UNDERSTANDING AND DIRECTING WHAT MODELS LEARN

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

LAURE JEAN THOMPSON

December 2020

Chapter 5 is based on the article "Computational Cut-Ups: The Influence of Dada" by Laure Thompson and David Mimno published in the *Journal of Modern Periodicals*, Vol. 8, No. 2, 2017. This article is used by permission of The Pennsylvania State University Press. The copyright notice for this material is as follows:

UNDERSTANDING AND DIRECTING WHAT MODELS LEARN

LAURE JEAN THOMPSON

Cornell University, 2020

Machine learning and statistical methods, such as unsupervised semantic models, make massive cultural heritage collections more explorable and analyzable. These models capture many underlying patterns of raw textual and visual materials, but neither model creators nor model users fully understand which specific patterns are learned by a given model nor under what conditions a particular pattern becomes more learnable. In this dissertation I address two core questions (i) what do models *actually* learn? and (ii) how can we direct *what* they learn? Instead of proposing new models, I focus on expanding the affordances, as well as our understanding, of existing ones that are used by scholars in the humanities and social sciences. In the first part of this dissertation, I study what models learn by way of expanding the ways in which they can be used. In the second part, I investigate how existing models can be directed *away* from known, uninteresting structures via corpus- and representation-level interventions. Throughout this work, I show how machine learning and statistical methods provide an opportunity to view collections from alien, defamiliarized perspectives that can call into question the boundaries of established categories. Likewise, I show how the uses of computational methods within humanities and social science scholarship can test, challenge, and expand the affordances of these methods. Ultimately, this dissertation highlights some of the many ways in which machine learning and the humanities help one another.

BIOGRAPHICAL SKETCH

Laure Thompson was born and raised in Renton, Washington. She attended the University of Washington where she discovered that computer science is full of interesting puzzles and that studying the ancient Mediterranean world is an interest worth cultivating. In the process of learning all the things, she graduated in 2013 with a B.S., *cum laude*, in Electrical Engineering and a B.S., *cum laude*, in Computer Science with minors in Classical Studies and Mathematics. Laure began her Ph.D. in Computer Science at Cornell University in Fall 2013. After several years studying theoretical programming languages and security, she found her research niche in the areas of natural language processing and machine learning with David Mimno as her advisor. Through this switch she found a way to productively combine her interests in computer science, classics, archaeology, and the humanities more broadly. She completed a minor in classical archaeology under the supervision of Caitlín Eilís Barrett; a minor that became unexpectedly relevant to this dissertation. In Fall 2020, Laure joined the College of Information and Computer Sciences at the University of Massachusetts Amherst as an Assistant Professor. And now, despite the challenges of a global pandemic, has finally written a dissertation.

ACKNOWLEDGMENTS

This Ph.D. has been a long, winding, and unexpected, but rewarding journey. Although the end looks far from what I had imagined some seven years ago, I am grateful for where it has taken me. None of this would have been possible without the many people who have supported me along the way. Apologies in advanced for all those unmentioned but not unappreciated.

First, I want to thank my advisor, David Mimno, for taking a chance on a fourth year Ph.D. student with no background in natural language processing or machine learning but a common interest in the Graeco-Roman world. Thank you for introducing me to the field of digital humanities and helping me navigate its intersections with computing. I am grateful for both your unending support and patience with my many projects that spanned across numerous domains and long lengths of time. I feel incredibly lucky to have you as my advisor and mentor.

Next, I would like to thank the other members of my committee: David Bamman, Caitlín Eilís Barrett, and Dexter Kozen. Thank you for your support and feedback that have greatly improved this dissertation. Thank you also to Dexter and Caitlín for tolerating the many changes that my academic plans underwent. Dexter, while this dissertation has little to do with programming languages, I hope you nonetheless enjoyed being on this committee. Caitlín, thank you for agreeing to be my minor advisor in classical archaeology long before my doctoral work had anything to do with classical archaeology. David, I hope you enjoyed being on a Cornell Ph.D. committee even though you were never able to join in person.

Additionally, I want to thank my many friends at Cornell and beyond that have supported me through my Ph.D. I especially want to thank Anna Gommerstadt and Anna Waymack. Without your friendships, I could never have completed this Ph.D. I also want to specifically thank my lab mates Maria Antoniak, Jack Hessel, Moontae Lee, Xanda Schofield, Melanie Walsh, and Greg Yauney. While I have only co-authored papers with a few of you, I have appreciated our many insightful discussions, some about research and some not.

Finally, I would like to thank my parents, David and Mary Thompson, for their unending love and support. Thank you for cheering me on through all of these years and only asking me occasionally "When are you going to graduate?" Who knew that my joking answer of 2020 would be the right one.

My Ph.D. work has been supported in part by the National Science Foundation through a Graduate Research Fellowship (DGE-1144153) and two grants (#1526155 and #1652536); and through resources provided by the HathiTrust Digital Library and the HathiTrust Research Center.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

The rise of digitization and the ubiquity of the World Wide Web have brought many exciting changes to humanistic scholarship. More texts and artifacts are available to scholars and the wider public than ever before. Digital collections, such as the HathiTrust Digital Library[1] and the Campbell Bonner Magical Gems Database,[2] not only allow broader access to materials, but also enable new ways of viewing and comparing objects and texts that may never coexist in physical space. Additionally, "born" digital materials, such as fanfiction [Milli and Bamman, 2016; Porter, 2018], online book reviews [Boot, 2017; Bourrier and Thelwall, 2020], and social media posts [Antoniak et al., 2019; Walsh, 2018], provide scholars with new ways to study a wide range of social and cultural phenomena that are less visible in extant, non-digital mediums.

But access is only part of the potential. These digital collections provide the opportunity to study materials at massive scales. These scales allow scholars to ask new questions about collection-level characteristics: How well does the well-studied canon differ from the wider set of materials? Are established concepts and categories clearly seen at these larger scales? How might these properties change over time or other contextual boundaries represented within these larger collections? These new lines of inquiry provide the opportunity to shift focus from the small, well-studied canon to the broader range of materials that have been less-studied or more commonly well-studied individually but not comparatively.

[1] https://www.hathitrust.org/

[2] http://classics.mfab.hu/talismans/

However, many of these digital catalogues are used much like their physical predecessors. Scholars use these sources to more easily gather the materials they would like to closely study; at which point, their research proceeds as it otherwise would with physical texts or objects sans the physical-material experiences that are not captured by digital reproduction. Fundamentally, the number of objects being studied is limited by time—there are only so many hours in the day—rather than the availability of materials. This is not the fault of scholars, but rather evidence of the lack of available methods to work with these collections at scale.

To fill this gap, scholars in the digital humanities and social sciences have adopted computational methods from statistics, machine learning, and natural language processing. These methods organize large collections into lower dimensional vector spaces such that the collections can be more easily manipulated and explored. These representations encode a document or artifact as a vector—a list of k numbers—that corresponds to a single point within a k-dimensional space. The distance between two vectors reflects the similarity of their corresponding documents or artifacts, but this similarity is dependent on what aspects of an item—features—that the particular model is capturing. With the right model, it is possible for scholars to operationalize their questions at scale.

But these computational methods were not designed with these scholars or their collections in mind. This fact does not prevent these methods from being useful tools to humanists and social scientists, but it adds additional complexities and barriers to their adoption and usability. There are benefits and perils in being removed from the original development of a method. While it is difficult to be aware of all the conditions necessary for a model to run successfully, there is a freedom from the assumptions of what a model *can* be used for. Tools can have many uses and naturally some are far removed from their original intended purpose. In this case, scholars, as users, are testing the affordances of models and fortunately (or

unfortunately) testing model limitations as well as the assumptions imposed by model creators.

However, even experts do not fully understand *what* underlying patterns these representations learn nor *which* patterns they are most likely to learn. This makes it difficult to know *when* these methods are applicable to a particular research question or corpora. If we knew which structures are learnable, it would be easier to know *when* a model is useful. Moreover, models are capable of learning many different structures, but not every structure is useful for a given line of scholarly inquiry. While it might be useful to organize texts by author, learning this structure is seldom useful when already known and can be problematic if it is mischaracterized as a cross-cutting pattern. Therefore it is not only important to identify what structures learn, but ways of directly influencing what models learn.

In this dissertation, I focus on better understanding what models learn and how we can directly influence what they learn. Instead of proposing new models, I instead focus on how we might expand the affordances, as well as our understanding, of existing models. By focusing on established models—ones already being used by scholars in the humanities and social sciences—I am intentionally prioritizing accessibility and usability. My goal is to improve current working practices by providing methods that can be easily adopted with minimal disruption to existing processes. To that end, I focus on methods that are simple, intuitive, and transparent that are compatible, but independent from the existing models in use.

This dissertation is organized into two core parts focusing respectively on the following questions:

1. What do models *actually* learn?

2. How can we direct *what* they learn?

Within each part, I cover both textual and visual material collections and correspondingly text- and image-based models. Each part is organized such that the chapters related to textual collections and models precede chapters related to visual collections and image-based models. Before diving into these core themes, I first build the necessary context in Chapter 2. I frame the general problem setting used throughout this dissertation and provide an overview of the various text- and image-based models that will be used within later chapters.

In Part I, I examine what models learn by focusing on new ways of using existing models. These new affordances expand our understanding of what these established models can and will learn. Chapter 3 shows how token-based contextual vector representations can be used to form clusters similar to the topics of topic models. Chapter 4 introduces new methods for studying how linguistic properties are spatially captured by type-level word embeddings. These general purpose methods can be used both for studying language and model-level variation. Switching to the visual domain, Chapter 5 proposes a methodological framework for adapting machine learning into a working tool for scholars of visual material. Features extracted from pretrained neural image models are used to computationally explore what makes the art movement Dada Dada.

In Part II, I explore how we might direct what models learn by examining corpus-level and representation-level interventions. These interventions focus on directing models *away* from known, unwanted structures rather than dictating *what* a model should learn *a priori*. After all, it is often easier to identify what we already know—and we find uninteresting—rather than the unknown structures that are possible to learn. These chapters rely heavily on humanities collections and the needs of these collections. The complexities and idiosyncrasies of humanities collections tend to "break" established models, or more accurately their (implicit) assumptions. Chapter 6 identifies and addresses the problem of topic models learning authorial

signals rather than cross-cutting themes. This issue is not restricted to humanities collections, such as science fiction novels, since the resulting topics for many text collections heavily correlate with known contexts (e.g. author, region, date). Rather, this issue tends to surface more often and problematically for digital humanists. Chapter 7 addresses similar issues for the image domain. It seeks to improve the framework proposed in Chapter 5 such that the image model representations (i.e. vectors) and our study of these representations do not depend on unwanted structure. This work focuses on producing more useful vectors spaces for studying magical gems, an art historic category of engraved gemstones from the Graeco-Roman world, that are not dominated by the medium (photograph, impression, drawing) of each image.

Finally, in Chapter 8, I reflect on how the humanities and machine learning truly help one another. While the argument that machine learning can help the humanities might be seen as an easy one, since such methods are already being used by scholars as research and pedagogical tools, this dissertation aims to highlight the untapped potential which can further enable scholars to study collections more easily and from new, alien perspectives. On the other hand, the converse is less apparent but just as strong. The many chapters in this dissertation demonstrate how humanities collections test and expand what machine learning models are capable of. I describe future directions of this dissertation and how these directions help pave a path to a (continued) symbiotic future for machine learning and the humanities.

2

BACKGROUND

Before we can begin to ask what models learn, we need a common frame of reference for what I mean by "models" and how we will use them. For the purposes of this work, we will consider a model to be a process that projects a collection of texts or images into an embedding space: a k-dimensional vector space. Such embeddings can be explicit output of a model, or implicitly extracted from a trained model's internal structure (e.g. penultimate layer of a neural network). We require that the distances within embedding spaces—typically Euclidean or directional (i.e. cosine distance)—reflect the similarity between the represented objects. The aspects of similarity preserved by an embedding space vary by model and are the learned structures of interest; the structures we want to identify and direct. While some models (e.g. topic models) have interpretable dimensions, many do not.

Generally, we will treat models as fixed processes, effectively black boxes. Our attention is not directed at the model itself, but rather its resulting embedding space. After all, we are interested in understanding *existing* models rather than building new ones. Like many users, we will use well-supported, off-the-shelf models and direct our attention to the input and output of such models.

2.1 USING VECTOR SPACES

Machine learning methods provide an opportunity to organize and study collections at scale, but also to view familiar collections from unfamiliar perspectives. In

this section, I will outline two general use cases of machine learning for humanities collections and scholarship.

