query expansion. Another observation from this limited set is that there are more posts describing personal experience in the ABMS results, while there are more cases of people attempting harm reduction with their fellow Redditors in the query expansion results.

We focus on the results from AIR for the rest of this article; interested authors could perform the same kind of analysis with query expansion results. As one of the primary goals of ABMS was to function without needing to know *any* domain relevant vocabulary, we choose to focus on the documents it suggests are relevant.

We show the score of the 2 million most relevant documents returned by our query in **Figure 10**. As expected for retrieving the most relevant documents, there is a large gap between the number of relevant documents, and less relevant or irrelevant documents. We section the results into quarters of the match score and collect the top 3 quarter scoring documents for further analysis. This results in a filtering down to 4431 documents from our initial set of 6,371,674. We note that it is possible there is relevant information in the remaining documents but given that our knowledge users – public health professionals – are unable to manually review documents numbering in the tens of thousands, we suggest this as an appropriate entry point.

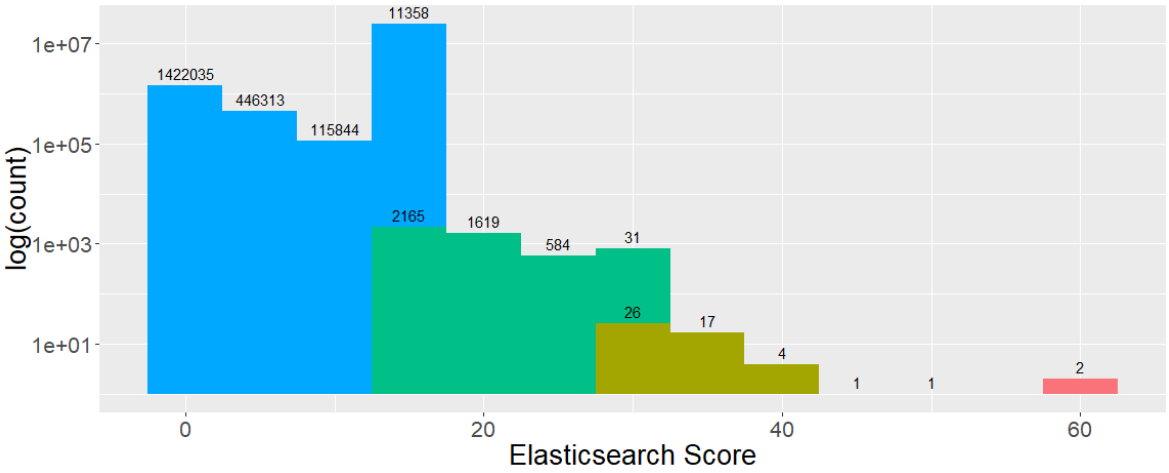


Figure 10. Distribution of Elasticsearch scores when using ABMS query. Y-axis is log transformed to allow smaller counts to be shown. Colours identify four separate ranges of scores, with indicated counts for each bin of scores.

To show trends inside of the results, we generated word clouds from the 2 million most relevant documents, and the top three quarters of the scoring set in **Figures 11 and 12** respectively. While it is unsurprising that some of the words are the same as those of the query, particularly in Figure 3, it is reassuring that these words still appear by raw frequency without the inverse document frequency adjustment that BM25 uses. Further, there is a clear change in words from the highest scoring subset

To go one step further in modelling the diversity of dialogue in our filtered step, we perform Latent Dirichlet Allocation (LDA) on the retrieved set of documents [27]. LDA is a topic modeling technique which outputs a set of words that are tied together under a common label, and these words collectively represent the 'topic'. We present the topics from both the 2 million result set and the top three quarters score set. These figures are in **Figures 13 and 14** respectively.

One of the opportunities offered by our system is the ability to count the number of documents that match our given criteria, which allow us to estimate an overall number of participants from /r/Ottawa which engage in opioid related dialogue. Although observation of our results suggests that not everyone who engages in these discussions engages in illicit use of these drugs, it provides an initial estimate that can be refined as needed to suit a PHS's needs. We suggest that the number is correlated to the true number, even if it is not without noise and error.

We count the number of unique authors from the top ¾ scoring posts, which is 1761 from the total of 4431 posts. This also means that the average number of posts per author, in this top scoring subset, is 2.516, with a standard deviation of 14.79. As an author cannot have a negative number of posts, this distribution is skewed to the right.

Figure 11. Word cloud showing the terms and frequency, encoded as size, of posts made all participants in /r/Ottawa. This image shows the baseline frequency of words before any filtering is applied to the set. English stop words and other common words like ‘Reddit’ were removed.

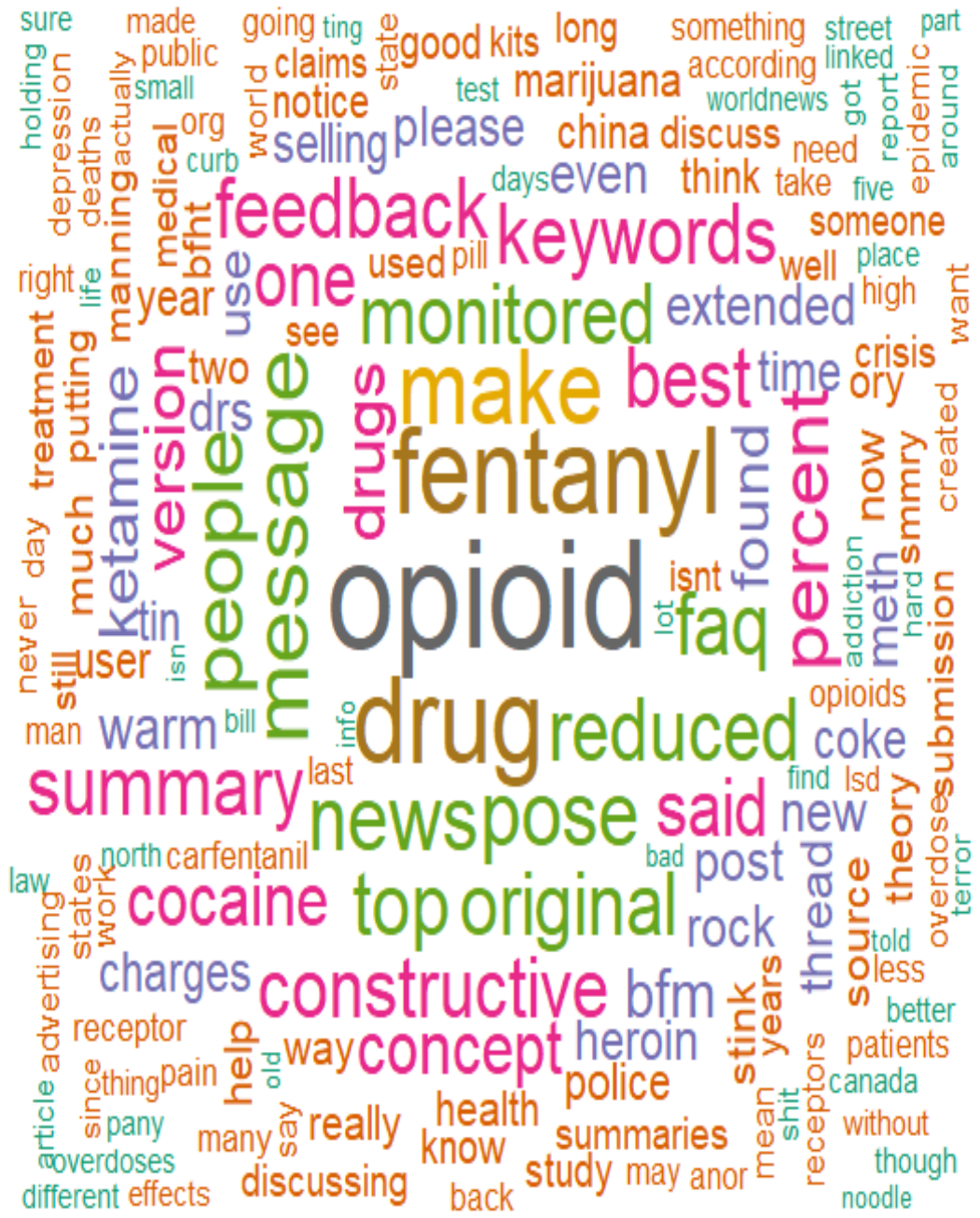


Figure 12. Word cloud showing the terms and frequency, encoded as size, of posts made by participants in /r/Ottawa after filtering to include only the top three-quarters of scores from using Archetype-Based Modeling and Search sourced dialogue. English stop words and other common words like ‘Reddit’ were removed.

A few accounts are more prolific than others. Of the 4431 posts, 554 are from an account called “autotldr”. This is a bot account that will post summaries of news articles – so while these posts are not made by an author directly, they are made in response to a news article with something related to our query about opioids. The original author’s post is less likely to be included if the headline does not include our keywords, and they are likely not in our 1761 unless they made other relevant posts. Thus, we choose to include these in our total activity.

Our entire /r/Ottawa corpus from 2017 has 19376 authors with no filtering for automated accounts. From analysis of the top 10 posts, not every post shows clear indication of opiates usage. While a systematic reviewing of these posts, outside the scope of this work, would be better to count the number of authors, we use this as an initial measurement heuristic. Of our top 10, which admittedly is an enriched sample, 4 of the 10 shows signs of illicit substance use. This suggests that 704 authors may engage in substance abuse, forming 3.635% of the /r/Ottawa participants. This number, rounded to 4%, matches the rate of prolonged opioid use following trauma or surgery [28]. This 4% of the online population contributes only 0.06435% of the total dialogue generated by /r/Ottawa participants, highlighting that this is not the only conversation that occurs.

We present collected thoughts on the usage of these techniques, identify ethical concerns, and identify future directions in which this work could be taken.

4.5 Discussion & Conclusion

Here we have demonstrated how we can combine techniques from information retrieval with knowledge that is discoverable from abstract labels using our machine learning & NLP technique. While we have previously shown utility from directly classifying the author representations as archetypal or not, this does not address issues of potentially problematic bias in an already stigmatized population. By building a human into the loop, we simultaneously incorporate the skills of PHSes and allow for bias to be mitigated by professionals.

From the kinds of results solicited in our reviewed top 10 list of posts, we suggest that AIR offers greater utility in finding posts that were made by members of the at-risk opiates using subpopulation that we set out to study. We attribute this as a result of AIR being trained on author representations, while query expansion is solely meant to find words with the most similar co-occurrence statistics. Given the nature our of discoveries and the implications for both future study, privacy, and public health, it is important to consider the origin of the data presented here.

Reddit, as a form of anonymous social media, presents a unique lens into the lived experience of the individuals whose writing we are analyzing. Anonymous writing has been shown to be distinct from what is shared tied to our real-world identities; classifiers are efficiently able to separate the two based on linguistics [29]. The motivation behind the writings is essential to consider for building a proper understanding appropriate for evidence-based interventions and assistance within public health.

Figure 13. Latent Dirichlet Allocation was used to model topics present in the /r/Ottawa participants corpus. Relative weighting (beta score) of the terms for the topic are shown.

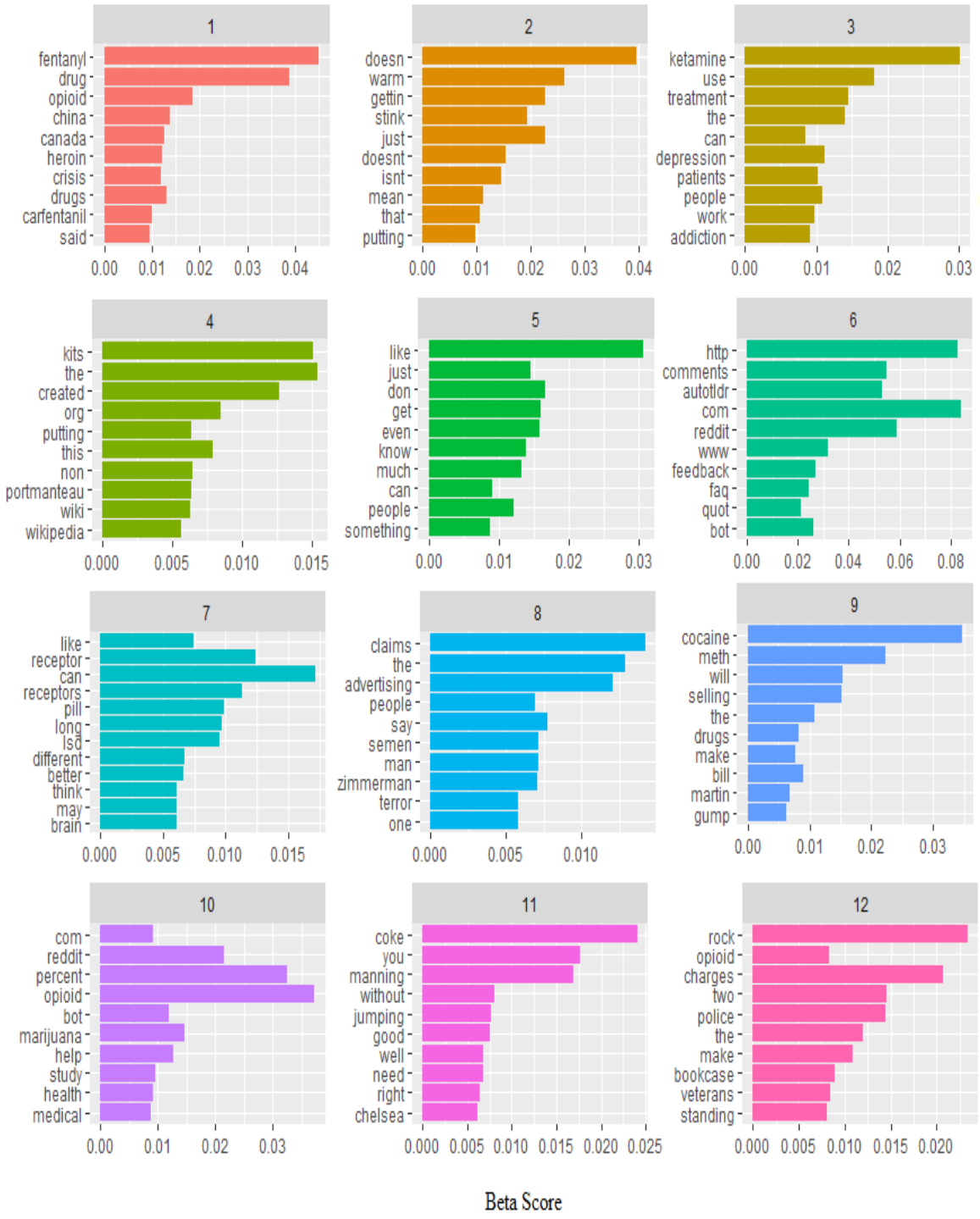


Figure 14. Latent Dirichlet Allocation was used to model topics present in the top three-quarters score from the AIR /r/Opiates query subset of the /r/Ottawa participants corpus. Relative weighting (beta score) of the terms for the topic are shown.