EXPLORATION. By design, vector spaces provide more efficient organization of large collections that rely solely on the input corpora and not additional metadata. These organizations make exploration explicitly spatial. This allows for item-level exploration wherein the most similar (and dissimilar) items can be quantitatively identified by their respective vector distances. But we can also lift exploration to collection-wide viewings by studying how items collectively cluster within the space. We can quantitatively identify distinct groupings using established algorithms such as k-means and agglomerative clustering.

We can also visualize our entire collection in two- or three-dimensional projections by applying dimensionality reduction methods, such as t-SNE [Maaten and Hinton, 2008] and UMAP [McInnes et al., 2018], to a collection's corresponding set of vectors. These reduced representations provide a simplified but viewable map of a collection. Note that in these visualizations relative but not absolute distances matter. Points that distinctively cluster within such a plot are indicative of prominent learned structures within the underlying embedding space. However, these indications are more intuitive than exact. It is worth comparing the clusters evident in these much lower dimensional projections with the item-level groupings identified by clustering algorithms applied to the original embedding.

With these exploratory processes, relationships between unconnected objects can be made visible. Many of these connections will be simplistic and naive, after all the underlying model is unaware of the broader contexts and histories of each item in the collection; however, this does not inherently reduce their value. Unlikely comparisons can lead to new scholastic insights. Obvious and simplistic patterns

can be markers of more complex and interesting phenomena. Embeddings are a tool to aid in scholarship, not to replace scholarship.

RE-EXAMINING CATEGORIES. While vector spaces are useful tools for organizing unlabelled data, they are also valuable for studying labelled data. Embeddings provide the opportunity to view established categories from alien, defamiliarized perspectives. From these decontextualized vantage points, we can recharacterize categories through their presence (and absence) within data-driven organizations.

We can study categories by augmenting our exploration process with added categorical labels. By coloring, or otherwise distinctively marking, each point in our two- and three-dimensional visualizations, we can see which categories are well-delineated and which are not. That is, which categories (if any) are prominently captured by the underlying model, and which groupings may be indistinguishable from one another. To support and expand the intuitions provided by these projections, we can compare the relative overlap between established categories and the groupings identified by clustering algorithms.

However, these exploratory tactics will only give us a partial understanding of how well these embedding spaces capture categories. Categories can be captured by an embedding but not necessarily dominate its geometric organization. Instead, we will use classifiers as a test for whether an embedding space can be used to identify a category. In natural language processing, this process is known as a *probing task* [Conneau et al., 2018].

But fundamentally we're not interested in accurately replicating a category, but rather examining and deconstructing one. Instead of solely focusing on a classifier's accuracy, we are instead interested in characterizing its successes and failures. Since many classifiers come with corresponding confidence scores, we can reexamine a known category from the items most correctly and incorrectly associated with a

particular label type. Perhaps paradoxically, we are testing a category's definitions by directly using its predefined labelings. What makes our process critical of a categorization is that we are not relying on the category's original selection process, but building a new characterization from a classifier's most confident predictions. This method will be used and described in greater detail in Chapters 5 and 7.

2.2 TEXT-BASED MODELS

Intuitively it is very easy to convert texts into numerical representations. Documents can be decomposed into salient components such as phrases, words, and morphemes. All textual models fundamentally rely on this discretization. The three model types I focus on within this dissertation—topic models, word embeddings, and contextual word embeddings—also rely on the core tenet of distributional semantics: "You shall know a word by the company it keeps" [Firth, 1957]. Namely, words cooccurring in similar contexts share similar meaning.

TOPIC MODELS. A statistical topic model [Deerwester et al., 1990; Blei et al., 2003] produces a k-dimensional vector space for both documents and words. Its dimensions–called topics—explicitly represent word distributions and because of this are largely interpretable unlike the dimensions of other models we will examine. Topics are typically summarized and interpreted by their top most probable words. For example, a topic with top words *music song sing singing sang play played songs played heard tune* represents a musical discourse, while a topic with top words *computer machine data system work program new information machines human computers* represents a discourse on computer science.

Just as topics are word distributions, documents are topic distributions. From the model's perspective, each word occurrence in a document is generated by sampling a topic t from the document's topic distribution and then sampling a word from t's word distribution. Since documents are explicitly characterized in terms of topics, we can also summarize topics by the documents they most frequently occur in.

Topic models are an established tool within the digital humanities in part because of well-supported implementations. A popular one is MALLET [McCallum, 2002], a Java-based package, that also has a well-supported R wrapper. Topic models have been used to study how topical discourses of a collection have changed over time. For example, Nelson [2010] studies how slavery, particularly fugitive slave ads, are reflected in the American Civil War-era articles of the *Richmond Daily Dispatch* and Barron et al. [2018] study the individuals, institutions, and ideologies of the French Revolution through parliamentary transcripts. Topic modeling has also been used to study whole fields of through their publications [Mimno, 2012; Sandy et al., 2019]. They have also been to recover documents for and about African American women that were otherwise lost or erased from the catalogue [Brown et al., 2019], to study the figurative language of poetry [Rhody, 2012] and to study small corpora of novels from a defamiliarized perspective free from established theories of the novel [Buurma, 2015].

Topic modeling is also amenable to non-textual data. For example, Mimno [2011] applies topic models to household archaeological data. Rooms in Pompeian households replace documents and object types are used in lieu of words. Schmidt [2012] applies topic models to the geodata of digitized ship log books. In this case, a ship's log is a document, geographic locations are words, and resulting topics can be visualized cartographically.

WORD EMBEDDINGS. As bag-of-word models, topic models ignore word order. They associate words by document boundaries alone. In contrast word embeddings trained with the `GloVe` [Pennington et al., 2014] and `word2vec` [Mikolov et al., 2013b; Mikolov et al., 2013a] algorithms use much smaller contexts. They use a sliding window with a fixed word length that produces contexts specific to each word occurrence in a document. As the name suggests, word embeddings encode words into a k-dimensional vector space. They do not inherently encode documents, although extensions exist such as the paragraph vector [Le and Mikolov, 2014].

In producing a k-dimensional vector space, these algorithms build two sets of vectors corresponding to words and their contexts. There is a one-to-one mapping between word and context vectors. While both of these algorithms construct context vectors, only `GloVe` uses them for the resulting word embedding. `GloVe` averages word vectors with their corresponding context vectors, while `word2vec` discards the context vectors entirely.

`word2vec` actually spans two separate models: skip-gram with negative sampling (`SGNS`) and continuous bag of words (`CBOW`). Both models construct their word and context vectors in an iterative, predictive fashion. For each window, both models rely on the center word's word vector w and the context vectors c_i of all other words in the window. `CBOW` uses the context vectors c_i to predict w, while `SGNS` uses w to predict the context vectors c_i. In both cases, w and c_i are pushed closer together. In `SGNS`, the word vector w is further trained through the use of *negative sampling*. In this technique a small set of context vectors c_{neg} are randomly selected and used as negative examples for the prediction task. As a result, w and c_{neg} are pushed further apart.

In natural language processing and machine learning, the primary use of word embeddings is as input for downstream tasks. Through the use of word embeddings as input features have improved performance on many natural language processing

tasks including dependency parsing [Chen and Manning, 2014], named entity recognition [Turian et al., 2010; Chiu and Nichols, 2016], and part-of-speech tagging [Al-Rfou' et al., 2013; Owoputi et al., 2013]. In contrast, scholars in the humanities and social sciences have studied the word embedding spaces directly. By examining how word vectors relate to one another within a word embedding trained on a specific collection, scholars have studied abstract concepts in eighteenth century literature Heuser [2016] and characterizations in popular nineteenth century novels [Grayson et al., 2016]. Word embeddings have also been compared to study change in discourse over time [Lange and Futselaar, 2018] and by author [Kerr, 2017].

CONTEXTUAL WORD EMBEDDINGS. A limitation of word embeddings is that each word type is represented by a single, static vector even though words can have multiple meanings such as *crane* representing a bird or a type of construction equipment and *bow* representing the front part of a ship or type of knot. Recently, deep learning models such as ELMo [Peters et al., 2018], BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], and GPT-2 [Radford et al., 2019] have begun to fill this gap. These models are pretrained on extremely large quantities of (English) texts and can be used to generate token-level word vectors that are context-sensitive.

ELMo uses a bidirectional Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] based architecture such that its two learning objectives are (1) predict a word w in sentence s using the series of words that precede w in s and (2) predict w using the series of words that follow w in s. GPT-2, BERT, and RoBERTa all have Transformer [Vaswani et al., 2017] based architectures, but use different learning objectives. GPT-2 predicts a word given the entire series of words that precede it. BERT has two learning objectives: (1) mask a small potion of words (15%) and predict them using the resulting masked text and (2) given two sentence

A and B, predict whether B is the sentence that follows A. RoBERTa is an adaption of BERT that only uses BERT's masking objective.

While these models have different architectures and learning objectives, they produce token-level representations that are transferable to new texts and tasks. Using context word embeddings as drop-in replacements for static word embeddings has yielded performance boosts for a wide range of natural language processing tasks. Contextual word embeddings have not been fully adopted by humanists both because of their relative nascency and the data and computing costs of training these models from scratch [Assael et al., 2019]. That being said, Sims et al. [2019] use contextual word embeddings as model input for detecting events in literature and Bamman et al. [2020] uses contextual word embeddings as model input for coreference resolution in English literature.

2.3 IMAGE-BASED MODELS.

Unlike text, images are more difficult to encode into lower-dimensional vector spaces. Pixels do not capture the same information as words. Luckily, the computational features learned by convolutional neural networks (CNNs) provide a promising direction.

A CNN is an image-processing tool often used in object detection. An image is passed through multiple layers that detect different patterns. The earlier layers detect more primitive forms such as lines and edges, while later layers identify more sophisticated patterns such as faces and books [Zeiler and Fergus, 2014; Mahendran and Vedaldi, 2015; Yosinski et al., 2015]. The final "top" layer produces predictions of likely objects within the image such as toasters, timber wolves, teapots, and trilobites [Deng et al., 2009].[1] The layers directly preceding the final

[1] http://image-net.org/challenges/LSVRC/2010/browse-synsets

layer capture visual features which are both high-level and usable by other image analysis tools [Razavian et al., 2014]. So, we can project an image collection into an embedding space using an off-the-shelf (pretrained) CNN; each image is converted into a vector by feeding the image as input to the CNN and extracting the vector from the penultimate layer (or other non-final layers).

With this general method for vectorizing images, scholars of visual materials are now able to apply quantitative methods that before were limited to text collections and numerical data sets. The Yale Digital Humanities has used extracted features for visualizing and exploring the Meserve-Kunhardt Collection, a collection of over 27,000 nineteenth century photographs.[2] Arnold et al. [2019] use extracted CNN features to study the visual style of two network era sitcoms, *Bewitched* and *I Dream of Jeannie*, while Rodriguez et al. [2020] use them for studying style and authorship in historical artwork. These extracted image vectors have also been used to identify images on social media related to the sales of human remains [Huffer and Graham, 2018], and to identify the style of damaged statuary to aid in restoration efforts [Wang et al., 2017].

[2] See `https://dhlab.yale.edu/neural-neighbors`.

Part I

WHAT DO MODELS ACTUALLY LEARN?

3

TOPIC MODELING WITH CONTEXT EMBEDDING CLUSTERS

This chapter is based on joint work with David Mimno.

3.1 INTRODUCTION

Contextualized word representations such as those produced by BERT [Devlin et al., 2019] have revolutionized natural language processing for a number of structured prediction problems. Recent work has shown that these contextualized representations can support *type-level* semantic clusters [Sia et al., 2020]. In this work we show that *token-level* clustering provides contextualized semantic information equivalent to that recovered by statistical topic models [Blei et al., 2003]. From the perspective of contextualized word representations, this result suggests new directions for semantic analysis using both existing models and new architectures more specifically suited for such analysis. From the perspective of topic modeling, this result implies that transfer learning through contextualized word representations can fill gaps in probabilistic modeling (especially for short documents and small collections) but also suggests new approaches for latent semantic analysis that are more closely tied to mainstream transformer architectures.

Topic modeling is often associated with probabilistic generative models in the machine learning literature, but from the perspective of most actual applications the core benefit of such models is that they provide an interpretable latent space that is grounded in the text of a specific collection. Standard topic modeling algorithms

operate by estimating the assignment of individual tokens to topics, either through a Gibbs sampling state or through parameters of variational distributions. These token-level assignments can then provide disambiguation of tokens based on context, a broad overview of the themes of a corpus, and visualizations of the location of those themes within the corpus [Boyd-Graber et al., 2017].