Work by Morrison has applied rhetorical genre theory to social media to understand the motivation behind participation in the communities found in places like subreddits [30]. Morrison studied how new mothers could use blogging as a medium to better understand their new roles and to adapt to the challenges they provide. Morrison also noted that there was social isolation in this new role. It is likely that stigmatized at-risk subpopulations experience similar kinds of isolation; this has been observed in homeless populations [31]. It is plausible that many opioid users are struggling with the social isolation of their addiction and the consequences of it and turn to anonymous social media for a sense of community that they do not find in their daily lives.

One takeaway from our study of this phenomenon is that drug usage is hard to classify into discrete, convenient categories. Trying to think of someone as solely an opioid drug user cuts out the co-occurrence or mixing with other drugs. A holistic approach to understanding the potential usage patterns and lifestyle of those who are at-risk would be wise to incorporate the understanding that any opioid, or combination of opioids, is unlikely to be the only substance at play in understanding their challenges.

An unavoidable ethical and privacy related challenge of analyzing this kind of data comes from the source—individuals that are part of a population. Care needs to be taken to ensure that there is not undue risk put onto those being studied. Considering this, we have ensured that no usernames or identifying information are presented in our results. It is possible that some posts retrieved could contain identifying information that would allow individuals to be identified. If someone decided to post something about a landmark by their house while intoxicated, this could deeply compromise their privacy. Individuals using AIR should consider who is able to access search results.

However, this information has all been posted in public forums. No special access was required to view it, and the usernames and entire posting histories would be viewable for those that searched for and found the content organically. This tumultuous environment is characteristic of the state of social media, highlighted by breaches like Cambridge Analytica [32]. However, given the massive penetration of social media into society, new challenges like this may be unavoidable. We offer here what we believe is an avenue for public good amidst the chaos, and hope other researchers find utility in supporting those at risk.

We demonstrate the ability of an Archetype-Based Modeling and Elasticsearch to retrieve documents that demonstrate archetypal behaviours, like opioid drug usage, from population level datasets. While we demonstrate a specific case study of the technique, we note its general applicability for researching other public health phenomena. We prototype using this system to measure a count of behaviour and suggest that 4% of /r/Ottawa participants show signs of using opioid drugs.

4.6 Future Work

One natural direction to take this work is to set up monitors for public health corresponding to the appropriate geographic catchment areas. By using the vocabulary identifying abilities of AIR with live data feeds of Reddit, Twitter, Facebook or other social media, it would be possible for PHSes to continuously monitor the incoming flood of data for cries for help, or other intervenable events.

For best performance of a technique with machine learning components, it is typical and helpful to tune the hyperparameters used by the technique. The best individuals to perform this tuning would be the PHSes, who were unfortunately unavailable because of the COVID-19 pandemic. Future work should thus include considerations for hyper-parameter tuning and how to communicate differences in information retrieved based on the tuning to PHSes.

To build further upon the potentials of AIR, there is a strong case for visual analytics. Big data and public health have previously been identified as a natural pairing for visual analytics [33]. AIR has natural opportunities for this, of which we identify three. First, by visualizing all words used in a word cloud or other form, searchers can see the extent of what they are querying with. Secondly, we can help mitigate the arbitrary choice of 200 words in the query with sensitivity encoding, which will allow searchers to see the words just outside the limit which can then be added or discarded as appropriate. Third, an interactive interface by which words can be added and removed by the searcher would solidify AIR as a cognitive tool for searching by combining its ability to solicit vocabulary at scale with the domain expertise of the searcher.

4.7 AIR Bibliography

[1] R. Tourangeau, B. Edwards, and T. P. Johnson, *Hard-to-survey populations*. Cambridge University Press, 2014.

[2] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, and P. Io, "The Pushshift Reddit Dataset." Accessed: Jan. 23, 2020. [Online]. Available: https://files.pushshift.io/reddit/.

[3] S. Pandrekar *et al.*, "Social Media Based Analysis of Opioid Epidemic Using Reddit," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2018, pp. 867–876, 2018, Accessed: Dec. 05, 2019. [Online].

[4] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – Challenges in topic discovery, data collection, and data preparation," *International Journal of Information Management*, vol. 39, pp. 156–168, Apr. 2018, doi: 10.1016/J.IJINFOMGT.2017.12.002.

[5] "National Report: Apparent Opioid-related Deaths in Canada - Data Blog - Public Health Infobase | Public Health Agency of Canada." https://health-infobase.canada.ca/datalab/national-surveillance-opioid-mortality.html (accessed Dec. 05, 2019).

[6] P. Seth, R. A. Rudd, R. K. Noonan, and T. M. Haegerich, "Quantifying the epidemic of prescription opioid overdose deaths," *American Journal of Public Health*, vol. 108, no. 4, pp. 500–502, Apr. 2018, doi: 10.2105/AJPH.2017.304265.

[7] W. Clyne, S. Pezaro, K. Deeny, and R. Kneafsey, "Using Social Media to Generate and Collect Primary Data: The #ShowsWorkplaceCompassion Twitter Research Campaign.," *JMIR public health and surveillance*, vol. 4, no. 2, p. e41, Apr. 2018, doi: 10.2196/publichealth.7686.

[8] R. Daniulaityte, L. Chen, F. R. Lamy, R. G. Carlson, K. Thirunarayan, and A. Sheth, ""When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets.," *JMIR public health and surveillance*, vol. 2, no. 2, p. e162, Oct. 2016, doi: 10.2196/publichealth.6327.

[9] R. Cherian, M. Westbrook, D. Ramo, and U. Sarkar, "Representations of Codeine Misuse on Instagram: Content Analysis.," *JMIR public health and surveillance*, vol. 4, no. 1, p. e22, Mar. 2018, doi: 10.2196/publichealth.8144.

[10] J.-P. Allem, E. Ferrara, S. P. Uppu, T. B. Cruz, and J. B. Unger, "E-Cigarette Surveillance With Social Media Data: Social Bots, Emerging Topics, and Trends.," *JMIR public health and surveillance*, vol. 3, no. 4, p. e98, Dec. 2017, doi: 10.2196/publichealth.8641.

[11] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, “The vocabulary problem in human-system communication,” *Communications of the ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987, doi: 10.1145/32206.32212.

[12] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding high-quality content in social media,” in *Proceedings of the international conference on Web search and web data mining - WSDM ’08*, 2008, p. 183, doi: 10.1145/1341531.1341557.

[13] G. D. Rosin, E. Adar, and K. Radinsky, “Learning Word Relatedness over Time,” Jul. 2017, Accessed: May 08, 2019. [Online]. Available: http://arxiv.org/abs/1707.08081.

[14] E. Sagi, S. Kaufmann, and B. Clark, “Semantic Density Analysis: Comparing word meaning across time and phonetic space,” 2009. Accessed: May 07, 2019. [Online]. Available: http://delivery.acm.org/10.1145/1710000/1705429/p104-sagi.pdf?ip=99.243.102.9&id=1705429&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1557290462_0607ff09b037412a4aa7afcc4b4c5e0c.

[15] G. Marchionini, “Exploratory search,” *Communications of the ACM*, vol. 49, no. 4, p. 41, Apr. 2006, doi: 10.1145/1121949.1121979.

[16] G. Marchionini, “Exploratory search: from finding to understanding,” *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.

[17] B. D. Davis, K. Sedig, and D. J. Lizotte, “Archetype-Based Modeling and Search of Social Media,” *Big Data and Cognitive Computing*, vol. 3, no. 3, p. 44, Jul. 2019, doi: 10.3390/bdcc3030044.

[18] M. T. Wiley, C. Jin, V. Hristidis, and K. M. Esterling, “Pharmaceutical drugs chatter on Online Social Networks,” *Journal of Biomedical Informatics*, vol. 49, pp. 245–254, Jun. 2014, doi: 10.1016/J.JBI.2014.03.006.

[19] R. Vessey, “Zappavigna, M. (2012). Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. London: Bloomsbury,” Springer, Cham, 2015, pp. 295–299.

[20] K. Brennan-Marquez, D. Susser, and K. Levy, “Strange Loops: Apparent versus Actual Human Involvement in Automated Decision-Making.” Oct. 02, 2019, Accessed: Nov. 19, 2019. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3462901.

[21] C. Gormley and Z. Tong, *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. “ O’Reilly Media, Inc.,” 2015.

[22] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, Dec. 2010, doi: 10.1561/1500000019.

[23] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] S. Amir, G. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace, “Quantifying Mental Health from Social Media with Neural User Embeddings,” 2017. Accessed: May 08, 2019. [Online]. Available: http://clpsych.org.

[25] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks Editor,” Kluwer Academic Publishers, 1995. Accessed: Jun. 14, 2019. [Online]. Available: https://link.springer.com/content/pdf/10.1007/BF00994018.pdf.

[26] F. Diaz, B. Mitra, and N. Craswell, “Query Expansion with Locally-Trained Word Embeddings,” May 2016, Accessed: May 08, 2019. [Online]. Available: http://arxiv.org/abs/1605.07891.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[28] A. Mohamadi *et al.*, “Risk factors and pooled rate of prolonged opioid use following trauma or surgery: A systematic review and meta-(regression) analysis,” *Journal of Bone and Joint Surgery - American Volume*, vol. 100, no. 15, pp. 1332–1340, 2018, doi: 10.2106/JBJS.17.01239.

[29] D. Correa, L. A. Silva, M. Mondal, F. Benevenuto, and K. P. Gummadi, “The Many Shades of Anonymity: Characterizing Anonymous Social Media Content,” *Ninth International AAAI Conference on Web and Social Media*, Apr. 2015, Accessed: Dec. 10, 2019. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10596.

[30] A. Morrison, “Social, Media, Life Writing,” in *Research Methodologies for Auto/biography Studies*, New York, NY: Routledge, 2019.: Routledge, 2019, pp. 41–48.

[31] A. Rokach, “Private Lives in Public Places: Loneliness of the Homeless,” *Social Indicators Research*, vol. 72, no. 1, pp. 99–114, May 2005, doi: 10.1007/s11205-004-4590-4.

[32] C. Cadwalladr and E. Graham-Harrison, “Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach,” *The guardian*, vol. 17, p. 22, 2018.

[33] K. Sedig and O. Ola, “The Challenge of Big Data in Public Helth: An Opportunity for Visual Analytics,” *Online Journal of Public Health Informatics*, vol. 5, no. 3, Feb. 2014, doi: 10.5210/ojphi.v5i3.4933.

5 Archetype-based Temporal Language Adaptative Stratification

5.1 Introduction

Using representation learning, information retrieval and search engines, Archetype-based Modeling and Search (ABMS) and Archetype-based Information Retrieval (AIR) enable exploratory search by expanding the way that queries can be formed [1], [2]. Normally, making a query for exploratory search requires the formation of keywords. Depending on the desired result of the search and the obscurity of domain-relevant vocabulary, the query formation task can be complex. Consider forming a query that finds posts made by members of a particular community, e.g., caffeine addicts. Some of query words may be obvious – coffee – but others may be more elusive. Through more detailed study, experts can often identify additional keywords that individuals might use to describe their state, such as being "jittery". We develop a new algorithm that uses and builds upon ABMS and AIR to describe the temporal nature of archetypal behaviours. For health behaviours, public health stakeholders wish to know about the prevalence, which ABMS and AIR assist, but also the progression and periodicity (i.e., cyclical every two months) of these behaviours. This chapter helps to fill that knowledge gap.

ABMS and AIR attempt to reduce the time, effort, and cognitive load of these problems by reducing the task of query formation to identifying text written by individuals match the desired archetype, i.e., caffeine addict. ABMS works by directly assigning a score to a machine-learning formed representation of an individual's written posts which represents their affinity to the archetype. All individuals in a search space can be assigned a score, sorted, and reviewed or otherwise counted based on a threshold. AIR works in a similar way, but lowers the non-trivial computational cost of generating an author representation of every individual in the search space. This is done by identifying the individual words which are most representative of the archetypal group as contrasted against a control or non-relevant group. These keywords can then be used as input to an information retrieval system – such as Elasticsearch using Lucene – and a score can be output using a similarity metric like BM25 [3], [4]. From a combination of these scores, we can identify individuals who match a given archetype.

These approaches, however, assume that the archetypal identity remains consistent throughout the time being considered. For many purposes, such as counting the number of matching individuals for a population size estimate, this assumption is valid or trivial if violated; the individual still displayed the behaviour, and so long as they are not counted multiple times if they cease and resume it, a single count suffices to include them. In other settings, such as trying to study depressive episodes as they appear on social media, it is precisely this transition which is of interest. As such, these previous techniques are not meant to handle looking at this gap. We develop a new algorithm

which we call *Archetypal Temporal Language Adaptive Stratification* (ATLAS). ATLAS builds upon ABMS and AIR to help users understand how an individual's behaviour, as understood by the written language they use to communicate it, changes over time. ATLAS provides 1) algorithmic assessment of a change in behaviour over time and 2) the words in their vocabulary that are representative of this change over the same time. Further, by aggregating individuals into a community, it can be used to measure change in a community's behaviour over time. We do not examine this case specifically, but by changing the aggregation of documents from aggregates of authors to aggregates of online communities, ATLAS can be applied without further modification. ATLAS builds upon AIR's ability to extract representative language that is associated with this change.

We summarize our contributions with ATLAS as follows:

- We develop and identify a new algorithm, ATLAS, which operates in two stages, ATLAS-Explore and ATLAS-Map.
    - ATLAS-Explore can be used to identify *individuals* who have changed labels over a given temporal period
    - ATLAS-Map is used to find, or map, the *language* that is representative of this change.
- We present a case study with social media data for the less sensitive parts of, and synthetic data for the sensitive parts of, ATLAS.
- We identify applications which could have societal and public benefit from the use of ATLAS

5.2 Background

There are two main areas of background that are important to consider for ATLAS. The first area is the technical area of machine learning and natural language processing with regards to time series modeling. While ATLAS requires transforming a continuous time series into a pre-defined series of discrete intervals, it is important to refer to previous work in the area to build a foundation from. The second area of background is the qualitative aspect of language and identity. However, many of these studies include quantitative and empirical measures of how language changes over time in humans. An understanding of both is necessary to justify why language change over time can reflect a behavioural change that is helpful for tasks such as understanding an individual's experience with addiction.

5.2.1 Time series Modeling with Machine Learning & Natural Language Processing

Time series modeling is a complex field; it must deal with all the complexities of typical data analysis while adding in the complexities of change over time. Textbooks have been written dealing with the specific challenges that are encountered when time is a variable to be considered [5]. The idea of using unsupervised learning to generate representations for use in time series modeling is not new, and has been well documented, including with deep learning variants and with applications in speech recognition systems [6]. Modeling a user via natural language has been previously performed [7]. This is done by using pre-trained word embeddings from techniques like GloVe to train representations with usr2vec [8], [9].

Owing to the complexities of the analysis, and likely due to limitations in sufficient amounts of data with a labelled temporal component, there is a gap in the literature about using natural language processing for the analysis of temporally stratified text. One possible explanation for this gap is the difficulties that come from the relations of word to one another changing over time, complicating the ability to do time series analysis over elongated periods of time [10]. To the authors knowledge, ATLAS is one of the first techniques to use NLP and time series analysis towards domain specific applications of practical interest to practitioners in public health and the social sciences.

Language diversity poses a problem for many kinds of NLP analysis, and language is notably diverse in social media [11]. The variety of the language poses a problem for NLP, particularly in forecasting, which is the typical time series analysis task. Non-canonical language is a well-established problem in NLP for which there is no universal solution [12]. While some word embeddings can attempt to infer meanings of words based on the composition of the word, this technique suffers when inferring out of vocabulary word embeddings [13]. As time goes on, the slang that is being used on places like social media is going to change, and the way that language is regarded will change as well. Memetics that are popular at one time can receive an entirely different reception after they have gone out of favour. This idea of personalization and community adaptation continues into a discussion of how language and identity are interwoven on digital platforms.

5.2.2 Identity, Language, and Language Change Over Time

The idea of language and its expression is central to the idea of human development, to the point where progression of language in adolescents is used as a marker of human development [14]. Language, too, changes as individuals age and socialize with others outside their age group; studies have found that exposure to the internet has had a measurable impact upon [15]. This has the implication that, over a long enough time, a change in vocabulary is associated with an individual's personal change.

While sometimes considered intuitive to the point of feeling trivial, the language used in online communications contains more about our identities than a fleeting conscious thought may impart. For example, a study of 75,000 volunteer's Facebook information found associations between living environment and the language used, about the language used by people suffering from various mental illness, and some gender specific differences which could be used to infer gender [16].

Consider something as simple as saying "Go Leafs Go!" on Twitter. From those three words, in the context they are given, it can be inferred that the person is talking about hockey, about the sports team the Toronto Maple Leafs, and by association to the geography of the sports team, has an above average chance of living in Canada, likely in Ontario, possibly in Toronto. This possibility of inferring has been shown in other studies; the language used online allows us to create affiliation to social groups, activities, hobbies, lifestyles and more [17].

As another researcher puts it, individuals on social media are creating a form of an autobiography of themselves, and the information contained in here can be very much associated with their development as a person, or informative of their day-to-day doings [18]. If a person's favoured sports team changes, then the language they use on social media to discuss sports will change to reflect that. This is the kind of association that ATLAS seeks to uncover. It, by no means, attempts to infer a causal relationship between the two, and this limitation of the approach is better served by other methods [19].

There is a body of evidence suggesting that the language used to communicate online or elsewhere is associated to a person's sense of identity; this identity evolves over time, and ATLAS is a tool to examine words representative of those changes by building on ABMS and AIR.

5.3 Methods

ATLAS is a hybrid machine learning, information retrieval and natural language processing algorithm that identifies authors who have undergone a change in archetypal behaviour and extracts keywords that represent that change. ATLAS is comprised of several components and representation learning stages, which we describe below. The stages are:

1. Timeframe Selection
2. Data Retrieval
3. Word Embedding Training
4. Classifier Training
5. Indexing
6. Scoring of Temporal Representations by Archetypal Query and Classifier
7. Identification of Archetypal-to-Nonarchetypal Transitionary Authors
8. Identification of Archetypal-to-Nonarchetypal Vocabulary

First, range of time must be selected to set the scope of ATLAS. A long enough time must be chosen for there to be sufficient data for training multiple representations. This interval could be daily, weekly, monthly, or longer, but it is limited by the amount of content that the author has generated. Without sufficient posts per author per interval, the representations will not have sufficient information to be compared to one another.

The second stage is data retrieval. The choice of documents and vocabulary are very significant to consider and pick carefully, as vocabulary that is not found within the chosen document set will not be identified later from being out of vocabulary. Downstream representations will be constructed as composites of words, and it is ultimately these words which will be retrieved to communicate changes in behaviour or archetype the algorithm found. An approximately equal amount of archetypal and non-archetypal text should be retrieved; an imbalance may be considered when the archetypal behaviour is known *apriori* to be rare.

The next stage is to construct a word embedding. This embedding should be trained on the entire set of posts above, although it can be trained on just the archetypal authors and their posts if need be. The most important part in training is to provide the word embedding with the full vocabulary used by these authors.

With a pre-trained word embedding, the next step is to generate representations of the author collections using a representation learning technique; this could be any, but the only published one currently is usr2vec [9], [21]. For each author, both archetypal and control, a representation of both their entire document collection and an individual one consisting of the temporally separated posts from each interval must be created. At the end of this step there will be a collection of author representations trained on all posts, and a temporally ordered set of representations trained on a time delimited subset of their entire posts.

Using the full-length author representations, the fourth step is to train a classification algorithm to separate the archetypal from non-archetypal case. Any classifier used must output a decision boundary or decision direction from which the similarity of the previously trained word vectors can be compared to.

Using the decision boundary between the two classes, a decision direction which is perpendicular to the decision boundary and points towards the archetypal representations is computed. This decision direction vector is compared to all word vectors using a similarity metric; cosine and dot product are typical. This produces a ranked list of keywords from the vocabulary of the individual word vectors. These are the words which are most closely aligned with the decision direction, suggesting they are most strongly aligned of all words to the archetypal authors.

Next, the text used to train the representations is indexed per author per interval using an indexing system like Elasticsearch [3]. Keywords from the previous step are used as a query here. The proper number of keywords to use depends on the vocabulary, and should be determined by the user of ATLAS. These keywords are used to query the index and a score is assigned based on an information retrieval scoring metric, i.e., BM25 [4]. This produces scores for the delimited author collections from the classifier and the information retrieval system.

With this collection of scores, it is necessary to set a threshold for each that satisfies the needs of the application. The classifier accuracy is one way to set this for the classifier; the information retrieval score needs to be manually analyzed and interpreted. It is necessary to set a threshold for both the archetypal and control label. This can be customized, but by default ATLAS uses a 50% classifier score 50% information retrieval score weighted average.

The next stage depends on whether the preference is to find individuals who had the archetypal behaviour and ceased having it, or if it is to find individuals who did not have the archetypal behaviour and then began to exhibit it. It will be assumed that the goal is to find archetypal behaviour which later becomes absent. All individuals who have a score about the threshold are collected from the set of all authors. Then, for every individual who is a member of this collection, ATLAS looks through the collections made by the same author for every time delimitation occurring after the original archetypal one. All individuals that pass this criterion with one or more occurrences are collected.

To follow are some considerations of the cases that could occur when comparing an archetypal representation of a given author to another representation that is non-archetypal. Given a representation r_i with representation r_j that follows at interval ct, where t is the interval and c is an integer greater than or equal to 1, there are the following cases possible (and more than one case is possible at a time):

- r_j immediately follows r_i, i.e., c = 1
- c > 1, and there exists one or more r_k inbetween r_i and r_j that are archetypal behaviour.
- c > 1, and there exists one or more r_k inbetween r_i and r_j that are non-archetypal behaviour.

A SVM is used to separate the pairs of representations and output a decision direction, and then the words most associated with that decision direction are extracted. For each pair, this list of words is output representative of the change between pairs. For visualization purposes, a word cloud where the word size is proportional to the magnitude of the word vector with the decision direction is effective.

One last comparison that can be made, if the author has more than two representations available for the time analyzed, is to repeat the same process with all the archetypal representations of the same author against all their non-archetypal representations.

We describe this algorithm in **Algorithm 4**, with the following terms: Word vectors $w_{1..n}$ where W is a word embedding consisting of individual word vectors w_i. Author vectors $q_{1..n}$ where Q is the collection of all author vector representations without a time separation, with *aq* for archetypal vectors and *bq* for non-archetypal vectors. The set of all time separated author representations is referred to as QT. Time interval delimited archetypal author representations are shown by aq_it_m. Non-archetypal vectors do not need to be split. The vocabulary of the word embedding is referred to as V. Score calculations are labelled $s(C)$ and $s(E)$ for classifier C score and keyword similarity metric score E. A tunable parameter γ between 0 and 1 controls the weighting of the classifier against the information retrieval metric; we default γ to 0.5. Thresholds are labelled *archetypal_threshold* and *nonarchetypal_threshold*. Filtered representations above the sensitivity are called $faq_{1..n}$.

Algorithm 4: Archetypal Temporal Language Adaptive Stratification (ATLAS)

Inputs: Set of archetypal documents with time stamps, *AD*, created by a set of authors $a_{1..n} \in AD$, set of non-archetypal documents, *BD*, authored by a set of users $b_{1..n} \in B$. Labels from being a member of either *AD* or *BD*. A time interval *t*. A positive integer amount of keywords *x* and *y* to use in scoring information retrieval results for representations and archetypal-to-nonarchetypal representations, respectively.

Terms:

Outputs: (1) All authors from *AD* that changed from archetypal to non-archetypal status, and the language representative of that change for each transition. (2) The language representative of the change from all archetypal representations to all non-archetypal representations.

Algorithm 4.1: ATLAS-Explore

- (1) Train word embedding W on documents $AD \cup BD$.
- (2) Train set of author representations Q for each $a_i \in AD$ and $b_i \in BD$
- (3) Train set of time separated author representations for each $aq_it_m \in QT$
- (4) Train a classifier C to separate aq from bq and retain all classification scores for aq.
- (5) Extract decision direction from C, calculate similarity of all $w_i \in W$ and sort
- (6) Index all AD using some information retrieval system i.e., Elasticsearch.
- (7) Take top x keywords and perform a query on AD with similarity score, i.e., BM25.
- (8) Identify all $aq_it_m \in QT$ with $\gamma s(C)+((1-\gamma)s(E)) > archetypal\_threshold$
- (9) With a list l that has as elements pairs of aq_it_m, collect items into the list as follows:
  For i = 1; i < length(faq); i++ do
  (10) for each representation j with the same author as i where $t_j > t_i$ & j has (0.5s(C)+0.5s(e)) < *nonarchetypal_threshold*, append pair i,j to l.
- Output list l, containing every pair of archetypal-to-nonarchetypal representations across all others
- Algorithmic Complexity: This algorithm scales mostly depending on the number of authors, and the number of total documents scaled by the authors. The scaling of training the word and author representation, the classification task, and the information retrieval tasks are dependent on implementation. The extraction of relevant vocabulary scales linearly with the number of words in the vocabulary. The retrieval of archetypal authors that have a shift to non-archetypal behaviour scales dependent on the number of authors, and the number of temporal divisions to iterate over.