A related but distinct objective is vocabulary clustering. These methods operate at the level of distinct word types, but have no inherent connection to words in context [e.g. Brown et al., 1992; Arora et al., 2013; Lancichinetti et al., 2015]. Recently, there has also been considerable interest in continuous type-level embeddings such as GloVe [Pennington et al., 2014] and word2vec [Mikolov et al., 2013a; Mikolov et al., 2013b], which can be clustered to form interpretable semantic groups. Although it has not been widely used, the original word2vec distribution includes code for k-means clustering of vectors. Sia et al. [2020] extends this behavior to contextualized embeddings, but does not take advantage of the contextual, token-based nature of such embeddings.

In this work, we demonstrate a new property of contextualized word representations: if you run a simple k-means algorithm on token-level embeddings, the resulting word clusters share similar properties to the output of an LDA model. Traditional topic modeling can be viewed as token clustering. Indeed, a clustering of tokens based on BERT vectors is functionally indistinguishable from a Gibbs sampling state for LDA, which assigns each token to exactly one topic. For topic modeling, clustering is based on local context (the current topic disposition of words in the same document) and on global information (the current topic disposition of other words of the same type). We find that contextualized representations offer similar local and global information, but at a richer and more representationally powerful level.

We argue that pretrained contextualized embeddings provide a simple, reliable method for users to build fine-grained, semantically rich representations of text collections, even with limited local training data. While for this study we restrict our attention to English text, we see no reason contextualized models trained on non-English data [e.g. Martin et al., 2020; Nguyen and Tuan Nguyen, 2020] would not have the same properties. It is important to note, however, that we make no claim that clustering contextualized word representations is the optimal approach in all or even many situations. Rather, our goal is to demonstrate the capabilities of contextualized embeddings for token-level semantic clustering and to offer an additional useful application in cases where models like BERT are already in use.

3.2 RELATED WORK

We selected three contextualized language models based on their general performance and ease of accessibility to practitioners: BERT [Devlin et al., 2019], GPT-2 [Radford et al., 2019], and RoBERTa [Liu et al., 2019]. All three use similar Transformer [Vaswani et al., 2017] based architectures, but their objective functions vary in significant ways. These models are known to encode substantial information about lexical semantics [Petroni et al., 2019; Vulić et al., 2020].

Clustering of *vocabulary-level* embeddings has been shown to produce semantically related word clusters [Sia et al., 2020]. But such embeddings cannot easily account for polysemy or take advantage of local context to disambiguate word senses since each word type is modeled as a single vector. Since these embeddings are not grounded in specific documents, we cannot directly use them to track the presence of thematic clusters in a particular collection. In addition, Sia et al. [2020] find that reweighting their type-level clustering by corpus frequencies is helpful.

Term	Model	Top Words
land	LDA	sea coast Beach Point coastal ***land*** Long Bay m sand beach tide Norfolk shore Ocean Coast areas dunes coastline
		land acres County ha facilities State location property acre cost lot site parking settlers Department Valley
	BERT	arrived arrival landing landed arriving arrive returning settled departed ***land*** leaving sailed arrives assembled
		land property rights estate acres lands territory estates properties farm farmland Land fields acre territories soil
	GPT-2	arrived landed arriving landfall arrive arrives arrival landing ***land*** departed ashore embarked Back sail
		land sea ice forest rock mountain ground sand surface beach ocean soil hill lake snow sediment dunes
metal	LDA	metals ***metal*** potassium sodium + lithium compounds electron ions hydrogen chemical atomic – ion gas atoms
		metal folk bands music genre band debut Metal heavy musicians lyrics instruments acts groups traditional
	BERT	metals elements ***metal*** electron element atomic periodic electrons chemical atoms ions atom compounds ion
		rock dance pop ***metal*** Rock folk jazz punk comedy Dance heavy funk alternative soul street club electronic
	GPT-2	rock pop hop dance ***metal*** folk hip punk jazz B soul funk alternative rap heavy disco electronic reggae gospel
		plutonium hydrogen carbon sodium potassium ***metal*** lithium uranium oxygen diamond radioactive acid

Table 3.1: Automatically selected examples of polysemy in contextual embedding clusters. Clusters containing "land" or "metal" as top words from BERT $L[-1]$, GPT-2 $L[-2]$, and LDA with with $K = 500$. All models capture multiple senses of the noun "metal", but BERT and GPT-2 are better than LDA at capturing the syntactic variation of "land" as a verb and noun.

In contrast, such frequencies are "automatically" accounted for when we operate on the token level. Similarly, clusterings of *sentence-level* embeddings have been shown to produce semantically related document clusters [Aharoni and Goldberg, 2020]. But such models cannot represent topic mixtures or provide an interpretable word-based representation without additional mapping from clusters to documents to words. It is widely known that token-level representations of single word types provide contextual disambiguation. For example, Reif et al. [2019] show an example

distinguishing uses of *die* between the German article, a verb for "perish" and a game piece. We explore this property on the level of whole collections, looking at all word types simultaneously.

There are a number of models that solve the topic model objective directly using contemporary neural network methods [e.g. Srivastava and Sutton, 2016; Miao et al., 2017; Dieng et al., 2020]. There are also a number of neural models that incorporate topic models to improve performance on a variety of tasks [e.g. Chen et al., 2016; Narayan et al., 2018; Wang et al., 2018; Peinelt et al., 2020]. Additionally, BERT has been used for word sense disambiguation [Wiedemann et al., 2019]. In contrast, our goal is not to create hybrid or special-purpose models but to show that simple contextualized embedding clusters support token-level topic analysis *in themselves* with no significant additional modeling. Since our goal is simply to demonstrate this property and not to declare overall "winners", we focus on LDA in empirical comparisons because it is the most widely used and straightforward, highlighting the similarities and differences between contextualized embedding clusters and topics.

3.3 DATA AND METHODS

We use three real-world corpora of varying size, content, and document length: Wikipedia articles (WIKIPEDIA), Supreme Court of the United States legal opinions (SCOTUS), and Amazon product reviews (REVIEWS). We select WIKIPEDIA for its affinity with the training data of the pretrained models. Because its texts are similar to ones the models have already seen, WIKIPEDIA is a "best-case" scenario for our clustering algorithms. If a clustering method performs poorly on WIKIPEDIA, we expect the method to perform poorly in general. In contrast, we select SCOTUS and

REVIEWS for their content variability. Legal opinions tend to be long and contain many technical legal terms, while user-generated product reviews tend to be short and highly variable in content and vocabulary.

WIKIPEDIA. In this collection, documents are Wikipedia articles (excluding headings). We randomly selected 1,000 articles extracted from the raw/character-level training split of Wikitext-103 [Merity et al., 2017]. We largely use the existing tokenization, but recombine internal splits on dot and comma characters but not hyphens so that "Amazon @.@ com" becomes "Amazon.com", "1 @,@ 000" becomes "1,000", and "best @-@ selling" becomes "best - selling".

SCOTUS. In this collection, documents are legal opinions from the Supreme Court of the United States filed from 1980 through 2019.[1] These documents can be very long, but have a regular structure.

REVIEWS. In this collection, documents are Amazon product reviews. For four product categories (Books, Electronics, Movies and TV, CDs and Vinyl), we select 25,000 reviews from category-level dense subsets of Amazon product reviews [He and McAuley, 2016; McAuley et al., 2015].

Corpus	Docs	Types	Tokens
WIKIPEDIA	1.0K	22K	1.2M
SCOTUS	5.3K	58K	10.8M
REVIEWS	100K	52K	9.4M

Table 3.2: Corpus statistics for number of documents, types, and tokens. Document and type counts are listed in thousands (K); token counts are listed in millions (M).

[1] https://www.courtlistener.com/

DATA PREPARATION. For SCOTUS and REVIEWS, we tokenize documents using the spaCy NLP toolkit.[2] Tokens are case-sensitive non-whitespace character sequences. For consistency across models, we also delete all control, format, private-use, and surrogate Unicode codepoints since they are internally removed by BERT's tokenizer. We extract contextualized word representations from BERT (cased version), GPT-2, and RoBERTa using pretrained models available through the huggingface transformers library [Wolf et al., 2019]. All methods break low-frequency words into multiple subword tokens: BERT uses WordPiece [Wu et al., 2016], while GPT-2 and RoBERTa use a byte-level variant of byte pair encoding (BPE) [Sennrich et al., 2016]. For example, the word *disillusioned* is represented by four subtokens "di -si -llus -ioned" in BERT and by two subtokens "disillusion -ed" in GPT-2 and RoBERTa. One key difference between these tokenizers is that byte-level BPE can encode all inputs, while WordPiece replaces all Unicode codepoints it has not seen in pretraining with the special token *UNK*. For simplicity, rather than using a sentence splitter we divide documents into the maximum length subtoken blocks. To make vocabularies comparable across models with different subword tokenization schemes, we reconstitute the original word tokens by averaging the vectors for subword units [Bommasani et al., 2020].

CLUSTERING. We cluster tokens using spherical k-means [Dhillon and Modha, 2001] with spkm++ initialization [Endo and Miyamoto, 2015] because of its simplicity and high-performance, and cosine similarities are commonly used in other embedding contexts. Although we extract contextualized features for all tokens, prior to clustering we remove frequent words occurring in more than 25% of documents and rare words occurring in fewer than five documents. Each clustering is run for 1000 iterations or until convergence, whichever comes first. For LDA, we

[2] https://spacy.io/

train models using Mallet [McCallum, 2002] with hyperparameter optimization occurring every 20 intervals after the first 50. For each embedding model, we cluster the token vectors extracted from the final layer $L[-1]$, the penultimate layer $L[-2]$, and the antepenultimate layer $L[-3]$. Vulić et al. [2020] suggest combining multiple layers, but no combination we tried provided additional benefit for this specific task. We consider more than the final hidden layer of each model because of the variability in anisotropy across layers [Ethayarajh, 2019]. In a space where any two words have near perfect cosine similarity, clustering will only capture the general word distribution of the corpus. Since Ethayarajh [2019] has shown GPT-2's final layer to be extremely anisotropic, we do not expect to produce viable topics in this case. For each test case, we build ten models each of size $K \in \{50, 100, 500\}$.

3.4 EVALUATION METRICS

We evaluate the quality of "topics" produced by clustering contextualized word representations with several quantitative measures. For all models we use hard topic assignments, so each word token has a word type w_i and topic assignment z. Note that we use "topic" and "cluster" interchangeably.

WORD ENTROPY. As a proxy for topic specificity, we measure a topic's word diversity using the conditional entropy of word types given a topic.

$$H(w \mid k) = -\sum_i \Pr(w_i \mid z) \log \Pr(w_i \mid z)$$

Topics composed of tokens from a small set of types will have low entropy (minimum 0), while topics more evenly spread out across the whole vocabulary will have high entropy (maximum log of vocabulary size; approx. 10 for WIKIPEDIA).

There is no best fit between quality and specificity, but extreme entropy scores indicate bad topics. Topics with extremely low entropy are overly specialized, while those with extremely high entropy are overly general.

COHERENCE. We measure the semantic quality of a topic using two word-cooccurrence-based coherence metrics. These coherence metrics measure whether a topic's words actually occur together. Internal coherence uses word cooccurrences from the working collection, while external coherence relies on word cooccurrences from a held-out external collection. The former measures fit to a dataset, while the latter measures generalization. For internal coherence we use Mimno et al. [2011]'s topic coherence metric,

$$\sum_i \sum_{j<1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_i)},$$

where D refers to the number of documents that contain a word or word-pair. For external coherence we use Newman et al. [2010]'s topic coherence metric,

$$\sum_i \sum_{j<1} \log \frac{\Pr(w_i, w_j) + \epsilon}{\Pr(w_i) \Pr(w_j)},$$

where probabilities are estimated from the number of 25-word sliding windows that contain a word or word-pair in an external corpus. We use the New York Times Annotated Corpus [Sandhaus, 2008] as our external collection with documents corresponding to articles (headline, article text, and corrected text) tokenized with spaCy. For both metrics, we use the top 20 words of each topic and set the smoothing factor ϵ to 10^{-12} to reduce penalties for non-cooccurring words [Stevens et al., 2012]. We ignore words that do not appear in the external corpus and do not consider topics that have fewer than 10 attested words. These "skipped" topics are often an indicator of model failure. Higher scores are better.

EXCLUSIVITY. A topic model can attain high coherence by repeating a single high-quality topic multiple times. To balance this effect, we measure topic diversity using Bischof and Airoldi [2012]'s word-level exclusivity metric to quantify how exclusive a word w is to a specific topic z,

$$\frac{\Pr(w_i \mid z)}{\sum_{z'} \Pr(w_i \mid z')}$$

A word prevalent in many topics will have a low exclusivity score near 0, while a word occurring in few topics will have a score near 1. We lift this measure to topics by computing the average exclusivity of each topic's top 20 words. While higher scores are not inherently better, low scores are indicative of topics with high levels of overlap.

3.5 RESULTS

We evaluate whether contextualized word representation clusters can group together related words, distinguishing distinct uses of the same word based on local context. Compared to bag-of-words LDA, we expect contextualized embedding clusters to encode more syntactic information. As we are not doing any kind of fine-tuning, we expect performance to be best on text similar to the pretraining data. We also expect contextualized embedding clusters to be useful in describing differences between partitions of a working collection.

BERT PRODUCES MEANINGFUL TOPIC MODELS. BERT cluster models consistently form semantically meaningful topics, with the final layer performing marginally better for larger K. Figure 3.1 shows that BERT clusterings have the highest external coherence, matching LDA for $K \in \{50, 100\}$ and beating LDA for

$K = 500$. For internal coherence, the opposite is true, with BERT on par with LDA for smaller K, while LDA "fits" better for $K = 500$. This distinction suggests that at very fine-grained topics, LDA may be overfitting to noise. BERT has relatively low word entropy, indicating more focused topics on average. Figure 3.2 shows the number of word types per cluster. BERT clusters are on average smaller than LDA topics (counted from an unsmoothed sampling state), but very few BERT clusters fall below our 10-valid-words threshold for coherence scoring. BERT clusters are not only semantically meaningful, but also unique. Figure 3.1 shows that BERT clusters have exclusivity scores as high if not higher than LDA topics on average. Since there is little difference between layers, we will only consider BERT L[-1] for the remainder of this work.

GPT-2 *can* PRODUCE MEANINGFUL TOPIC MODELS. As expected, the final layer clusterings of GPT-2 form bad topics. These clusters tend to be homogeneous (low word entropy) and similar to each other (low exclusivity). They also highlight the differences between our two coherence scores. Since these clusters tend to repeatedly echo the background distribution of WIKIPEDIA, they perform relatively well for internal coherence, but poorly for external coherence. Since the final layer of GPT-2 has such high anisotropy, we cannot expect vector directionalities to encode semantically meaningful information. In contrast, the penultimate and antepenultimate layer clusterings perform much better. We see a large improvement in external coherence surpassing LDA for $K = 500$. Topic word entropy and exclusivity are also improved.

For $K = 500$, GPT-2 L[-3] has surprisingly low mean internal coherence—the worst scores in Figure 3.1 by a significant margin. The number of topics below the 10-valid-words threshold is similar to BERT, so this result is comparable. We posit that this layer is relying more on transferred knowledge from the pretrained GPT-2

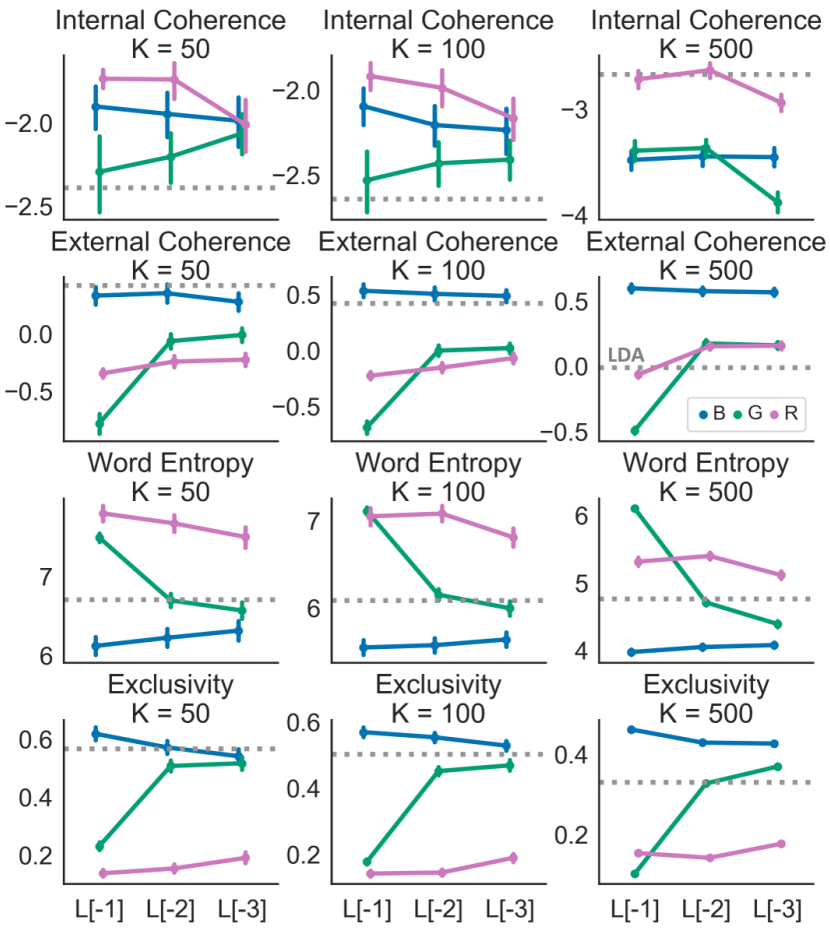

Figure 3.1: Contextual embedding clusters produce mean internal and external coherence scores comparable to LDA (dashed line). BERT clusters (blue) have high mean external coherence, better than LDA for large numbers of topics. BERT clusters contain more unique words, while RoBERTa (red) and GPT-2 (green) $L[-1]$ clusters tend to repeat similar clusters. BERT clusters have the highest word concentrations.

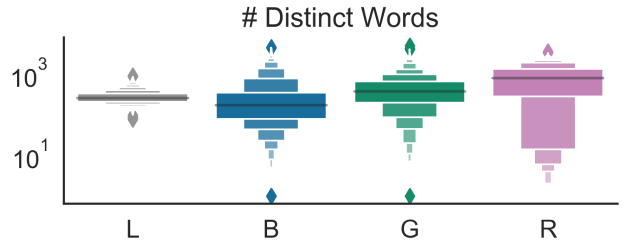

Figure 3.2: Distinct words per cluster for LDA, BERT $L[-1]$, GPT-2 $L[-2]$, and RoBERTa $L[-1]$ for $K = 500$. Although the average BERT cluster covers fewer word types, RoBERTa has more clusters with very few (< 20) word types.

model than the working collection. Because of this less explained behavior, we will only consider GPT-2 L[-2] going forward.

ROBERTA CLUSTERS ARE NOTICEABLY WORSE. Given BERT's success and GPT-2's partial success, we were surprised to find that RoBERTa cluster models were consistently of poor quality, with very low exclusivity scores and high word entropies. Although RoBERTa scores fairly well in coherence, this is not indicative of collectively high quality topics because of the correspondingly low exclusivity scores. As shown in Figure 3.1, RoBERTa has the highest average number of distinct words per cluster, but also many of clusters that contain very few distinct words. For $K = 500$, 25–50 clusters are skipped on average for different layer choices. For example, one topic consists entirely of the words *game, games, Game*, another just *ago*, and one simply the £ symbol. The remaining tokens are thus limited to a smaller number of more general topics that are closer to the corpus distribution.

While it is commonly accepted that RoBERTa outperforms BERT for a variety of natural language understanding tasks [Wang et al., 2019], we find the opposite to be true for semantic clustering. There are a number of differences between BERT and RoBERTa, but our experimental results do not mark a clear cause. The tokenization method is a very unlikely source since GPT-2 uses the same scheme.

CONTEXTUALIZED EMBEDDING CLUSTERS CAPTURE POLYSEMY. A limitation of many methods that rely on vocabulary-level embeddings is that they cannot explicitly account for polysemy and contextual differences in meaning. In contrast, token-based topic models are able to distinguish differences in meaning between contexts. There has already been evidence that token-level contextualized embeddings are able to distinguish contextual meanings for specific examples

Model	Perc.	Entr.	Top Words (noun verb adj other)
LDA	5%	0.69	Valley Death valley Creek California mining °
	25%	0.97	army forces soldiers campaign troops captured
	50%	1.11	society News Week Good Spirit Fruit says Doug
	75%	1.28	Washington Delaware ceremony Grand Capitol
	95%	1.53	critics reviews review positive mixed list
BERT	5%	0.00	1997 1996 1995 1937 1895 1935 96 1896 1795 97 09
	25%	0.61	Jewish Israel Jews Ottoman Arab Muslim Israeli
	50%	0.86	captured defeated attacked capture attack siege
	75%	1.09	hop dance hip B R Dance Hip Z Hop rapper Jay
	95%	1.48	separate combined co joint shared divided common
GPT-2	5%	0.00	2004 2003 2015 2000 2014 1998 2001 2013 2002 1997
	25%	0.42	Atlantic Pacific Gulf Mediterranean Caribbean
	50%	0.73	knew finds discovers learned reveals discovered
	75%	1.02	Olympic League FA Summer Premier Division
	95%	1.42	positive mixed critical negative garnered favorable

Table 3.3: Contextualized embedding clusters are more syntactically aware than LDA. Topics ranked by the entropy of POS distribution of the top 20 words $K = 500$.

[Wiedemann et al., 2019; Reif et al., 2019], but can they also do this for entire collections?

Instead of manually selecting terms we expect to be polysemous, we choose terms that occur as top words for clusters with dissimilar word distributions (high Jensen-Shannon divergence). While dissimilarity is not indicative of polysemy—different topics can use a term in the same way—it narrows our focus to words that are more likely to be polysemous. Table 3.1 show topics for two such terms "land" and "metal". All models distinguish *metal* the material from *metal* the genre, but BERT and GPT-2 are also distinguish *land* the noun from *land* the verb.

CONTEXTUALIZED EMBEDDING CLUSTERS ARE MORE SYNTACTICALLY CONSISTENT THAN LDA TOPICS. Contextualized word representations are known to represent a large amount of syntactic information that is not available to traditional

bag-of-words topic models [Goldberg, 2019]. We therefore expect that token-level clusterings of contextualized word representations will have more homogeneity of syntactic form, and indeed we find that they do.

As a simple proxy for syntactic similarity, we find the most likely part of speech (POS) for the top words of each cluster. We use this method because it is easy to implement; inaccuracies should be consistent across models. To measure the homogeneity of POS within each topic, we count the distribution of POS tags for the top 20 words of a cluster and calculate the entropy of that distribution. If all 20 words are the same POS, this value will be 0, while if POS tags are more diffuse it will be larger. We find that BERT and GPT-2 clusters have consistently lower entropy. In Table 3.3, we see that the 25th percentile for LDA topics has entropy 0.97, higher than the median entropy for both BERT and GPT-2. We find that these results are consistent across model sizes. Although contextualized embedding clusters are more homogeneous in POS, LDA may appear more so because it is dominated by nouns. For LDA, nouns and proper nouns account for 43.7% and 33.4% respectively of all the words in the top 20 for all topics, while verbs make up 8.5% and adjectives 6.7%. These proportions are 39.0%, 25.3%, 14.9%, and 8.1% for BERT and 37.0%, 23.4%, 16.9%, and 9.5% for GPT-2.

COMPRESSION IMPROVES EFFICIENCY WHILE MAINTAINING QUALITY. We have established that we can effectively learn topic models by clustering token-level contextualized embeddings, and we have shown that there are advantages to clustering at the token rather than vocabulary level. But for token-level clustering to be more than a curiosity we need to address computational complexity. Vocabularies are typically on the order of tens of thousands of words, while collections often contain millions of words. Even storing full 768-dimensional vectors for millions of tokens, much less clustering them, can be beyond the capability of many po-

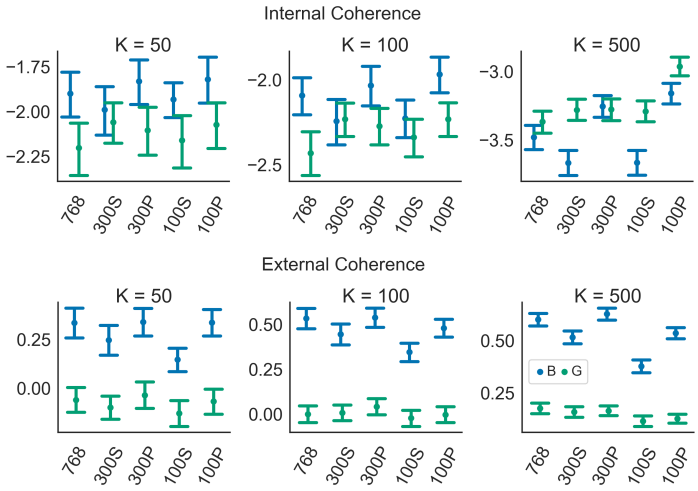

Figure 3.3: Mean internal and external coherence for reduced features for BERT and GPT-2. Features reduced with PCA tend to have higher coherence than SRP.

tential users' computing resources. Therefore, we investigate the effects of feature dimensionality reduction to reduce the memory footprint of our method.