Algorithm 4.2: ATLAS-Map

- (1) For every pair aq_it_m, aq_jt_n, $\in l$:

  (1.1) Train a classifier to distinguish between the two representations.

  (1.2) With the decision direction from the classifier, extract top *y* keywords

  (1.3) By similarity metric, i.e., cosine or dot product with words $w_i \in W$

  (1.4) Generate a word cloud of the top *y* keywords with word size proportional to the magnitude of *y* with the decision direction via the similar metric.

- (2) For every author $a_i \in AD$ with at least one archetypal-to-nonarchetypal pair, do:

  (2.1) Collect all $aq_it_m \in QT$ with $(0.5s(C)+0.5s(E)) > archetypal\_threshold$

  (2.2) Collect all $aq_jt_n \in QT$ with $(0.5s(C)+0.5s(E)) < nonarchetypal\_threshold$

  (2.3) Train a classifier to distinguish between the collection of archetypal representation against nonarchetypal representations

  (2.4) Extract the top *y* keywords from author vocabulary using a similarity metric between the decision direction and $w_i \in W$

  (2.5) Generate a word cloud of the top *y* keywords with word size proportional to the magnitude of *y* with the decision direction via the similar metric.

- Algorithmic Complexity: This section of ATLAS scales dependent upon the number of representations which have one or more matches to a representation which changed label. The scaling of the classifiers will be dependent on this amount. The extraction of relevant vocabulary scales dependent on the number of words in the vocabulary. The creation of the word cloud will be dependent on *y*.

5.4 Partially Synthetic Case Study: Depressive to Non-Depressive Transitions on Reddit

We now present a case study that demonstrates ATLAS as applied to a collection of social media documents that are relevant for the task of better understanding depression as it appears on Reddit. In this case study, we construct archetypal representations of depressed authors from Reddit data retrieved from Pushshift. From the construction of archetypal representations, we use ATLAS to show how these representations can be used to identify individuals that experience transitions between, loosely, depressive, and non-depressive states. We supplement with synthetic data at the last stages to avoid the ethical complexities of looking at individual's transitions from depressive to non-depressive states and reporting on it in a published work without ethical approval.

5.4.1 Synthetic Case Study Methods

We collect all the authors on Reddit that posted one or more times on /r/Depression in October through December 2019. Further, all posts made by these authors were collected. This forms the archetypal set for our analysis. For control and contrast, a combination of authors sampled from /r/aww, /r/CasualConversation, /r/totallynotrobots and /r/AskReddit were pulled to form the control set. This set was chosen to obtain dialogue that would have less depressive content than a random subset, and still be representative of typical behaviour by individuals on social media. Totally not robots was selected as it contains humans attempting to imitate automated accounts; the content is both non-depressive and may associate automated accounts more with the control case. In total, 81,118 /r/Depression archetypal authors were collected, and 143,737 control authors were collected. While the relatively small class imbalance could be fixed by discarding control authors, we opted to retain some imbalance with the intuition that depressive content is rarer than non-depressive content. Thus, the model being trained to perform better in imbalanced scenarios could be beneficial.

Using GloVe, a word embedding was trained on all the posts made by the /r/Depression authors. This ensures that the complete vocabulary of these authors is available at later stages, which is important when trying to find the vocabulary associated with a change between depressive and non-depressive states. A total vocabulary size of 164,298 was found, with each word having a corresponding vector representation. With the pre-trained word embedding, representations of each author from the control and archetypal sets were generated using usr2vec. A linear SVM was trained and tuned on this set using a 75/25% training/test split. After evaluating the performance, the SVM was retrained on the entire set. With the trained SVM, the cosine similarity between the SVM's decision direction and each individual word vector from the /r/Depression authors' vocabulary. These are ordered and used as the basis of a keyword search when documents are later indexed.

Next, we synthetically created an author document that is split in a temporal manner. We imagine this author to be making posts to a social media platform such as Reddit. Given that these documents are synthetic, the distinction is entirely arbitrary, but we can imagine each set of documents as belonging to a few weeks each. For our first synthetic author, we generated four documents, two of which displayed depressive behaviours and two of which do not. The synthetic narrative for the first author is someone with a self-identified gaming addiction, experiencing sleep disturbances, spending a lot of time on social media. In the first document, these are identified and in their peak. In the second document, the author begins to engage in resilient behaviours like going onto the platform less, discusses spending more time outside, and trying to exercise more. In the third document, the author discusses having trouble sleeping again, not finding the motivation to exercise, their experience going down a metaphorical rabbit-hole on YouTube, playing games more, and using alcohol as a sleep aid. In the fourth and final document, the author describes finding more balance in their lives again.

Representations of each temporally separated collection of documents were then trained using the pre-trained word embedding from the /r/Depression authors. The documents belonging to each author were then indexed in combination with the control set documents into an Elasticsearch index. This was done to provide document frequencies for the various terms – in a non-synthetic study, there would be a corpus of interest that contained the transitions between depressive and non-depressive states. The control documents are being used to provide the context and word frequencies that would be found in this corpus of interest. The author documents are then scored using the SVM classifier and scored with BM25 from the query formed from the classifier results. In a non-synthetic example, the representations scoring above a threshold and having a transition between depressive and non-depressive representations would be extracted; here, the synthetic documents are identified as having these and extracted for analysis. The synthetic documents are included in **Appendix Table 2** for review.

For both authors, an SVM was used to maximally separate the transitions from depressive to non-depressive representations. The words associated with each direction were extracted. An SVM was also used to separate the combination of the depressive representations from the non-depressive representations and the words associated with the depressive direction extracted. This concludes the methodology of the case study; we now present the results.

5.4.2 Synthetic Case Study Results

When training the SVM to separate the /r/Depression author representations from our control set authors, the SVM achieved an accuracy of 86.4% on training and 81.3% on the test set. The SVM was retrained on the entire set and achieved an accuracy of 86.2%. This suggests there is a signal by which the two representations can be separated; without this sign, it is not recommended to proceed to further stages of ATLAS.

The magnitudes from the cosine analysis of this are shown in **Figure 15**. Briefly, the alignment of any individual word is less in this analysis than in previous studies, suggesting each individual word does not contribute strongly to the archetypal label. This is commented on in the discussion. The 400 most significant words as determined by cosine magnitude with the decision direction are shown in **Figure 16**. The amount of 400 is chosen as it is the most that the information retrieval system used, Elasticsearch, can handle in the query field.

For the synthetic author representation, we show the words associated with the first depressive case and the transition to the non-depressive case in **Figure 17**. The words associated with the second depressive case and transition to non-depressive case are shown in **Figure 18**. Lastly, the word associations when comparing both depressive representations against both non-depressive representations is in **Figures 19 and 20** for the depressive and non-depressive words, respectively. The interpretation of the words in these figures is subjective, which transitions us to the case study discussion.

5.4.3 Synthetic Case Study Discussion

When previously performing ABMS to examine words with opioid usage, the magnitudes of the cosine of each word vector with the decision direction was also examined. In that case, the magnitudes went as high as 0.6, suggesting that some of the words were 3x more associated with the archetypal case than in the depressive analysis. Hopefully, this makes sense – depression is a complex mental phenomenon, whereas discussing opioids involves, at some point, referring to a noun that is highly associated with the topic. Fentanyl, for example, is not often mentioned when discussing a trip to the pub. Practically, we suggest this means the query scoring approach for the topic of depression benefits from the inclusion of more words to capture more of the distributed association across many words. The only limitation is the number of words that the information retrieval system can handle in a single query, which we found to be approximately 400 in our case.

The words that are associated with the depressive case can be complex to interpret. Some of the words can be more obvious ‘creep’ seems self-explanatory, loneliness is sensible, deaths and foul language (i.e., ‘shit’) can also make some sense when trying to imagine someone that is not in a good state of mind. Others, it requires more of a leap of logic. This could include the video games mentioned, such as Minecraft and Fortnite, which are popular games played by many people across the globe. The context in which these words are used would help illumine the association they have to a depressive representation, and unfortunately this can be difficult to convey without showing direct posts. While entirely speculative, we suggest that from examining the word cloud in **Figure 16**, some of the common topics that are driving individuals to believe they are depressed can be extracted. There are words associated with family, dating, school, incarceration, politics, drug activity, and finances, for example.

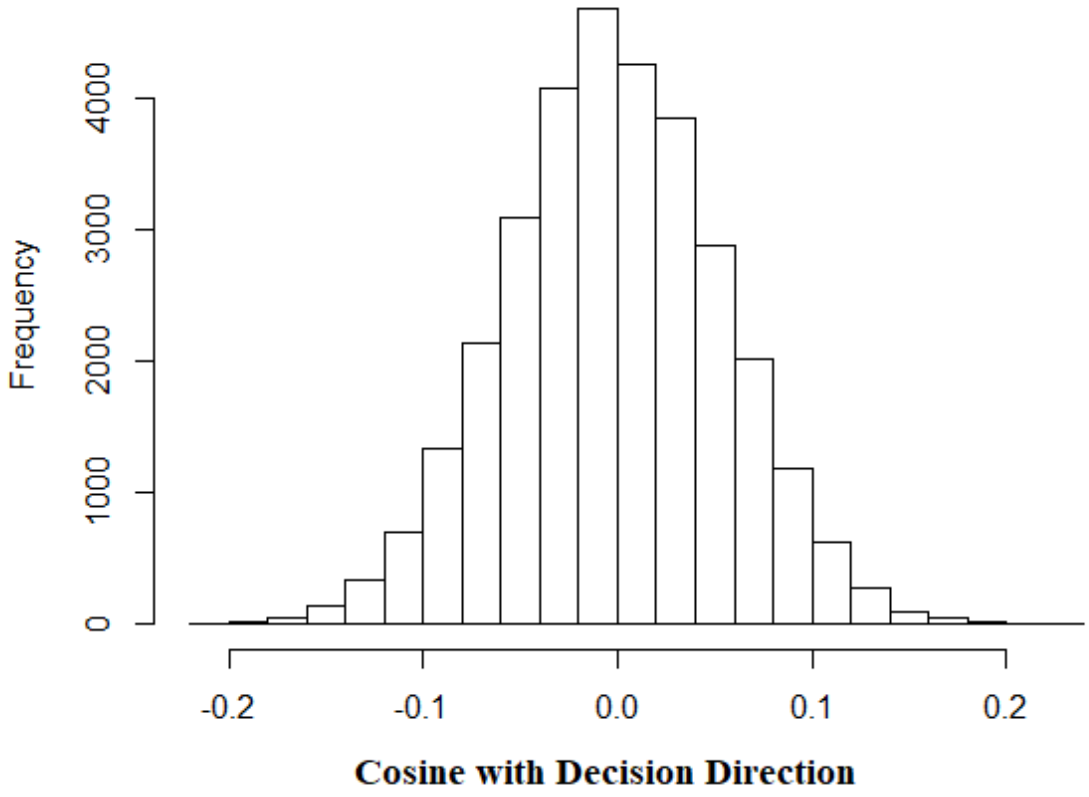


Figure 15. Histogram of cosine magnitudes from comparing word vector cosine similarity to the separating hyperplane between depressive and non-depressive author representations. The majority of words have no association, while some are positive aligned, and some are negatively aligned.

In **Figure 17 and 18**, we see how well ATLAS can perform in a relatively low data scenario. Each author representation was trained on 10 documents containing a short amount of text each. From this, there is still evidence of the topics that were discussed in the method section reappearing in the depressive case. One of the limitations that is evident from these two figures is that, as there is only one case to learn from for each label, there is some unexpected vocabulary in the non-depressive scoring side. However, looking at the control documents made, this language does appear; it was typically about the synthetic author discussing their problems and their improvement for them, but this nuance is not captured as no part of this particular system goes into a deeper understanding of the linguistic intent behind the words and their association; this could be a promising avenue for future work.

Figure 16. Word cloud of the words most associated with the depressive representations trained from Reddit authors active in /r/Depression, as determined by cosine similarity of the corresponding word vector with the SVM trained decision direction. Word size is proportional to the magnitude of the cosine with the decision direction.

In **Figure 19 and Figure 20**, we see that when the two representations are combined – so a total sample size of 4, with two depressive and two non-depressive representations. This comes as little surprise, as these techniques are not meant to be working in the case of separating one representation from another – even though SVMs are useful for this as maximum margin classifiers. In **Figure 19**, we see the words that we expect from the synthetic narrative. 'CS' is a short form for another common game (counter strike), and we see Facebook emerge as part of the social media. In **Figure 20**, we see some of the non-depressive narrative emerging as well. Exercise can be seen, as well se other positive behaviours. Interestingly, alcohol does not make it into the word cloud; this may be because part of the synthetic document involves a college story about a roommate feeding a house plant beer, and the plant surprisingly thriving. One part of **Figure 20** that bears mentioning is the number of depressive words like insomnia – the synthetic document had the individual talking about recovering from them, making this association reasonable to the learner but less so to human interpretation.

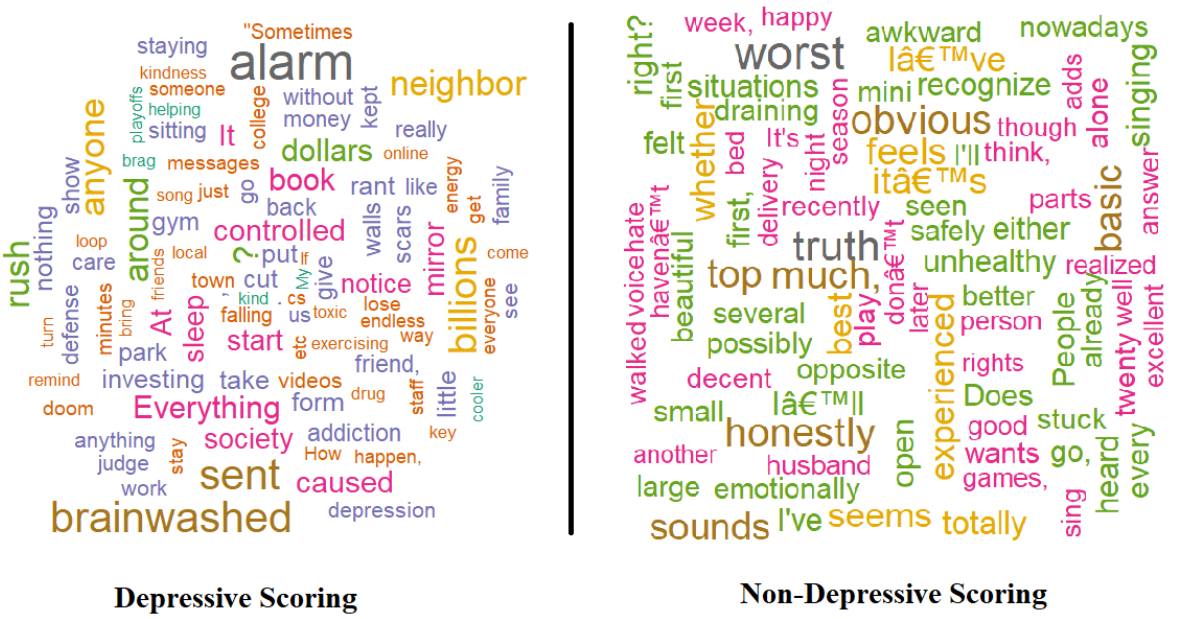


Figure 17. Vocabulary associated with the depressive representation and non-depressive representation of the synthetic author in the case study presented here.

We conclude this case study with the suggestion that ATLAS performs well even in low data environments, giving it great utility in sifting social networks for transitions between representation states like this one. Increasing the explainability and fidelity are both desirable, but any enhancements to these are out of scope for this case study and are good focal points for future work. The results here show that ATLAS has promise in extracting keywords that are useful for experts wanting to understand the transition between states, and would otherwise not be able to review collections of documents from social media that are too large except in the case of dedicated teams over long periods of time.

5.5 Discussion

ABMS and AIR can identify individuals who match an archetypal behaviour; ATLAS addresses the task of finding changes in these behaviours over time. The key to the task is adding in a temporal measure that is sensible and allows for the identification of individuals who have experienced the desired change. In the methods section, we briefly describe how ABMS or AIR can be used to identify candidate individuals that have changed from archetypal to non-associated or even opposed behaviours. This labeling could be harmful if there is stigma associated with the behaviours and could be used to amplify stereotypical bias; pre-existing stereotypical linguistic bias persisting is well established [24]. The same potential concerns exist in ABMS and AIR, and ethical concerns and cautions are identified in each algorithm, but those methods are focused on providing high level summaries of the search.

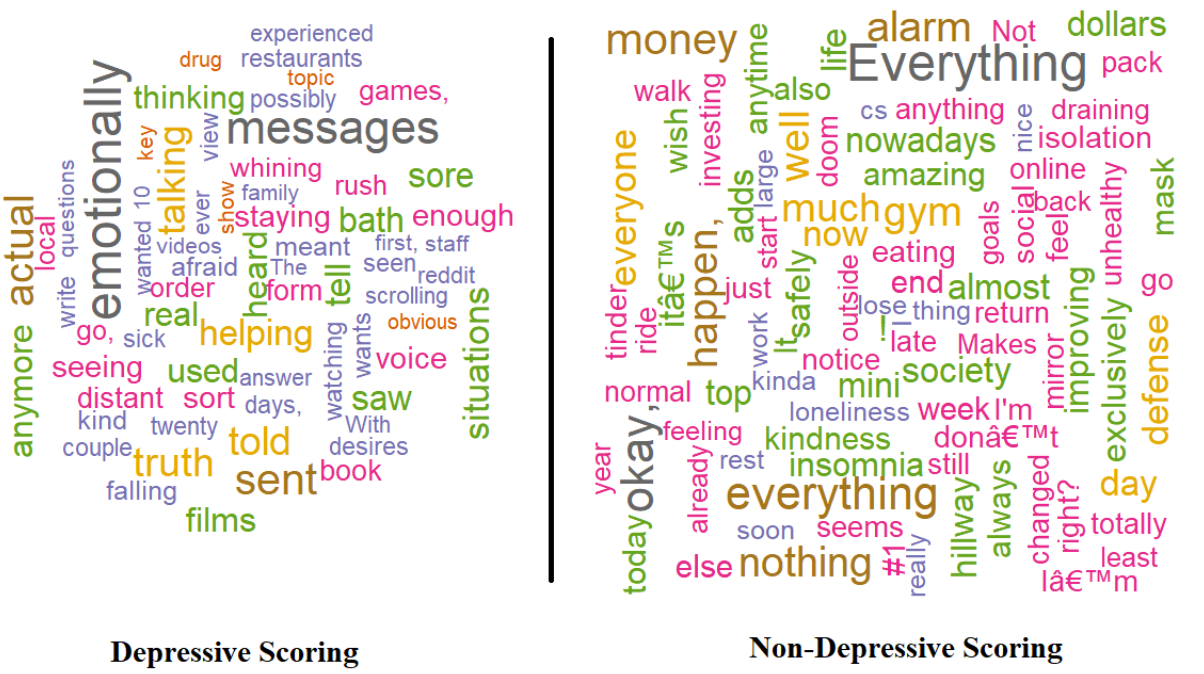


Figure 18. Vocabulary associated with the depressive representation and non-depressive representation of the synthetic author in the case study presented here.

In contrast, ATLAS is an automated way of surveying, analyzing, and assisting the interpretation of, individual behaviours. Owing to this, we suggest ATLAS requires consideration of the social impact before using it at scale on real data. As an algorithm that can take in generic input, it can be used for several purposes. Some of these could be beneficial to society, such as understanding behaviours of addiction through social media, understanding resilience from those who have broken addictive behaviours, or monitoring decreases in diversity of vocabulary that are potentially associated with conditions of cognitive decline. The gradual decay of vocabulary is observed in Alzheimer's, meaning this could be a passive surveillance that assists in early detection [25]. However, it could also be used nefariously to measure the effectiveness of targeted online ad campaigns at changing an individual's political viewpoints (or even to see if certain ads or messaging are able to depress them, make them more anxious, or worse).

ATLAS may be applied to many different problems. As it stands, many organizations would be able to use a technique like ATLAS without ethical oversight or approval. As the data is freely available on social media – repositories like Pushshift contain a trove of such information, and Twitter has made all of its historical data available for academic research – there is potential for misuse even within academic applications [26].

One of the complications of this is that individuals have posted this data on a globally connected medium, with the intent that anyone who wishes could go see, comment, or otherwise interact with their posts.

Figure 19. Words associated with the depressive representations of the synthetic documents included in the case study

It is also plausible to say that many of these individuals never thought that a system like ATLAS would use them for an experiment – in the most benign case – that is more familiar to the realm of social science than computer science; a line that is blurred with some new applications [27].

It is our suggestion that ethical approval be sought through formal procedures even if not required owing to the nature of the technique. However, we also realize that ethical approval can be a lengthy process, and is not guaranteed success, we first detail some of the technical limitations to ensure it is worth pursuing approval before doing so.

Figure 20. Words associated with the non-depressive representations of the synthetic documents included in the case study.

5.5.1 Limitations

ATLAS, like many machine learning techniques, is data hungry. Both ABMS and AIR, its predecessors, made use of approximately 36,000 authors for their studies. For ATLAS, the training stage to find the relevant vocabulary made use of 224,855 authors. The total representation training time consumed approximately 35 CPU years. This was without generating any temporally stratified representations – to do the same process again for all representations would double the processing time, meaning this analysis would have consumed a total of 70 CPU years. We were very fortunate to have the support of both Compute Canada and CIFAR for work on understanding depression on social media to support out computational needs, but this will not be the case for every research group.

Further, the time stratification stage also relies on those authors being consistently active over a longer period. This is true for some authors, but not all. Some behaviours may even be consistent with an absence from social media – such as breaking a social media addiction. While author representations can be trained on low amounts of posts, their ability to accurately model an author's digital presence increases with more data. As one of the best ways to confirm the success or failure of ATLAS's modeling is to manually review these posts, a low number of posts makes it more difficult to confirm the presence or absence of the desired behaviour. The synthetic case study shows that something potentially interesting can be discovered even with a low amount of data, but it certainly would be helpful to have more. Even going from a sample size of 2 to 4 showed considerable improvement in the relevance of the words, and unexpected vocabulary in either set at low sample sizes can knock out otherwise pertinent behaviours.

ATLAS is limited by its ability to be transparent in the decisions being made at the algorithmic level. Providing too much information about what ATLAS is doing can quickly cause those assessing the algorithm to be overwhelmed, while insufficient amounts can cause the primary assessment to be made on personal opinions and suffer from the resulting echo chambers. A balance has to be struck between simply interpretable and overwhelmingly complex.

The last limitation we highlight is the presence of labelled data. While repositories like PushShift have radically changed the amount of data that is available, there are still many topics that will not have an associated subreddit – a division of the social media site Reddit which is specialized around a given topic, for example, Depression or Opioid usage in /r/Depression or /r/Opiates, respectively. If the behaviour does not have a corresponding subreddit that is enriched in relevant posts or authors with matching the behaviour, even if there is an abundance of unlabeled controls, it is not advisable to proceed with ATLAS. It may be possible to infer labels in a semi-supervised manner, but this is left to the discretion of individual practitioners and experts in semi-supervised learning.

5.5.2 Ethical Considerations

As it stands, in many places, using this technique for any given research purpose would not require an extensive consultation with an ethical team, if any. This precedence often stems from an established view that there is no expectation of privacy on social media. This has been challenged a few times but never overturned – there is no reasonable expectation of privacy for someone that is posting on social media in many jurisdictions [28],[29]. Of the many ethical considerations when using social media data, this concept of privacy for social media content can be one of the most perplexing. Content is shared on social media with the intent of being seen; many people purposefully try to get as much attention or following on social media as possible.

However, with new techniques that can change the kinds of information which can be extracted from these, we suggest extreme caution; if we are being overcautious, we hope that the ethical community will provide better guidance. We are certainly not Ethicists.

One suggestion from us is that ATLAS ventures into what can be considered a naturalistic observation study; this is a study where researchers go out into a public forum and observe unaware participants in order to see what their behaviour is like when they are not being observed [30]. When these studies are run, they do have formal ethical considerations to protect the participants. Further, they require approval which is contingent upon some expectation of professional behaviour and oversight on the study itself. We suggest that analyzing a change in someone's behaviour and extracting potentially related keywords has something in common with these naturalistic studies, and the ethical consideration merits consideration at least as deep as the learning itself.

5.5.3 Future Work

Despite these challenges, there are applications for which ATLAS could be beneficial. The study of vocabulary decline in Parkinson's disease is one such application, and we describe a scenario that would hopefully satisfy an ethics board. First, the data collection could benefit from social media for the language learning step, but any actual analysis of the trends and habit of the persons is more personal and should not be included without prior consent. Reaching out to people through social media to ask for consent is often unsuccessful, but has been done, and could identify multiple participants. It is also possible that several people with the disease would be willing to volunteer their social media accounts; all platforms have a way for the individuals to download their data these days, and depending on the frequency of their activity, this could provide the necessary amounts of time labelled data for ATLAS to function. This would have to be done in coordination with clinics that study and treat Parkinson's disease, but this is sensible to evaluate the changes in vocabulary and see if they have a clinical interpretation to begin with.

ATLAS would benefit from a more comprehensive analysis of how accuracy changes the results. While the partially synthetic case study here made a study of the details of the accuracy (i.e., false positive, false negative) unhelpful, future work will include these and how they influence the overall results.

The output of such a project in the long term, to have some use beyond the investigation of the decline in vocabulary, could potentially be an application that monitors for this decline over time and alerts the individual with the application as an early warning system. To avoid the perils of a central repository which could suffer a breach and expose this data, individuals who have reason to believe they are predisposed or at risk for the condition could install it themselves and benefit from passive

surveillance. Updates, relevant vocabulary, and classification models could be imported from a central hub.