The hidden layers of deep learning models are known to learn representations with dimensionalities much lower than the number of neurons [Raghu et al., 2017]. We apply two methods for dimensionality reduction: principal component analysis (PCA) and sparse random projection (SRP) [Li et al., 2006; Achlioptas, 2001].[3]

We find that reducing our token vectors to as few as 100 dimensions can have little negative effect. Figure 3.3 shows that reduced PCA features produce improved internal coherence and little significant change in external coherence, but reduced SRP features are worse in both metrics, especially for BERT. We note that more clusters pass the 10-valid-words threshold for reduced PCA and SRP features. Instead of skipping 10 BERT clusters and 7 GPT-2 clusters on average, no clusters are skipped for PCA reduced features and only 1 for 100-dimensional SRP features. For 300-dimensional SRP features there is only a significant drop for GPT-2 with 3 skipped on average. This decrease in skipped topics indicates that overly specific

[3] All implementations from [Pedregosa et al., 2011].

topics are being replaced with less specific ones. We hypothesize that dimensionality reduction is smoothing away "spikes" in the embeddings space that cause the algorithm to identify small clusters. Finally, larger dimensionality reductions decrease concentration and exclusivity, making clusters more general.

PCA is significantly better than SRP, especially for more aggressive dimensionality reductions. We find that mean-centering SRP features does not significantly improve results. An advantage of SRP, however, is that the projection matrix can be generated offline and immediately applied to embedding vectors as soon as they are generated. To overcome the memory limitations of PCA, we use a batch approximation, incremental PCA [Ross et al., 2008]. Using 100 dimensions and scikit-learn's default batch size of five times the number of features (3840), we find no significant difference in results between PCA and incremental PCA. For the remainder of experiments we use 100-dimensional vectors which correspond to the top 100 components produced by incremental PCA.

PRETRAINED EMBEDDINGS ARE EFFECTIVE FOR DIFFERENT COLLECTIONS. While BERT and GPT-2 cluster models produce useful topics for WIKIPEDIA, will this hold for collections less similar to the training data of these pretrained models? Does it work well for collections of much shorter and much longer texts than Wikipedia articles? We find that both BERT and GPT-2 produce semantically meaningful topics for SCOTUS and REVIEWS, but BERT continues to outperform GPT-2. As with WIKIPEDIA, we find that contextualized embedding clusters have the largest advantage over LDA for large K. Figure 3.4 shows that for $K = 500$ BERT and GPT-2 clusters have significantly higher external coherence scores on average than LDA topics for SCOTUS and very similar scores for REVIEWS. For smaller K, LDA has the highest external coherence scores followed by BERT. Internal coherence is more difficult to interpret because of the variability in exclusivity.

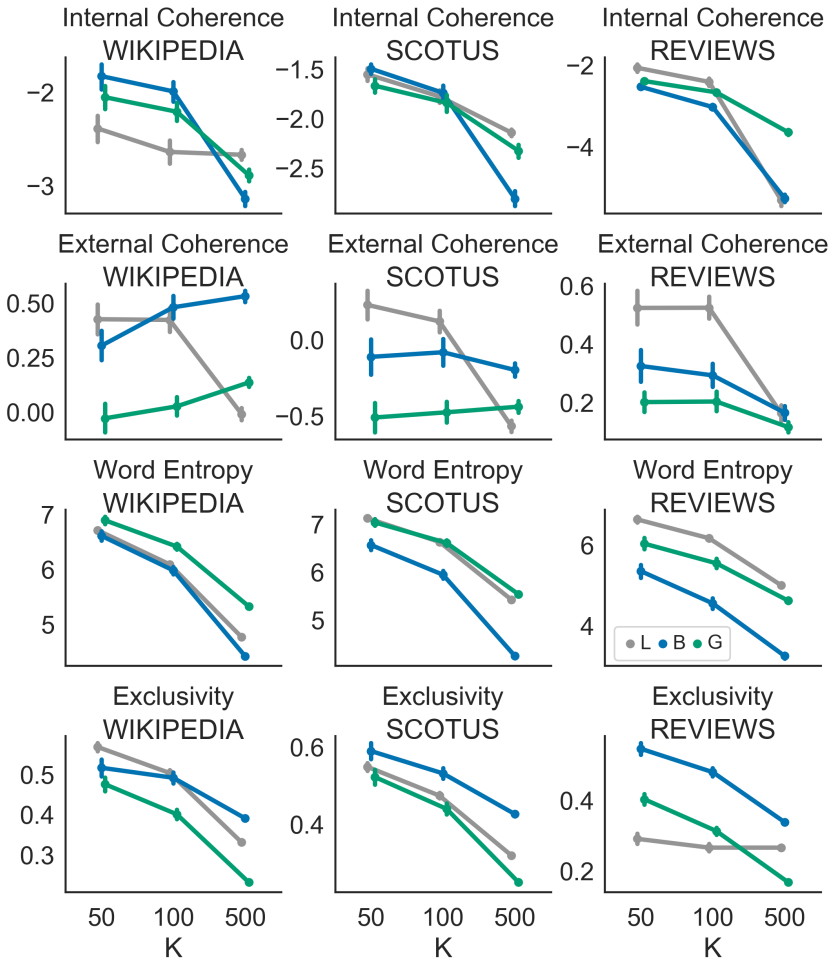

Figure 3.4: BERT and GPT-2 produce coherent topics for less familiar (w.r.t. pretraining) collections. BERT consistently produces more unique clusters. LDA external coherence drops for $K = 500$.

With $K = 500$, BERT clusters have substantially worse internal coherence scores. In contrast, GPT-2 clusters tend to experience a smaller drop in scores, but this can be partially explained by their much lower average exclusivity. We find that BERT consistently produces the most unique topics for Scotus and Reviews. BERT consistently has significantly higher mean exclusivity scores for both Scotus and Reviews, while GPT-2 tends to have scores as good as LDA for $K \in \{50, 100\}$, but significantly lower for $K = 500$.

CONTEXTUALIZED EMBEDDING CLUSTERS SUPPORT COLLECTION ANALYSIS. Token-level clusterings of contextualized word representations support sophisticated corpus analysis with little additional complexity. In practice, topic models are often used as a way to "map" a specific collection, for example by identifying key documents or measuring the prevalence of topics over time [Boyd-Graber et al., 2017]. A key disadvantage therefore of vocabulary-level semantic clustering is that it is not *grounded* in specific documents in a collection. Token-level clustering, in contrast, supports a wide range of analysis techniques for researchers interested in using a topic model as a means for studying a specific collection. We provide two case studies that both use additional metadata to organize token-level clusterings *post hoc*.

Given a partition of the working collection, we can count tokens assigned to each topic within a given partition to estimate locally prominent topics. Table 3.4 shows the three most prominent topics for the four product categories in REVIEWS from a BERT cluster model with $K = 500$. Many of these topics are clearly interpretable as aspects of a particular product (e.g. *full albums*, *individual songs*, *descriptions of songs*). Two topics contain mostly prepositions. While we could have added these words to a stoplist before clustering, these less obviously interpretable clusters can nevertheless represent distinct discourses, such as descriptions of action in Movies (*into, up, over, through, ...*) or descriptions of physical objects in Electronics (*use, up, than, off, ...*). We can also find the topics that are *least* associated with any one product category by calculating the entropy of their distribution of tokens across categories. These are shown in Table 3.5, and appear to represent subtle variations of subjective experiences: *overkill*, *possibilities*, *reasons*, and *time periods*. We emphasize that this analysis requires no additional modeling, simply counting.

For partitions that have a natural order, such as years, we can create time series in the same *post hoc* manner. Thus, we can use a BERT clustering of SCOTUS

books	book books author novel novels work Book fiction by authors
	read reading copy Read reads Reading readable reader reread
	problem children problems course power lives mystery
electronics	use up than off used back over using there about work need
	screen quality sound device power battery unit system software
	setup remote battery mode card set range input signal support
movies	movie movies films flick theater Movie flicks game cinema film
	into up over through between off down than about around
	film movie picture screen documentary films Film cinema
cds	album albums record release Album LP releases records effort
	songs tracks hits tunes singles material stuff Songs ballads cuts
	lyrics guitar vocals voice bass singing solo vocal sound work

Table 3.4: Unlike vocabulary-level clusters, token-level clusters are grounded in specific documents and can be used to analyze collections. Here we show the most prominent BERT topics ($K = 500$) for the four product categories in REVIEWS. This analysis is purely *post hoc*, neither BERT nor its clustering have access to product labels.

than beyond Than twice upon much except besides half times less unlike nor
day days Day morning today date daily Days month 19 basis night period
can Can able possible manage could knows lets capable allows can't easily s
when When once time whenever Once soon everytime upon during Whenever
too enough Too overly taste beyond tired sufficiently plenty somehow Enough
because since due Because Since considering cause meaning given thanks
went took got happened came started did turned ended fell used kept left
would 'd Would might d normally I'd imagine otherwise happily woulda
up off Up ready upload ups along end forth uploading used down away
instead rather either matter Instead opposed other depending based choice

Table 3.5: The ten BERT topics ($K = 500$) from REVIEWS with the most *uniform* distribution over product categories.

1980–2019	Top Words
	union employment labor bargaining Labor workers job strike
	gas coal oil natural mining mineral fuel mine fishing hunting
	compensation wages pension wage salary welfare compensate
	discrimination prejudice unfair bias harassment segregation
	medical health care hospital physician patient Medical
	market competition competitive markets compete demand
	election vote voting electoral ballot voter elected votes elect
	violence firearm gun violent weapon firearms armed weapons
	patent copyright Copyright Patent patents trademark invention

Table 3.6: Grounded topics allow us to analyze trends in the organization of a corpus using a BERT topic model with $K = 500$. Here we measure the prevalence of topics from 1980 to 2019 in US Supreme Court opinions by counting token assignments. Topics related to *unions* and *natural resources* are more prevalent earlier in the collection, while topics related to *firearms* and *intellectual property* have become more prominent.

to examine the changes in subject of the cases brought before the US Supreme Court. Table 3.6 shows time series for nine manually selected topics from a BERT clustering of SCOTUS with $K = 500$, ordered by the means of their distributions over years. We find that topics related to *labor and collective bargaining*, *oil and gas exploration*, and *compensation* have decreased in intensity since the 1980s, while those related to *medical care* and *elections* have remained relatively stable. It appears that *competitive markets* was a less common subject in the middle years, but has returned to prominence. Meanwhile, *discrimination* has remained a prominent topic throughout the period, but with higher intensities in the 1980s. Additionally, topics related to *gun violence* and *patents and copyright* appear to be increasing in intensity.

3.6 CONCLUSION

We have presented a simple, reliable method for extracting mixed-membership models from pretrained contextualized word representations. This result is of inter-

est in several ways. First, it provides insight into the affordances of contextualized representations. For example, our result suggests a way to rationalize seemingly *ad hoc* methods such as averaging token vectors to build a representation of a sentence. Second, it suggests directions for further analysis and development of contextualized representation models and algorithms. The significant differences we observe in superficially similar systems such as BERT and RoBERTa require explanations that could expand our theoretical understanding of what these models are doing. Why, for example, is RoBERTa more prone to very small, specific clusters, while BERT is not? Furthermore, if models like BERT are producing output similar to topic model algorithms, this connection may suggest new directions for simpler and more efficient language model algorithms, as well as more representationally powerful topic model algorithms. Third, there may be substantial practical benefits for researchers analyzing collections. Although running BERT on a large-scale corpus may for now be substantially more computationally inefficient than running highly-tuned LDA algorithms, passing a collection through such a system is likely to become an increasingly common analysis step. If such practices could be combined with online clustering algorithms that would not require storing large numbers of dense token-level vectors, data analysts who are already using BERT-based workflows could easily extract high-quality topic model output with essentially no additional work.