As a last note, we highlight that an understanding of the complexities of ATLAS makes a great case for the development of visual analytics and interactivity applications. Each component of ATLAS is individually complex, and owing to its modularity, the most effective way of generating a given representation will change as methods improve. A way of visualizing the relationships between words, between authors, and between authors ATLAS has identified as undergoing these changes will increase the interpretability of these systems and move it away from the so called 'black box' that many similar applications exist in.

5.6 Bibliography

[1] B. D. Davis, K. Sedig, and D. J. Lizotte, "Archetype-Based Modeling and Search of Social Media," *Big Data Cogn. Comput.*, vol. 3, no. 3, p. 44, Jul. 2019.

[2] G. Marchionini, "Exploratory search," *Commun. ACM*, vol. 49, no. 4, p. 41, Apr. 2006.

[3] C. Gormley and Z. Tong, *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc.," 2015.

[4] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends® Inf. Retr.*, vol. 3, no. 4, pp. 333–389, Dec. 2010.

[5] G. Kitagawa, *Introduction to Time Series Modeling with Applications in R*. CRC press, 2020.

[6] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, no. 1, pp. 11–24, Jun. 2014.

[7] R. Kass and T. Finin, "Modeling the User in Natural Language Systems," *Comput. Linguist.*, vol. 14, no. 3, pp. 5–22, Sep. 1988.

[8] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[9] S. Amir, G. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace, "Quantifying Mental Health from Social Media with Neural User Embeddings," *Machine Learning for Healthcare Conference*, Nov 2017.

[10] G. D. Rosin, E. Adar, and K. Radinsky, “Learning Word Relatedness over Time,” *Proceedings of the 2017 Conference on Empircal Methods in Natural Language Processing* Jul. 2017.

[11] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics – Challenges in topic discovery, data collection, and data preparation,” *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, Apr. 2018.

[12] B. Plank, “What to do about non-standard (or non-canonical) language in NLP,” *arXiv* 2016.

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics* Jul. 2016.

[14] C. E. Sims, S. M. Schilling, and E. Colunga, “Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals,” *Front. Psychol.*, vol. 4, no. NOV, p. 871, Nov. 2013.

[15] G. Baxter and W. Croft, “Modeling language change across the lifespan: Individual trajectories in community change,” *Lang. Var. Change*, vol. 28, no. 2, pp. 129–173, Jul. 2016.

[16] H. A. Schwartz *et al.*, “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,” *PLoS One*, vol. 8, no. 9, p. e73791, Sep. 2013.

[17] R. Vessey, “Zappavigna, M. (2012). Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. London: Bloomsbury,” Springer, Cham, 2015, pp. 295–299.

[18] A. Morrison, “Social, Media, Life Writing,” in *Research Methodologies for Auto/biography Studies*, New York, NY: Routledge, 2019.: Routledge, 2019, pp. 41–48.

[19] J. Pearl, “Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution,” Jan. 2018.

[20] B. Athiwaratkun, A. G. Wilson, and A. Anandkumar, “Probabilistic FastText for Multi-Sense Word Embeddings,” *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 1–11, Jun. 2018.

[21] S. Amir, B. C. Wallace, H. Lyu, P. Carvalho, and M. J. Silva, “Modelling Context with User Embeddings for Sarcasm Detection in Social Media,” *arXiv* 2016.

[22] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks Editor,” Kluwer Academic Publishers, 1995.

[23] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. W. Maintainer, “The e1071 Package,” 2005.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” *arXiv*, Aug. 2019.

[25] E. Aramaki, S. Shikata, M. Miyabe, and A. Kinoshita, “Vocabulary Size in Speech May Be an Early Indicator of Cognitive Impairment,” *PLoS One*, vol. 11, no. 5, p. e0155195, May 2016.

[26] J. M. Baumgartner, “Pushshift API.” 2018.

[27] J. Grimmer, “We are all social scientists now: How big data, machine learning, and causal inference work together,” in *PS - Political Science and Politics*, 2014, vol. 48, no. 1, pp. 80–83.

[28] B. Mund, “Social Media Searches and the Reasonable Expectation of Privacy,” *Yale J. Law Technol.*, vol. 19, 2017.

[29] M. A. Moreno, N. Goniu, P. S. Moreno, and D. Diekema, “Ethics of social media research: common concerns and practical considerations,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 16, no. 9, pp. 708–713, 2013.

[30] M. Gorup, “Ethics of Observational Research,” in *Handbook of Research Ethics and Scientific Integrity*, Springer International Publishing, 2020, pp. 475–491.

6 Conclusion

This dissertation has established three new algorithms for exploratory document search. They are well suited to analysis of individual's behaviour of on social media. First, in the pursuit of identifying opioid drug abuse related content on social media, we created Archetype-based Modeling and Search (ABMS) to handle the task of searching social media for documents with unknown vocabulary. Documents could be displayed as normal for a search technique or counted for the purpose of assessing the number of documents matching a given profile. Existing techniques were specialized around look-up-based keyword searches, and had limited or no ability to adapt to an environment where the vocabulary was evolving, heavy on slang, and largely unknown. ABMS was successful in supporting the exploratory search task, but the amount of computation required to process new corpuses made it prohibitive to apply at scale.

This motivated the development of the second algorithm, Archetype-based Information Retrieval (AIR), which was a modification of ABMS. This modification drastically cut down on the amount of computation to search compared to ABMS, but retained its ability to find documents with unknown vocabulary and present it to the end user to review. This technique was used to identify opioid drug usage related content in /r/Ottawa, and an estimate of the overall prevalence of the opioid drug user archetype estimated.

The third algorithm in this dissertation, Archetype-based Temporal Language Adaptative Stratification (ATLAS), combined the scoring methods of both ABMS and AIR to provide a more adaptive scoring, and has practical utility when examining some archetypal behaviours by considering their temporal aspect. A case study used Reddit content to establish the ATLAS model, and synthetic data to show its utility.

This chapter, which serves as a conclusion of the dissertation, is divided into three sections: 1) a review of the chapters and some of the corresponding contributions, 2) the general contributions this dissertation makes to the body of scientific literature and 3) potential areas for future research.

6.1 Dissertation Summary

In Chapter 3, we presented ABMS, an algorithm which takes in a set of archetypal and control documents and then produces a machine learning model which can be used for exploratory search of a corpus with unlabeled documents to identify similar ones to the archetypal case. This algorithm used new notions of similarity made possible by first creating vector representations of word, which capture a soft notion of semantic similarity by using word co-occurrence statistics to group similar words together into an approximated ontology of related words and concepts. The utility of the system was shown in finding candidate documents in a social media forum of interest to partners at Public Health Ottawa.

In Chapter 4, we described an extension to ABMS, AIR, which was an adaptation of the same core idea of using vector representations of words aggregated in the form of author representations to find word similarities that may be unknown, but modified so that new documents did not have to have a representation calculated for them. This was done by using some of constituent machine learning algorithm properties to allow the extraction of the most impactful words in distinguishing the author representations from each other. This had two major impacts; first, the keywords extracted could be used in a search engine which used the state-of-the-art approaches from information retrieval, which the direct classification from ABMS lacked; and second, there was no longer a need to compute a representation for every unknown document, which was a non-trivial computational task. This system was used to assess the same social media forum of relevance to Public Health Ottawa, and an estimate of the number of individuals who matched the opioid drug usage profile we developed given.

In Chapter 5, we described an algorithm that combines both the information retrieval scoring system from AIR and the direct machine learning scoring system from ABMS in a system called ATLAS. ATLAS is further distinguished from the other two in that it considers the temporality of archetypal behaviours as they are represented on social media or in other document collections. For example, an opioid problem or a depressive episode can have a temporal component – people can begin with the behaviour, recover or display resilience, and cycle between the two states potentially *ad infinitum*. With this in mind, ATLAS provides an automated way to extend the searching capabilities of ABMS and AIR to a temporal context, and provides insight into the words which are associated with the transition between states. This was demonstrated by building a model that was able to identify depressive text, and tested on synthetic data.

6.2 General Contributions

The broadest application of this research is to the task of exploratory search. As a research problem, exploratory search is complex, novel, and intriguing because it must deal with the uncertainty of information completeness. Whenever the results of an exploratory search are being discussed, it must deal with the fact that the ground truth is unknown; test sets and synthetic samples can help, but their ability to generalize to a truly unknown case is limited. For example, when examining opioid dialogue, a very limited subset of people will know most of the slang related to illicit opioid drug usage. Even in these cases, they may not compose a keyword-based search that would retrieve all the relevant documents to what they are looking for. In the case of much social media, there exists no resource that completely annotates every post made by every author.

For example, to manually curate every post made by every author in in /r/Opiates would be a gargantuan task owing to the amount of time that would have to be spent reviewing each post and then attempting to compare similarities. ABMS, AIR, and ATLAS, provide a way to explore these datasets which are too large for manual review

without a significant expenditure of time manually reviewing. ABMS provides a method which focuses on comparing words by their vector representation, which allows for more complex interactions comparisons than with the Boolean operators of a lookup search. AIR provides a way to both explain how ABMS works, by finding the keywords which best explain its decision to classify a document as belonging to an archetypal group or not, and by providing these keywords to be used in concert with information retrieval heuristics such as BM25.

ATLAS takes the process of analyzing authors by whether they are belonging to an archetypal group or not by adding the ability to separate the representations by a given temporality. In the case of some archetypes that are of interest to public health surveillance, these behaviours may only exist for a given period or may go through cycles. For example, individuals may go through periods of depression that may or may not have a cyclical behaviour. ATLAS provides an automated way to find these transitions for further analysis. This offloads the cognition of separating an individual into discrete time periods for analysis and provides an automated way to look for periods of archetypal behaviour.

When considering which algorithm to use, ABMS is best suited for when non-temporal searching is being performed in an environment where computational expense and explainability are not as important. If either explainability or compute time is a concern, AIR is better suited. It is always possible to use both and compare. ATLAS is the algorithm of choice for looking for changes over time, and the decision whether to use exclusively ABMS or AIR is informed as normal. Using both is once again optimal. For a given topic, it will be necessary to compare the results from ABMS or AIR to determine which is more optimal. The accuracy of the SVM to separate the author representations gives an indicator of performance; low accuracy suggests there is not sufficient information to extract archetypal information. Using a histogram of word vector similarities by cosine or other similarity metrics will elucidate the importance of given language; opioid vocabulary was more distinct and gave highest magnitude values in the vicinity of 0.4, while depressive language had less unique words and gave values around 0.2. Lower values indicate that a lot of collective dialogue is needed to distinguish, while higher values suggest there is fairly unique language. This fairly unique language leads into the idea of sublanguage.

These three algorithms collectively contribute to the idea of sublanguage [1]. Sublanguage refers to the language used by specific populations that is exclusive or has meaning exclusive to them. Slang and memetics, considered so, are examples of sub language. The concept used here, Archetype, is superordinate to the idea of sublanguage. All Archetypes have the potential for sublanguage, but Archetypes also consider the idea of increased or decreased frequency of language, topical collections, and are united behind more abstract ideas of archetypes. Sublanguages are typically associated with a population. Frequently, this population can be the basis of an archetype. By using AIR, it is possible to identify candidate sublanguage which can later be verified.

6.3 Limitations and Future Work

Circumscribing the totality of topics like opioid drug usage and depression requires more complexity than the methods of learning word and author representations can capture at present. This thesis demonstrates the utility of going and finding strong trends which can be useful for analyzing massive collections of text in a more sophisticated manner.

To say that these methods are an example of a complete exploratory search is not possible. There is almost always a way to suggest that there were unseen behaviours in the unknown corpuses being considered that would not be captured by these techniques. The techniques are ultimately trying to model an emergent phenomenon of human behaviour as it is captured through their writings on social media. Describing the totality of what depression is and is not, is outside the scope of this, although the techniques must model approximations of it.

When considering social media, there are many complexities which limit the effectiveness of ABMS, AIR, and ATLAS. Humans do things like lie on the internet, misrepresent themselves, harass one another, and purposefully try to wreak havoc. Certainly, there will be examples in the corpuses we used where individuals were talking about opioid drug usage in /r/CasualConversation, and there can be little doubt that depressed individuals on Reddit go to /r/CasualConversation and exist there too. The success of the techniques relies in trying to find datasets which are enriched relative to each other, and then trying to find the language that explains that enrichment. There is no assumption of causation here, which is a strong limit of the techniques ability to generalize beyond the kinds of ways demonstrated in this dissertation.

Future work will consider how to compare the results of these algorithms against that of formal ontologies. For many specialized topics, there is either not a formal ontology created, or it has to be repurposed to a more adapted task. A pharmaceutical ontology of opioid drugs, for example, may not generalize well to the behaviours found by those using opioid drugs. Comparison to semantic web data and formal ontologies will allow for a stronger comparison, and highlight the strengths and weaknesses of the algorithms described. As an extension of this, it poses an interesting research question how different social media transfers to other social media contexts. Data in this dissertation is retrieved from Reddit and has been applied to Twitter, but norms, posting limits and other factors make a detailed study of differences worthwhile.

As these algorithms facilitate the retrieval of a considerable amount of new data, it has also been considered to develop guidelines for usage that other partners and practitioners would be amenable to, and which strike the right balance between careful usage and providing a public good. Consultation and agreement with guidelines from the Association of Internet Researchers would be a sensible starting point for this effort [2].

For this technique to be more accessible to researchers outside of technical disciplines such as computer and data science, it would be beneficial to create an interface which assisted the steps in the algorithm. This will allow the techniques described to be applied to a larger set of questions that may be beneficial to researchers from social science. The trio of algorithms described have utility in several public health domains, where it could be used to further study any mental disorders which have communities of sufferers providing data about the lived experience of it online. Regarding ATLAS, it would be interesting to see if causal reasoning could be applied between the sets of representations to describe archetypal transitions in a casual framework to better our understanding of these individual's lived experience. These algorithms, functioning as intended, are meant to be providers of information that can assist the decisions of those who are best poised to help.

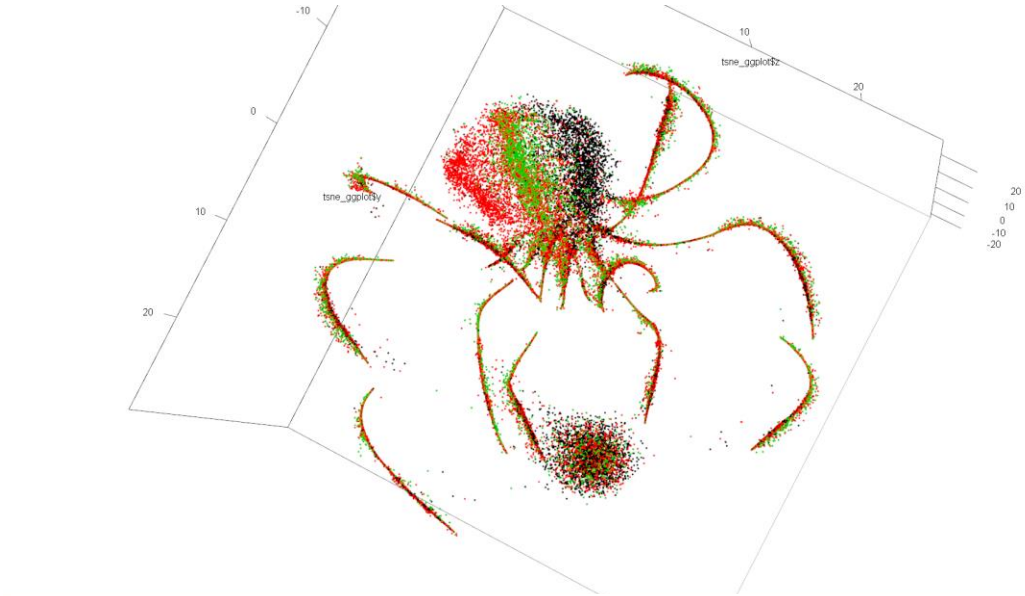
6.4 Bibliography

[1] K Richard, and J Lehrberger. "Sublanguage". *de Gruyter*, 1982.

[2] A Markham, and E Buchanan. "Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee". *Association of Internet Researchers*. Dec 2012. Retrieved from: https://aoir.org/reports/ethics2.pdf

fin

Appendices



Appendix Figure 21. This is a *t*-SNE Visualization of the author representations used in Chapters 3 and 4 of this thesis. The red coloured dots are Reddit authors from the subreddits /r/Opiates, the black coloured ones from /r/CasualConversation, and the green ones from /r/Ottawa.

Tables:

Appendix Table 2. Top ten posts returned by an **Archetype-based Information Retrieval (AIR)** query, and top ten results of a 200-word query expansion on the term 'opioids'. Posts are missing punctuation due to text cleaning prior to their indexing.

Top 10 posts retrieved by AIR for posts made by /r/Ottawa participants	
Rank	Query: Top 200 AIR Identified Words
1	"Hard drugs seized by police in the bust included fentanyl meth MDMA morphine bath salts hydromorpone heroin cocaine ketamine and xanax "

2	"Grade 9 was just drinking and smoking weed Grade 10 got addicted to speed and did molly shrooms acid ketamine xanax morphine oxys coke regularly but recreationally Grade 11 speed addiction turned into meth addiction Continued recreational drug use and tried a shit ton more chems DMT 2CB nBOME PCP 4FA Valium whatever Opiate and benzo use became more regular and drinking became a problem Grade 12 quit drinking and started heroin ODed my first time IVing and quit for a while Lot more recreational drugs Quit meth the summer after high school fell back into heroin addiction first semester of college"
3	" Dissociative is a subclassification of Hallucinogen There are three subclasses Psychedelic Hallucinogen 5 HT2aR agonists like Psilocybin and LSD Dissociative Hallucinogen NMDA receptor antagonists like PCP and Ketamine kappa opioid receptor agonists like Salvinorin A in salvia and Deliriant Hallucinogen mAChR antagonists like Muscimol the active alkaloid in Amanita muscaria and scopolamine found in things like Datura and Deadly Night Shade "
4	"Xanax is a benzodiazepine not an opioid Some addictive quality but nowhere near opioids "
5	"What The bupe is what is causing that because of its half life it just sits on your receptors for days after your last dose provided you are a daily user of more than a few days Its sitting there and will block any other opiate unless its powerful enough like hydromorphine IV to cut through the bupe Naloxone is strictly in suboxone to discourage IV injection of suboxone strips Ingested sublingual the naloxone is instantly rendered useless as the amount in a sub film isnt enough to have any real effect on your receptors "
6	"Isn t Narcan an opioid blocker They give kits away for free here in Ontario to anyone to help fight the epidemic with Fentanyl being used to cut coke and heroin "
7	"it s fent i used to get shit like that sometimes it was pink it s like 95 cutting agent and a bit of fentanyl or maybe even carfentanil watch out"
8	"Unfortunately not all pressers think as far ahead or have as much common sense as you just demonstrated I mean look at the fentanyl crisis You d think dealers wouldn t want to damage their cred and sell proper Heroin diacetylmorphine but instead to make a little more short term they put fent in it when long term it puts off customers from ever coming back while even killing other customers But yeah ur right In a common sense world you d think we wouldn t have to even debate or question what s in these pressed bars Thank god with these ones I have right now I feel pretty confident it s alprazolam or at least a short acting benzo Which is what I want Feel sorry for those who got the fake xanax presses without anything in them aka chalk

	taste and does nothing Feel even worse and more sorry for those who got the batch with phenazepam or fentanyl in them Eek "
9	This is an automatic summary original http www sfgate com news science article China makes deadly opioid carfentanil a 10935986 php reduced by 79 gt Beijing is banning carfentanil and three similar drugs as of March 1 China 039 s Ministry of Public Security said Thursday closing a major regulatory loophole in the fight to end America 039 s opioid epidemic gt China said the March 1 ban will also apply to carfentanil 039 s less potent cousins furanyl fentanyl acryl fentanyl and valeryl fentanyl gt Legally used as an anesthetic for elephants and other large animals carfentanil burst into the North American drug supply last summer causing hundreds of unsuspecting drug users to overdose gt Though Beijing has said U S assertions that China is the top source of fentanyls lack evidence the two countries have deepened cooperation as the U S opioid epidemic intensifies gt That same month China began evaluating whether to ban carfentanil and the three other drugs gt quot One news I just got is that the carfentanil and furanyl fentanyl etc opioid analogs will be controlled in China on March 1 effective quot one vendor called Ete wrote in an email Summary Source http smmry com http www sfgate com news science article China makes deadly opioid carfentanil a 10935986 php FAQ http www reddit com r autotldr comments 31b9fm faq autotldr bot Version 1 65 139588 summaries so far Theory http www reddit com r autotldr comments 31bfht theory autotldr concept Feedback http www reddit com message compose to 23autotldr PM s and comments are monitored constructive feedback is welcome Top five keywords Drug 1 China 2 fentanyl 3 carfentanil 4 U S 5 Post found in r news http np reddit com r news comments 5unuhi china carfentanil ban a gamechanger for opioid NOTICE This thread is for discussing the submission topic Please do not discuss the concept of the autotldr http www reddit com user autotldr bot here
10	"It s a dissociative hallucinogen and a very unique one at that Hitting only the kappa opioid receptor which is the yang to mu opioid receptors that pain killers hit yin in that it reliably causes dysphoria The opioid system seems like its the evolutionary this stimulus good That stimulus bad It s a bizarre one because every other dissociative hallucinogen that I can think of is active via the PCP mechanism aka NMDA antagonism like PCP ketamine and probably nitrous I wouldn t say I enjoy salvia but I certainly find low doses interesting There s a unique body sensation that comes with doing it To me it feels like your body is being intersected by an infinite number of infinitely thin sheets of metal not that in a painful way but in a cold and detached kind of way Very hard to describe LSD psilocybin mescaline etc are classical psychedelic hallucinogens all of which seem to share the main Serotonin 2a receptor agonism Interestingly the reason that LSD lasts so long is that it binds to the 5 HT2a receptor in such a way that it actually folds part of the receptor over itself trapping it inside The only way the neuron can clear the

	receptor is by taking the whole receptor back into the body of the neuron and destroying it a process that takes about 12hrs "
Rank	Query: 200 most similar words to 'opioids' by word vector cosine similarity using /r/Opiates pretrained GloVe embedding
1	"That s not really true about benzos being like alcohol in a pill And saying that isn t really an answer to why junkies aka opiate users like benzos lol smh Junkies probably like benzos because they take the edge off Something which most downers have in common Benzos are also quite useful when it comes to dealing with opiate withdrawal Imo if I had to pick one medication to deal with opiate withdrawal it would be a benzo or clonodine Oxy is similar to heroin yes as both are opiates opiate agonists When you compare the two they feel act very similar on the body apart from maybe the duration of their effects However the same can t be said when comparing alcohol and booze Alcohol isn t just primarily a GABA agonist though like benzos are It also targets many other receptors They are both depressants affect some of the same GABA receptors but that s it in terms of similarities "
2	"There are a lot of groups of drugs and they re not mutually exclusive Drugs can be classified by their effects upper downer or chemical composition opioid benzodiazepine Here s a quick overview of the main ones off the top of my head Stimulants are drugs that stimulate your CNS causing euphoria energy and focus This is a broad class of drugs classified by their effects stimulation such as cocaine meth or MDMA These drugs act on dopamine serotonin and norepinephrine to cause these effects Inside this class are smaller groups of drugs classified by their chemical composition such as amphetamines Meth Adderall Vyvanse Most stimulants including cocaine and amphetamines act primarily on dopamine and norepinephrine Stimulants like MDMA or mephedrone that cause strong feelings of attachment to people empathy are called empathogens or entactogens These drugs are not necessarily chemically related they just share similar effects They cause these effects mainly by releasing serotonin Depressants are drugs that depress the central nervous system classified by effects Examples of these are alcohol GHB and three main chemical groups opioids benzodiazepines and barbiturates barbiturates are no longer common Natural opioids like morphine and codeine are found in the opium poppy Semi synthetic opioids like oxycodone or hydromorphone are synthesized from opium Synthetic opioids like fentanyl can be made without opium They all have very similar effects sedation relaxation and euphoria They are also notorious for being extremely addictive Benzodiazepines are almost exclusively found as prescription anti anxiety medications or street pressed replications of those pills eg Xanax Valium They cause drowsiness alcohol like intoxication and suppress anxiety very effectively They are also extremely addicting Psychedelics are a large group of drugs that alter our perception of the senses usually mainly by

	distorting our vision Common examples are LSD and psilocybin mushrooms They also strongly alter our headspace and thought process These drugs share similar effects but not necessarily chemical composition There are however several chemical groups of psychedelics Tryptamines are a group of chemicals including psilocybin mushrooms DMT mescaline and others Many of them are naturally occurring Their visuals are very organic feeling and the headspace is strong Ergolines are drugs like LSD and it s many derivatives ALD 52 1p LSD that are derived from the ergot fungus The visual effects are often described as more digital mathematical or electric The headspace tends to be lighter than tryptamines Phenethlamines are synthetic psychedelics often with stimulating effects Visual effects and headspace can vary wildly between phenethlamines Dissociatives are the last main group of recreational drugs They all share dissociating effects a feeling of disconnection between the mind and body Most dissociatives including ketamine and PCP are arylcyclohexylamines chemical group Dissociatives are hard to explain if you ve never done them so I won t go into it I ve left a few things out but I think it s pretty good for a meth rant done on my phone "
3	"From what I m reading on the internet fentanyl is an opioid medication Same family as heroin morphine and hydromorphone As mentioned the dosage required for an effect is extremely small and is normally used in health care for extreme pain According to Wikipedia it s 50 100 times more potent than morphine I think dealers are cutting it into their heroin opioids since it has similar effects but requires a way smaller dose So you could sell super strong heroin with less heroin stronger effects and more profit "
4	"Unless you OD and get brain damage or do something self injurious then you re fine Opioids don t kill brain cells the only temporarily change the way the brain works Codeine is a relatively weak opioid compared to other more power opioids such as hydrocodone Vicodin oxycodone morphine and most notoriously heroin The oxycodone is a stronger opioid with more potent addictive effects It s withdrawal symptoms are consequently more unpleasant You are walking a dangerous path and I urge you to exercise extreme caution As for your cognitive performance you re most likely suffering from PAWS post acute withdrawal syndrome Your opioid habits have caused your brain to change in order to become more accustomed to your use of drugs Because you used those drugs so often your brain chemistry changed to adapt to the use of the drugs Since you re no longer using the drugs you re brain is in disarray and you feel like shit as a result With time your brain will return to equilibrium that is the state it was before you used drugs The longer you abused drugs the longer it will take to get better Do yourself a favor and please engage in thorough harm reduction research before you abuse any substance At your young age the brain is still developing and can be quite vulnerable If you re really worried talk to a doctor or psychiatrist and be honest You have the rest of your life ahead of you Drugs are no joke You can die or permanently harm yourself from