4

CONTINENTS OR ARCHIPELAGOS? MEASURING THE LINGUISTIC GEOMETRIES OF WORD EMBEDDINGS

This chapter is based on joint work with Maria Antoniak and David Mimno.

4.1 INTRODUCTION

One of the most important recent developments in natural language processing is the surprising power of continuous representations. Low-dimensional vector embeddings trained with simple algorithms can predict linguistically meaningful properties of language [Turian et al., 2010; Chen and Manning, 2014]. In this work we consider *how* such restricted models represent these complex properties. Are embeddings able to support complex down-stream tasks because they provide suitably rich inputs that later models can combine in complex ways, or because their geometry is primarily organized around these properties?

As a test case we focus on a specific linguistic phenomenon, part of speech (POS). We select POS because it is well-studied phenomena and computational tools exist for identifying POS with relatively high accuracy in many languages. We also focus on a specific family of algorithms, lexical word embedding models such as `word2vec` [Mikolov et al., 2013a; Mikolov et al., 2013b] and `GloVe` [Pennington et al., 2014]. We select these algorithms not because they are state-of-the-art, but because they are powerful enough to represent linguistic features [Cotterell and Schütze,

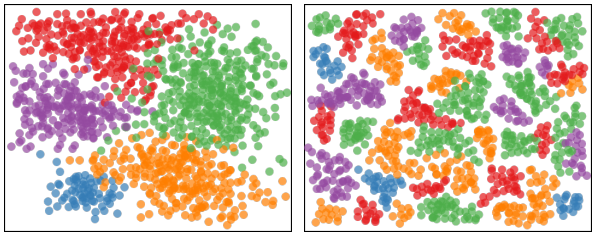

Figure 4.1: Hypothetical visualizations of continental (left) and archipelagic (right) representations.

2015] yet simple enough that we can efficiently train tens of thousands of models on specific collections in multiple languages.

POS can be defined by syntactic and morphological distributions [Schachter and Shopen, 2007] and good embedding spaces place words occurring in similar contexts close together, so we expect these models encode POS information. But in vector spaces with tens or hundreds of dimensions there may be many ways to encode such patterns. Are parts of speech nicely captured in *single* low-dimensional subspaces ("continents") or are they encoded more locally in small, dispersed clusters ("archipelagos")? Figure 4.1 shows hypothetical visualizations of these two possibilities.

In this work we use a variety of techniques to probe the geometry of parts of speech within embedding spaces. We measure how these geometric properties vary by algorithm and language. Finally, we address how parameter settings and the calibrated removal of specific word classes from the training text affect these geometric representations.

This work is distinct from popular probing tasks that determine, for example, whether a specific neural network is capable of representing a linguistic pattern; the fact that word embeddings represent POS is not controversial. Rather, we seek to determine at a detailed, numerical level *how* embeddings represent POS infor-

mation, and whether we can modify algorithmic parameters or input collections in a way that changes these POS representations.

4.2 DATA: PART OF SPEECH FROM MULTI-LINGUAL PARALLEL TEXTS

One attractive feature of POS as a linguistic phenomenon is that there exists a small number of common tags that, while not universal, have been argued to exist across large numbers of languages [Petrov et al., 2012]. These tags can be split into two groups. *Open* classes represent "content" words such as nouns and verbs and frequently incorporate new terms. *Closed* classes represent "function" words such as determiners and conjunctions and rarely if ever add new terms [Schachter and Shopen, 2007]. We focus on open classes as they dominate the vocabulary of the languages under inspection. We use closed-class words as a treatment variable and measure the effect of their presence or absence on the geometry of open-class words.

Prior work uses embeddings to predict word-level POS, providing extrinsic evidence that word embeddings contain syntactic and semantic information [Al-Rfou' et al., 2013; Owoputi et al., 2013; Schnabel and Schütze, 2014; Lin et al., 2015]. However, our goal is not to build a superior POS tagger but to measure the spatial encoding of POS.

Open Classes	Closed Classes
Adjective (ADJ)	Adposition (ADP)
Adverb (ADV)	Conjunction (CONJ)
Noun (NOUN)	Determiner (DET)
Proper Noun (PROPN)	Numeral (NUM)
Verb (VERB)	Pronoun (PRON)

Table 4.1: Key parts of speech considered in this work.

We want to compare the geometry of POS over multiple languages. In order to make meaningful comparisons we must factor out semantic differences in the training texts and use the same POS tagging method for all texts. So, we use translated parallel text from European Parliament proceedings from April 1996 through November 2011 taken from the Europarl dataset [Koehn, 2005] tagged using the spaCy toolkit.[1] The tagger uses averaged perceptron [Collins, 2002] with Brown cluster features [Koo et al., 2008]. This tagging method does *not* rely on continuous embeddings.

The combination of Europarl corpora and spaCy pre-trained POS taggers provide us with tagged parallel text over seven languages: German (de), Dutch (nl), English (en), French (fr), Italian (it), Portuguese (pt), and Spanish (es). We use this specific order as a rough proxy for language similarity. The Dutch POS tagger in spaCy deals poorly with proper nouns, but we include it as an additional comparison with that warning. Although we are not able to make conclusions about non-Western-European languages, we can still demonstrate that our findings are not restricted to one language. For each language, we segment sentences using spaCy, remove sentences with fewer than five tokens, and perform word-level tokenization and POS tagging. Finally, we reduce our working vocabulary to words occurring at least 20 times; the resulting POS distributions are shown in Figure 4.2.

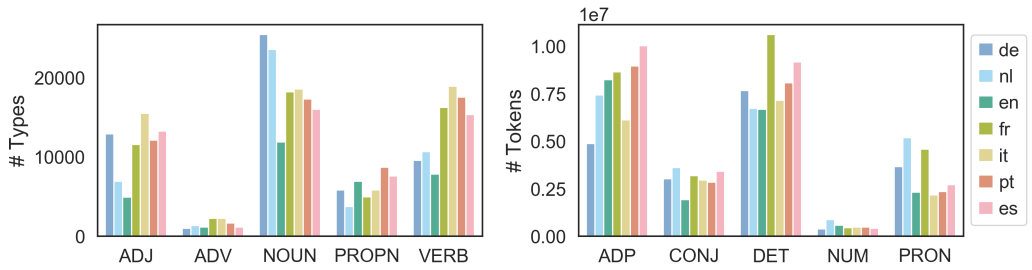

Figure 4.2: Distributions of open- and closed-class POS tags after rare words have been removed but before any other treatment.

[1] https://spacy.io/

POS ambiguity is a serious challenge for non-contextual embeddings ("I lead the charge" vs. "I charge the lead-acid battery"). We record ambiguity for word w over POS tag p with a probability distribution $\Pr(p \mid w)$ from the POS tag frequencies in each language corpus. We consider both open- and closed-class POS types in the Universal Dependencies universal POS tag set as well as an "X" tag which can indicate code-switching and foreign words. We collapse coordinating and subordinating conjunctions into a single tag. From these, we filter to those POS types that occur across all of our languages; these tags are listed in Table 4.1. For each word type in a corpus, we de-noise its observed POS distribution by removing POS that make up less than 2% of its observed distribution.

Lang.	# Sents.	Avg. Sent. Leng.	# Types	POS Entropy
de	2,210,003	21.1	45,551	$.09 \pm .20$
nl	2,474,902	21.7	32,602	$.19 \pm .28$
en	2,234,816	24.6	23,620	$.16 \pm .24$
fr	2,296,291	25.8	31,976	$.29 \pm .33$
it	2,211,793	23.1	38,610	$.25 \pm .34$
pt	2,272,752	23.7	34,969	$.26 \pm .32$
es	2,107,208	26.2	36,129	$.21 \pm .30$

Table 4.2: Summary statistics for the seven languages under consideration.

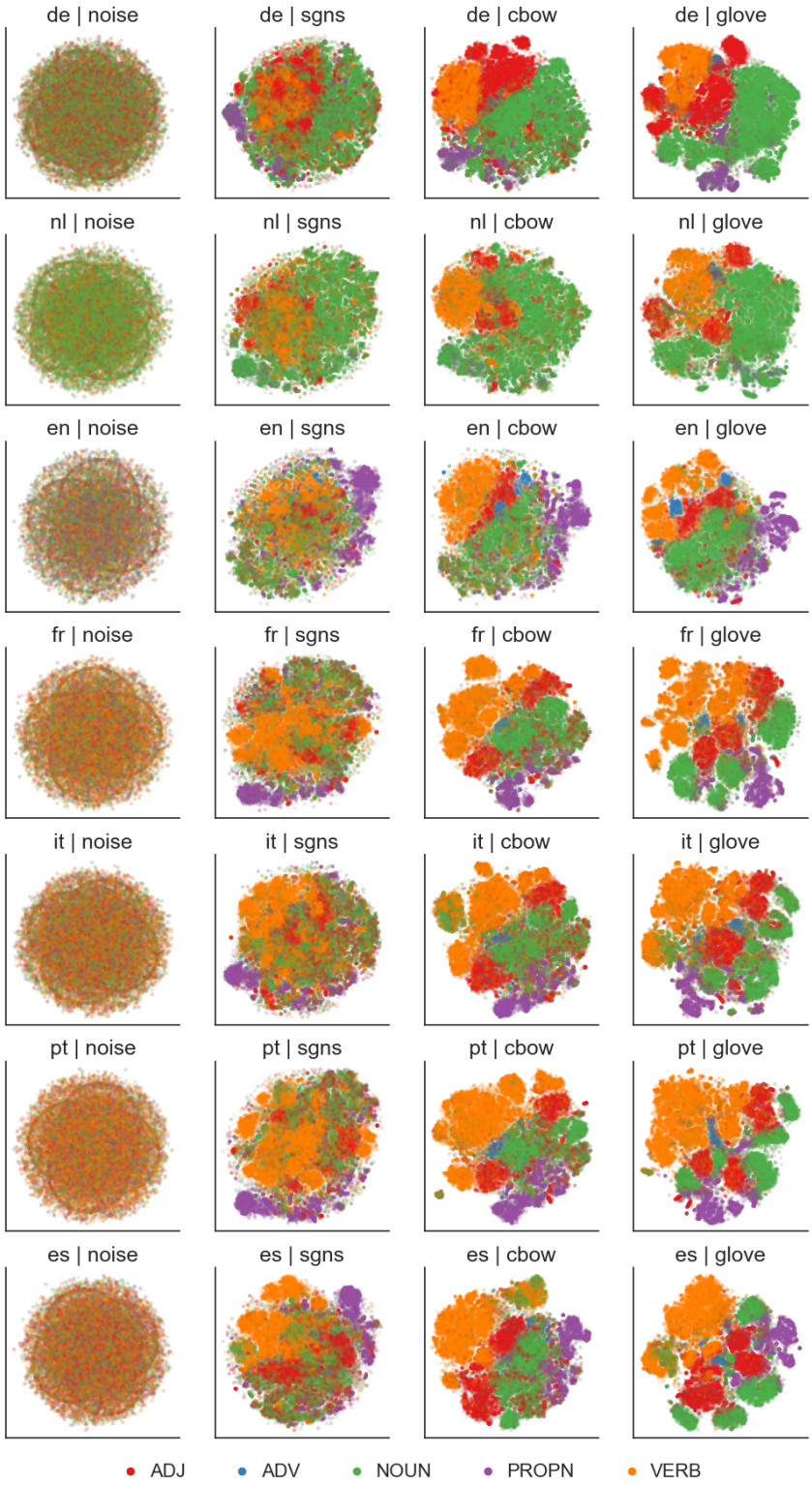

Figure 4.3: t-SNE projections of open-class words for each algorithm and language. POS clusters differently across language and embedding algorithm. For all languages, `GloVe` forms the most distinct POS formations and `CBOW` forms the most concentrated ones.

4.3 PROOF OF CONCEPT: VISUALIZING POS

For an initial intuition, we examine two-dimensional t-SNE projections [Maaten and Hinton, 2008] for each language and embedding algorithm. t-SNE projections are a simplification of the actual linguistic geometries and should be approached with caution [Wattenberg et al., 2016], but should be comparable relatively if not in absolute terms. We focus on three algorithms: global vectors for word representations (`GloVe`), skip-gram with negative sampling (`SGNS`), and continuous bag of words (`CBOW`). We use the `GloVe` implementation provided by Pennington et al. [2014]. For `SGNS` and `CBOW`, we use the implementations provided by the `fastText` library [Bojanowski et al., 2017] with subword learning turned off. For all algorithms we fix the number of training iterations to 15 and use a window size of 5, a vector size of 100, and 5 negative samples for `CBOW` and `SGNS`. For all other unspecified model parameters, we use the respective implementation defaults. We ℓ_2-normalize all vectors. Figure 4.3 shows projections for the three algorithms as well as a baseline of 100-dimensional Gaussian noise. Each point represents a word and its color represents that word's most probable POS. For ease of comparison, the projections have been rotated so that the mean vector of verbs (orange) points up and to the left.

POS ORGANIZATION VARIES BY ALGORITHM. Across all languages, `GloVe` appears to have the most distinct POS formations. While `CBOW`'s are slightly less distinct, they are more concentrated than `GloVe`'s. In contrast, `SGNS`'s POS clustering shows strong *local* clustering but lacks the global structure of `GloVe` and `CBOW`. So, `CBOW` has the most continent-like formations, but there is not a one-to-one mapping between clusters and POS classes. `GloVe`'s formations look more continental

than archipelagic given the size of its clusters. Meanwhile, SGNS seems to be an archipelago of sorts but its POS "islands" lack clear separation.

POS ORGANIZATION VARIES BY LANGUAGE. Languages from the same language family tend to have similar POS formations. Germanic language (de, nl, en) projections are dominated by noun clusters, whereas Romance language (fr, it, pt, es) projections are dominated by verb clusters, perhaps due to more flexible noun compounding in Germanic languages and richer verbal morphology in Romance languages. We also see that Romance languages tend to be more archipelagic than Germanic languages. All languages have a prominent cluster of proper nouns (except Dutch, which we believe is due to errors in tagging). Given that each language-specific corpus contains the same content, this might indicate that proper nouns are content-dependent rather than language-dependent.

4.4 QUANTIFYING POS FORMATIONS

While t-SNE projections can give us an intuition into how POS is organized within an embedding, we must confirm our observations through quantitative measurements of continental and archipelagic characteristics. Quantitative measurements of the geometries of word embedding spaces have been tangentially explored in work that seeks to align separately trained embeddings, but these works are primarily concerned with translation rather than intrinsic syntactic properties of the embedded space [Lample et al., 2018; Hartmann et al., 2018; Kementchedjhieva et al., 2018]. Inspired by probing tasks used for understanding the properties of sentence embeddings [Conneau et al., 2018], we propose to measure the clustering tendencies of POS classes through classification tasks that rely on local and global

structure. The difference is that we already know that these embeddings encode POS information; we instead want to characterize *how* POS classes are represented within the space of these embeddings.

To evaluate the POS predictions of a classifier we define $\text{pred}_c(w)$ as the predicted POS for classifier c on word type w. In order to account for POS ambiguity we define a weighted prediction accuracy for classifier c and POS class p as

$$\text{Acc}_c(p) = \frac{\sum_w \mathbf{1}(\text{pred}_c(w) = p) \cdot \Pr(p \mid w)}{\sum_w \Pr(p \mid w)},$$

where $\mathbf{1}(x)$ is the indicator function. We produce an overall classifier score by averaging the POS-level accuracy scores. In practice, the maximum value for this metric is determined by the level of ambiguity in the language. We therefore define a reference upper bound UB, which samples a POS prediction for word w by sampling from the actual smoothed POS distribution $\Pr(p \mid w)$.

GLOBAL STRUCTURE: NEAREST CENTROID. If an embedding space is primarily organized around POS (the "continents" hypothesis), then it should be sufficient to represent each POS as a single centroid, and predict POS for any point in the embedding space by finding the nearest of these centroids. We construct a centroid for each POS using the word types exclusively from that class.

LOCAL STRUCTURE: NEAREST NEIGHBORS. If an embedding space is primarily organized around some factor other than POS, and consists of large numbers of locally-consistent clumps (the "archipelagos" hypothesis), then we would expect that a word should share the POS of its near neighbors, and that this local information should be much more informative than single class-level centroids. Prior work has used the variation of nearest neighbors of open-class words to study the

similarities between spaces learned via different training algorithms [Pierrejean and Tanguy, 2018]. We use k nearest neighbors classification to measure the local POS clustering with $k = 5$. If nearest neighbor accuracy is substantially higher than centroid accuracy, we treat that as evidence for a lack of global structure. In contrast, if they are similar we treat this as further evidence for the "continent" hypothesis.

4.5 RESULTS

We use 10-fold cross validation for each trained embedding, maintaining the test splits across each trained instance. All significance tests involve a t-test paired by split/run numbers. As we are making a large number of comparisons, we describe a difference as "significant" if $p < 1 \times 10^{-7}$. We will limit many figures to German, English, and Portuguese for space reasons; results are typically comparable within language families.

4.5.1 *Effects of Algorithm and Language*

We confirm that language and embedding algorithm affect the spatial organization of POS classes. As seen in Figure 4.4, we find that all three algorithms perform significantly better than random noise for both nearest centroid and nearest neighbor classification. For nearest centroid, `CBOW` produces significantly higher accuracy scores than `GloVe` and `SGNS` across all languages. For all languages except French, `SGNS` produces significantly higher accuracy scores than `GloVe`. For nearest neighbors, `SGNS` produces significantly lower accuracy scores than `GloVe` and `CBOW` across all languages. We find that `GloVe` tends to score higher than `CBOW`, but only

significantly for Dutch, English, French, Portuguese, and Spanish. These results agree with what we observed in the t-SNE plots: all three embedding algorithms have clear archipelagic POS formations, with varying degrees of continental form. `CBOW` is the most continental and surprisingly `GloVe` is the least but contains some continent-like forms.

It is possible that centroid-based classification is influenced by the overall geometry of the embedding space, which is known to vary by algorithm. [Mimno and Thompson, 2017] find, for example, that `SGNS` vectors cluster in a cone due to negative sampling. Chandrahas et al. [2018] find that different knowledge graph (KG) embedding methods result in different geometric properties (conicity and vector spread). We therefore measure the distance from the centroid of all ℓ_2-normalized vectors in an embedding to the centroid of each POS. Adverbs and proper nouns have the largest centroid distance for all models and languages. Centroid distance is largest for `GloVe`, indicating that its POS centroids are widely separated, so any reduction in performance of centroid-based classification is due to variability around the centroids, not by the closeness of centroids.

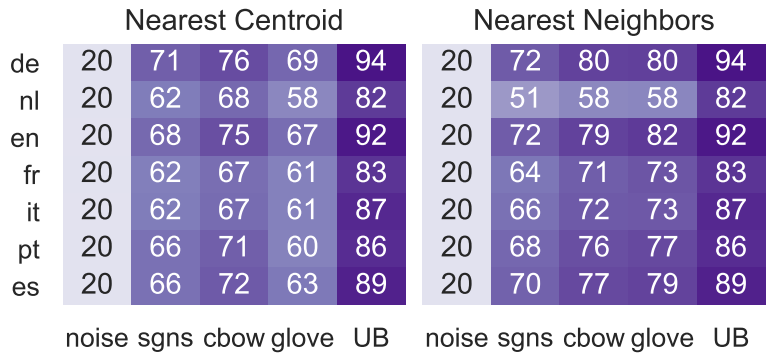

Figure 4.4: Mean scores for nearest centroid and nearest neighbor classification. Overall, `CBOW` has the highest scores for nearest centroid and `GloVe` has the highest for nearest neighbor.

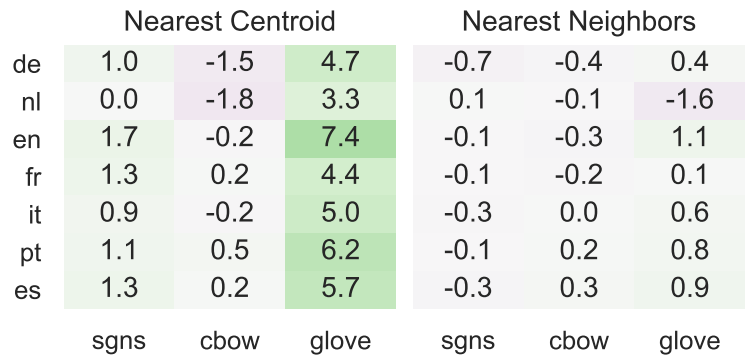

Figure 4.5: Increasing the number of training iterations from 15 to 100 has no significant effect on Nearest Neighbor accuracy scores, but SGNS and especially GloVe benefit from more training.

4.5.2 *Effects of Parameter Settings*

In this section we consider the factors that affect how each algorithm encodes POS information, holding the collection fixed.

TRAINING TIME. All of the algorithms that we consider are in some way stochastic, making small steps in the direction of a gradient. It is possible that any variation we observe is simply a factor of the convergence properties of the stochastic algorithms. To determine whether POS formations improve with additional training time we increase the number of training iterations to 100.[2]

In Figure 4.5, we observe that increasing the number of training iterations affects our classification tasks in different ways. For nearest neighbors, changes are either not significant or are inconsistent across languages. For nearest centroid, GloVe scores significantly improve for all languages, to the point where they are comparable to CBOW. SGNS scores experience a small but significant increase for all languages except Dutch. In contrast, CBOW scores see no consistent change across

[2]Iteration numbers are not necessarily comparable, but they provide a rough estimate of computation.

languages: German and Dutch scores worsen, Portuguese scores increase, and the remaining four languages experience no significant difference. These results indicate that additional training predominantly affects global organization of POS rather than local neighborhoods. In the case of GloVe (and possibly SGNS), POS "islands" are shifting within the embedding space such that they are closer to other same-class "islands" during the additional training iterations.

WINDOW SIZE. All of the embedding algorithms we consider are driven by the cooccurrence of words within a sliding window. Any information we learn about words is therefore sourced from these contexts. It is therefore natural to ask how the *size* and *weighting* of context windows affects the properties of the resulting embeddings.

Changing window size and construction is known to impact embedding quality [Levy et al., 2015; Levy and Goldberg, 2014]. In fact, different algorithms explicitly prefer different window sizes. SGNS and CBOW use a default window size of 5, while GloVe uses a default of 15. Moreover, window size affects how well embeddings encode POS information [Bansal et al., 2014; Lin et al., 2015]. But again, we want to understand how window size alters the *geometric* encodings of POS.

As expected, changes in window size significantly affect POS formations at both the local and global level. As seen in Figure 4.6, GloVe classifier scores increase with window size, while CBOW and SGNS scores decrease. For both classifiers, the highest overall accuracy scores for all languages correspond to SGNS and CBOW embeddings trained with a window size of 1. As window size increases, SGNS scores plummet and are surpassed by the rising GloVe scores. While CBOW scores also decrease as window size increases, they do so at a slower rate than SGNS. GloVe scores initially increase as window size increases, but hardly change once window size exceeds 5 or 10. For nearest centroid scores, CBOW is never surpassed by GloVe. In fact, for

a window size of 100, CBOW accuracy scores are significantly larger than GloVe for all languages except French and Italian. In contrast, for nearest neighbor, GloVe surpasses CBOW for all languages beyond window size 5 or 10.

These results support the hypothesis that syntactic information is mostly captured locally: neighboring words (at least for the languages we examine) are all you need to learn open-class POS. More importantly, we find that smaller window sizes make word2vec-style algorithms (e.g. SGNS and CBOW) more continent-like. This echoes Bansal et al. [2014] and Lin et al. [2015]'s finding that continuous skip-gram (i.e. SGNS without negative sampling) embeddings are most useful for POS prediction for small context windows. Surprisingly, we find that GloVe's POS formations become more defined and global as window size increase towards 10.

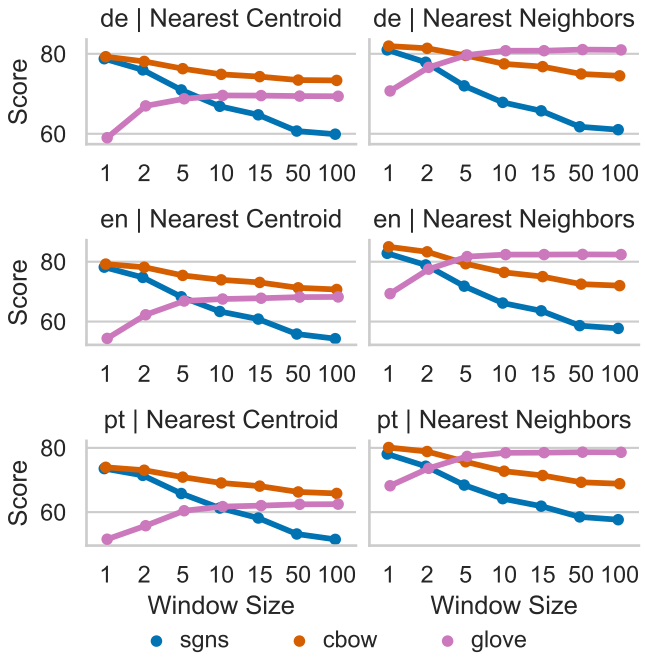

Figure 4.6: Window size versus classifier score. GloVe scores increase with window size while CBOW and SGNS scores decrease.