	them The way many young people treat drugs so nonchalantly and with little regard for their formidable potency is troubling Edit I d worry more about your oxy habit Don t touch that shit at your age Smoke weed with your friends "
5	"Great question Research points to a link between prescription drug abuse particularly prescription painkillers like oxycodone morphine hydromorphone to the eventual use of heroin One US study indicates that four out of five heroin users started using opioids in prescription drug form Over prescribing of prescription drugs has been well documented in North America Canada and the United States consume 80 of the world s prescription painkillers Prior to the arrival of fentanyl we were already in the grips of a prescription drug crisis With many thousands of users addicted to prescription drugs and turning to heroin when their supply of prescription drugs is curtailed fentanyl s arrival created a perfect storm It entered both the prescription drug market in the form of counterfeit pills and in the growing heroin market Kids steal drugs from their parents medicine cabinet If they re stealing powerful prescription painkillers that s a potential gateway to heroin use and addiction "
6	" gt Doctors should not prescribe medication with harmful physical side effects merely because their patients want it Define want People with depression want antidepressants These medications like almost every other medication often have harmful side effects Should doctors also stop providing antidepressants in all but the most extreme circumstances "
7	"Over the past few years two problems have arisen regarding opioids and opiates both in the USA and in my home country of Canada The first problem more prevalent in the USA is the overprescription of opiates for pain management which has caused countless deaths and addictions This problem is still going on today as hundreds of millions of opiate prescriptions are filled every year in the USA alone The second problem which has become huge in Canada and I believe also the USA has been the lacing of street drugs particularly heroin with extremely potent analogues such as fentanyl and carfentanil Despite mainly being added to heroin as a cut fentanyl has turned up pressed into benzo pills ecstasy pills as well as found in cocaine and meth The widespread fentanyl problem has caused countless overdoses through north america and personally I haven t taken any drugs besides weed in ages and would only take them in the future if I had a reliable test kit Naloxone kits are everywhere now in Canada and the news is always focused on some other kid with tons of potential who s life was cut short by an OD My question to you guys is where do you see both of these issues going What do you think would be the logical conclusion to both of these issues Also a bonus question What do you think should be done to combat each of these problems "

8	"When you say narcotics are you including all the prescription opioids out there Doctors are partly to blame pharm companies push their products and in the end so many who never intended to become addicts eventually do Many chronic pain sufferers are turning to marijuana for more efficient pain management with less side effects and no chance of overdosing from Psychology Today Contrary to popular mythology prescription drugs are more lethal than illegal or street drugs Prescription drug abuse and addiction kill far more people in the U S every year than all illegal drugs combined The unprecedented rise in overdose deaths in the U S parallels a 300 percent increase since 1999 in the sale of powerful painkillers such as Vicodin and OxyContin These drugs were involved in 14 800 overdose deaths in 2008 more than cocaine and heroin combined 2 "
9 (2 in ABMS set)	"Grade 9 was just drinking and smoking weed Grade 10 got addicted to speed and did molly shrooms acid ketamine xanax morphine oxys coke regularly but recreationally Grade 11 speed addiction turned into meth addiction Continued recreational drug use and tried a shit ton more chems DMT 2CB nBOME PCP 4FA Valium whatever Opiate and benzo use became more regular and drinking became a problem Grade 12 quit drinking and started heroin ODed my first time IVing and quit for a while Lot more recreational drugs Quit meth the summer after high school fell back into heroin addiction first semester of college"
10	"yeah actually benzos valium xanax klonopin are given to alcoholics to manage their withdrawals some people combine benzos and alcohol though and their withdrawals are really bad this is in comparison to opiates opiods oxycodone heroin vicodin where the withdrawals while extremely unpleasant almost rarely if ever kill someone I think the most likely way to die from opioid withdrawals is from suicide "

Appendix Table 2. Synthetic documents used to train depressive to non-depressive author representations for the analyses described in Chapter 5.

synthauth1 You ever have one of those days where you look at the clock and realize you spent the entire day Minecraft? Lol this is totally an addiction .
synthauth1 One of the worst parts about insomnia for me is the isolation and loneliness that come with being the only person awake . Makes me feel so distant from society
synthauth1 Omg what a day! You ever have one of those days where it seems like the world is out to get you? I don't know what's in the air today but I've just had enough of it
synthauth1 Go leafs go! Looking forward to seeing how the rest of the season plays out
synthauth1 How deep is your deepest Youtube dive? I've been so deep into it that I think the lizard people are being controlled by floridaman . Is this what it's like to go crazy? Maybe I already was .
synthauth1 It feels like everyone else around me is brainwashed into doing exactly what society wants them to do at all times . You ever feel like you don't have a place? Maybe this is what it's like to actually be woke lol
synthauth1 Getting up every day and doing the same thing over and over without any change is really draining . Is this what life is now? Go on Youtube, go on Twitter, play some games, read some news, stay up too late at night watching bad films .
synthauth1 It used to be flipping through channels, now it's flipping through Minecraft, fortnite, cs go, until the morning alarm comes and it's time to haul my sorry fucking ass to work

synthauth1 I wish I knew the first thing about investing so I could get out of the 9-5 wage drone life . I don't see things improving anytime soon though synthauth1 Isn't it kind of haunting that there's billions of dollars going to the richest in our economy while there's protests for basic human rights going on . I'm sorry to be doing so much whining but all of this doom scrolling has me falling into a pit . Maybe I need to cut back on the Twitter for awhile .
synthauth2 It's amazing what a couple of nights of good sleep can do . I feel so much better and refreshed . Is this what normal people feel like? synthauth2 Went on a walk in the park today and checked out some of the local foodtrucks . Thirty minutes of walking means mini donuts are okay, right? synthauth2 I never realized how much money I was spending on delivery until recently . It really adds up! Taking the twenty or thirty minutes to walk to one of the local places has been really rewarding for me . synthauth2 I can hear my neighbor singing through the walls again . I am really happy about it because I haven’t heard her sing in a long time . She has a beautiful voice . Tonight it sounds like her husband is joining her as well . synthauth2 Running on the treadmill, if I felt like I had to stop Because my legs were getting sore and feeling shaken, should I have kept pushing? I would slow it down to a stop . When they open again I’m gonna go back . Looking forward to it synthauth2 Get on a pushbike . Chose a hill and cycle up it as fast as you possibly can until you feel you've had a good workout . Stating "I feel the need . . . the need for speed" at the start is optional

synthauth2 I’m buying a bike soon to ride around town a few times a week . I saw some people on reddit who do statewide traveling on their bike and bring a tent with them and that sounds awesome so I’m gonna do that eventually too .

synthauth2 When I used to work out, I would meditate first, never after . I’m really curious about this—so for at least a few days, it changed my life

synthauth2 I used to never eat out alone because I was afraid the staff would see me as some lonely weirdo, but now that I've done it a few times from necessity, I absolutely love it! Usually, I'll pack a book or my laptop, find exactly where I want to go and order exactly what I want to eat .

synthauth2 Sometimes, usually when i'm standing in front of a mirror i get a weird feeling like i'm actually a living being . Was wondering If you guys have experienced this as well .

synthauth3 So I come from a very toxic family and it’s caused me to to turn inward which I thought was the key to my problems . I have also cling to very unhealthy relationships which made me turn inward even more . My #1 problem has been food . Not eating too much, but eating too little . Sometimes I just don’t have the will to eat, and eating alone is not very motivating .

synthauth3 I know I'm responsible for myself and all that, but the truth is humans were not meant to be alone . I see it on the opposite end where people develop other unhealthy coping mechanisms for loneliness like overeating, promiscuity, spending addictions, drug and alcohol addictions etc .

synthauth3 I hate how everything is online nowadays . I don’t want to meet people through Facebook or tinder or Snapchat etc . What are some good ways to meet cool people in person by yourself? I’m not super outgoing but I’m not at all awkward either . I go to the gym but I hardly have long conversations with strangers there . I don’t go

out by myself but sometimes I want to . . . but I've seen other people go out alone and they were just kinda sitting there alone . . . I feel like that'd be me .

synthauth3 Due to staying at home for nearly a year because of covid, I sort of just, don't give a shit anymore . I rant on my irl account, sometimes over 10 posts a day for the same topic, and I just don't care whether people are gonna judge me or not anymore

synthauth3 I feel like I'm living in a Bon Jovi song . . . How does it go? "Sometimes you tell the day by the bottle that you drink" . At least it's helping me sleep

synthauth3 I finally found the energy to get out of bed and take a bath today . It's been 6 days . May not be a huge win but it's a win and I'll take it!

synthauth3 I have no idea who i am anymore . I have no goals . I have no desires . Yet i am still afraid to die, so i'm still stuck in this endless loop of nothingness . Does anyone feel the same? I don't even know what i should do anymore

synthauth3 I did it . I spent an entire day watching Youtube videos . I don't even remember what they were about . Everything is kind of of condensing into a slow drone of nothingness .

synthauth3 Why does no one notice I'm depressed? How much more fucking obvious can I be? Do I need to show everyone my scars so they recognize depression is an actual thing and no one's making it up? Do I need to write a fucking book with every fucking thought I've had so they can fucking understand I'm fucking sick and dying inside?

synthauth3 So many questions are flooding my head . It's like I hear one and before I can answer it another interrupts me before I can finish that thought, and before I know, I'm drowning in questions . I've been told to make time to think, but honestly the moment I stop to I start to lose it

synthauth4 Hi all, I just wanted to quickly remind us that a small act of kindness will make you feel better . I've just sent a few messages to friends that I haven't been talking to in awhile .

synthauth4 In the last week, I've gone out to restaurants . I've walked the streets safely without a mask . Heck, last night I even considered making love . But nothing has signaled a return to normal more than Kadri being suspended for the playoffs .

synthauth4 It's all because we were in a funk for most of April . Now we're out of the funk and back to our normal selves . People are starting to remember that we have a decent defense group, great top 6, and excellent goaltending and depth .

synthauth4 I had a large fern in college that was watered exclusively with stale coffee and leftover beer dregs . It flourished for a year until we put it outside to get some sunlight and real water . Was dead a week later .

synthauth4 I don't care about being cuddled or human warmth . All I care about is someone being there for me emotionally . I honestly don't care if it's in real life or online . But hey, that's just me .

synthauth4 There's so many great people out there who don't realize there's somebody that loves them, more than an acquaintance, more than a friend, more than a best friend, more than anything

synthauth4 It's almost like night and day . Things just seem more normal . I feel like I have control over what I'm doing again, rather than being a passenger to my own life . Is this what being normal is like?

synthauth4 I actually love making people feel happy as well! We should form a squad of person helping people! We will call them the anti sadness squad! With both our collective power we should be able to make thousands happy!

synthauth4 I'm always in my own head, whether I'm daydreaming being way cooler than I am in situations that would never happen, or thinking deeply about some philosophical topic till I'm convinced my view is superior . Looking back on several social situations I think I brag too much, and I've been told so by friends .

synthauth4 It's nice to be getting back into the habbit of exercising again . The endorphin rush is a welcome change from those days where I didn't want to do anything .

Curriculum Vitae

Name: Brent Douglas Davis

Post-secondary Education and Degrees:

Western University
London, Ontario, Canada
2011-2014 HBSc. Biochemistry of Immunity and Infection

Western University
London, Ontario, Canada
2014-2015 Diploma in Computer Science

Western University
London, Ontario, Canada
2015-2021 Ph.D in Computer Science

Honours and Awards:

Province of Ontario Graduate Scholarship
2019-2020

Mitacs Accelerate
2019-2020

Teaching Assistant Award, Western University
2019-2020

Related Work Experience

Teaching Assistant
Western University
2015-2020

Lecturer
Western University
2018-2021

Lecturer
Huron University College
2020-2020

Publications:

(Cited Pre-print) Exautomate: A user-friendly tool for region-based rare variant association analysis (RVAA) BD Davis, JS Dron, JF Robinson, RA Hegele, DJ Lizotte BioRxiv, 649368

(Cited Pre-Print) Decision-Directed Data Decomposition. BD Davis, E Jackson, DJ Lizotte. arXiv preprint arXiv:1909.08159

Archetype-based modeling and search of social media. BD Davis, K Sedig, DJ Lizotte. Big Data and Cognitive Computing 3 (3), 44

Loss-of-Function CREB3L3 Variants in Patients With Severe Hypertriglyceridemia. Jacqueline S Dron, Allison A Dilliott, Arden Lawson, Adam D McIntyre, Brent D Davis, Jian Wang, Henian Cao, Irina Movsesyan, Mary J Malloy, Clive R Pullinger, John P Kane, Robert A Hegele. Arteriosclerosis, Thrombosis, and Vascular Biology 40 (8), 1935-1941.

The influence of depression-PTSD comorbidity on health-related quality of life in treatment-seeking veterans. C Forchuk, A Nazarov, R Hunt, B Davis, K St. Cyr, JD Richardson. European Journal of Psychotraumatology 11 (1), 1748460