WINDOW WEIGHTING. The relative weighting of words within the sliding context window is subtly different across algorithms. `word2vec`-style algorithms weight each word token within the context window equally, but sample an *effective* context window size for each word token uniformly from 1 to the maximum window size. As a result, if the maximum window size is five, in expectation a word that is one word position away from the query word will be sampled five times more often than the word five positions away, corresponding to a *linear* weighting proportional to distance. In contrast, the standard `GloVe` implementation weights neighboring words using a *hyperbolic* weighting, such that a word t positions away from the query word will have weight proportional to $1/t$. This difference means that even with large window sizes `GloVe` still puts substantial weight on nearby words. Lison and Kutuzov [2017] find that sublinear weighting has no consistent effect on word similarity, but might changes in window weighting affect the learning of syntactic information?

In order to compare the effect of window weightings, we modify the reference implementations of both families of algorithms to use the weighting scheme of the other. For `GloVe` we implement a linear window weighting, while for `word2vec`-style algorithms we modify `fastText` to sample window sizes with probability proportional to the inverse of the token distance to emulate a hyperbolic weighting, in expectation.

Overall, hyperbolic weighting tends to improve accuracy scores across languages and algorithms, while linear weighting worsens them. As seen in Figure 4.7, we see that hyperbolic weighting significantly improves accuracy scores for `SGNS` for window sizes 5 and higher. But, these scores still decrease as window size increases. `CBOW` exhibits a similar but much smaller behavior: hyperbolic weighting either causes at most a slight improvement. In contrast, linear weighting significantly

decreases GloVe's scores once a window size of 5 is reached. At this point, linear weighted GloVe's scores decrease at a similar rate to standard SGNS.

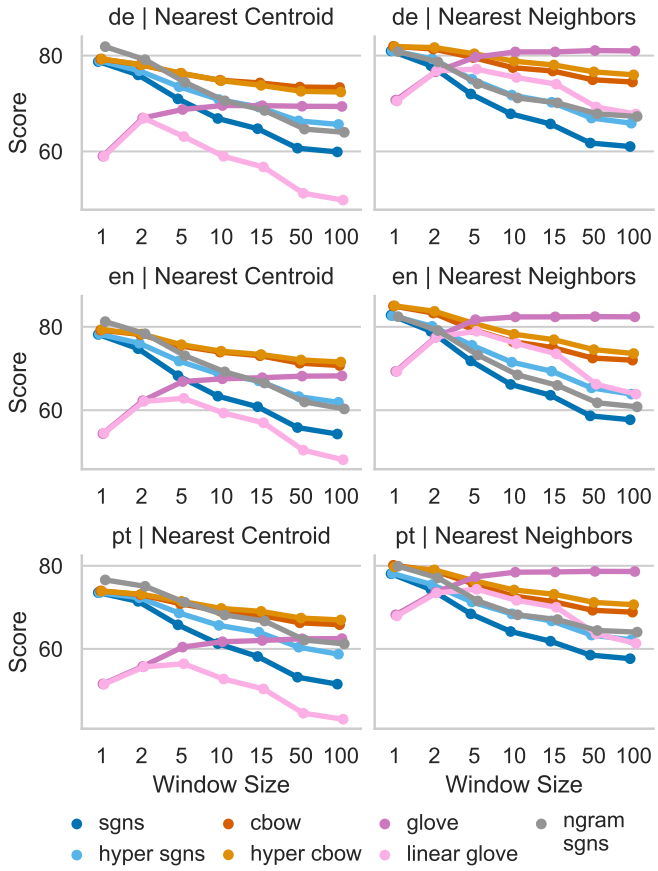

Figure 4.7: Window size versus classifier score including algorithms with altered window weighting or subword information. Hyperbolic weighting tends to improve scores, while linear weighting worsens them. For SGNS, using subword information results in similar increases to using hyperbolic window weighting.

SUBWORD INFORMATION. Morphology often indicates POS information that can be used in machine learning [Pinter et al., 2019]. Adding character n-grams as features using the fastText algorithm [Bojanowski et al., 2017] scores significantly higher than SGNS for all languages and a large majority of window sizes. As seen in Figure 4.7, fastText scores tend to decrease at a similar rate as SGNS scores and are

often quite similar to the scores of SGNS with hyperbolic weighting. Additionally, fastText produces the most continental POS formations for a window size of 1.

4.5.3 *Effects of Data Modifications*

We have shown that modifying algorithm parameters affects the encoding of POS, but what about modifications to the input texts? The difficulty of aligning the embeddings of language pairs has been shown to be dependent not only on the training algorithms but also on qualities of the training corpora Hartmann et al. [2018]. Word frequency, the presence or absence of specific documents, document length, and corpus size can result in surprising instabilities of cosine similarities [Hellrich and Hahn, 2016; Tian et al., 2016; Antoniak and Mimno, 2018; Wendlandt et al., 2018]. Prior work [Tang et al., 2016; Lison and Kutuzov, 2017] has shown that removing function words worsens syntactic learning while improving semantic learning.

In this section we selectively remove tokens from closed classes and measure the difference between the resulting embeddings of open classes. As a baseline modification we randomly shuffle the tokens in each sentence. We expect this treatment to have a more destructive effect on POS classification than any subsampling treatment that preserves word order. We examine three types of corpus-based interventions. In the first, we remove in expectation 50%, 75%, 90%, and 100% of tokens tagged as a closed-class POS. In the second, we remove all tokens from each closed class, one class at a time. In the third, we remove all closed-class tokens *except* for one class, one class at a time.

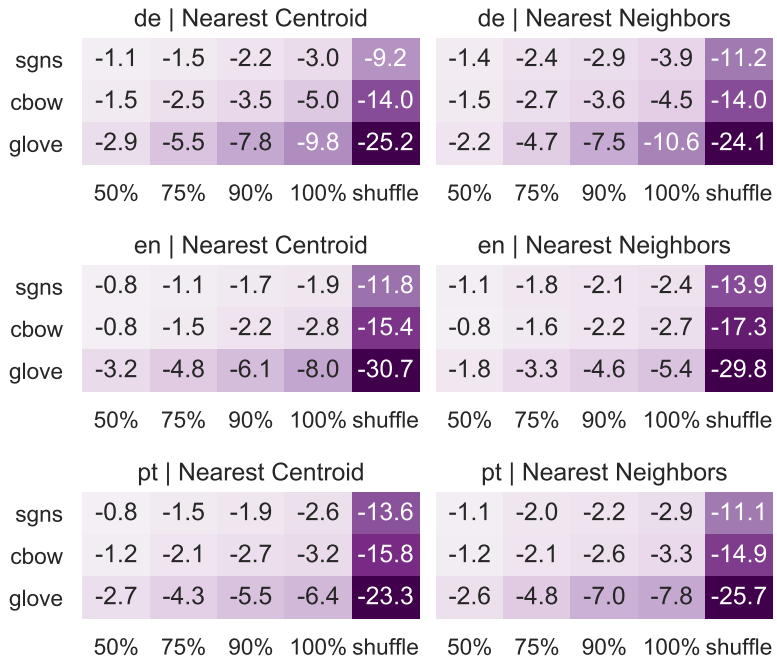

Figure 4.8: Change in classifier scores versus subsampling of closed-class tokens. Removing 50% or more closed-class tokens harms scores for all models and languages.

REMOVING CLOSED-CLASS POS HARMS OPEN-CLASS POS FORMATIONS. For all algorithms, the removal of 50% or more of closed-class tokens harms nearest centroid and nearest neighbor scores. As we would expect, more harm is caused by larger deletions. In Figure 4.8, we find that all models experience significant decreases in score, but `GloVe`'s change is significantly larger. Unlike `GloVe`, these algorithms already remove significant numbers of frequent words, which tend to be closed-class. The `fastText` implementation discards a token of word type w at the rate $1 - (\sqrt{\lambda} + \lambda)$, where λ is the ratio between a user-specified threshold (default 0.0001) and the overall frequency of w.[3] This process is vital for good performance: without it both nearest centroid and nearest neighbor scores are

[3]This formula differs from the one presented in Mikolov et al. [2013b], but matches the one used in current `word2vec`implementations.

similar to random noise. But it also removes close to 75% of tokens for all closed classes in all languages, with the exception of numbers.

Examining POS-level scores (omitted due to space limitations), we find that adjective, noun, and verb formations are most affected by our interventions. Proper nouns tend to be the least affected (except for German) with their local and global structure experiencing the least harm. Adverbs tend to maintain their global form, but their local structure deteriorates. In Romance languages, the local clustering of adverbs is especially affected.

THE INFLUENCE OF CLOSED-CLASS POS VARIES. We measure the importance of a closed-class POS type in two ways. By removing all of the class's tokens, we can measure how unique this POS is with respect to the other closed class POS types. In contrast, by removing all closed-class POS *except* this class, we can measure the individual influence of this class on open-class POS formations. We expect the first type of intervention to have less of an effect than the second; the second removes dramatically more tokens. Due to the interference between `fastText` downsampling and our modifications, we focus only on the results of `GloVe`.

In Figure 4.9, we see that numerals are redundant with the other closed-class POS, but that their presence does provide a very small improvement over the absence of all closed-class tokens. Determiners and adpositions have the largest influence on open POS formations, the removal of either significantly worsens global and local clustering. Pronouns and conjunctions have much smaller impact and tend to influence global formations over local ones.

Examining POS-level scores, we find that the effect of these interventions varies by language. As we might expect, the presence of determiners affects noun formation. For French, the removal of determiners has a uniquely large effect on the local and global structure of nouns. But for English and Portuguese, adpositions have

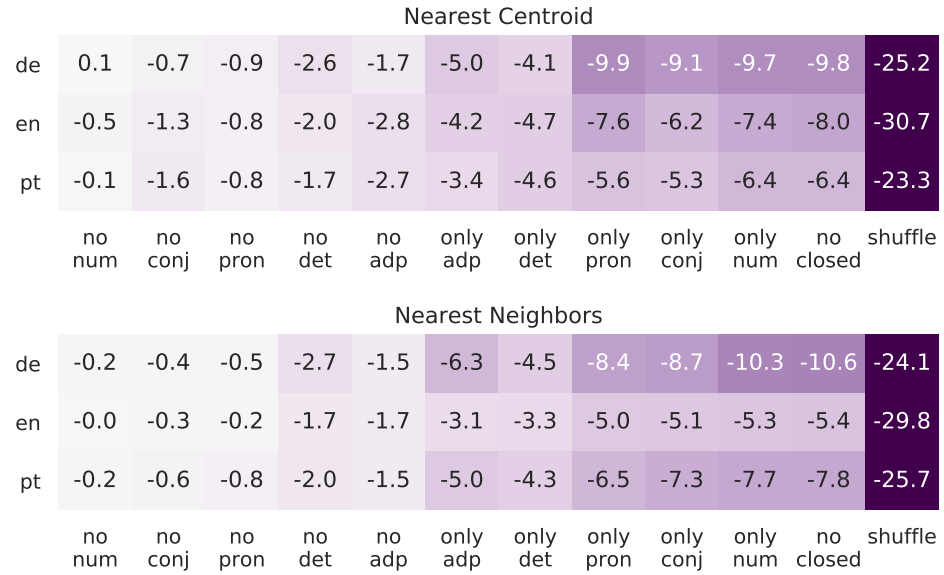

Figure 4.9: Change in GloVe classifier scores versus removing one or all-but-one closed class. Determiners and adpositions have the largest individual influence on scores, while numerals have the least.

a stronger effect on the global noun formations than determiners, but both have equally strong influence on local noun formation. A more systematic analysis of the effects of closed-class POS on embeddings is beyond the scope of this work, but could provide a fascinating tool for comparative linguistics.

4.6 CONCLUSION

In this paper we introduce general methods for measuring the geometric structure of word vector features. These methods enable us to investigate the geometries of POS within word embeddings. While we find that no algorithm is entirely a continent or an archipelago, CBOW best satisfies the "continent" hypothesis while GloVe and SGNS best satisfy the "archipelago" hypothesis.

We confirm that window size and weighting greatly impact open-class POS formations. GloVe's hyperbolic window weighting is crucial for encoding POS information: linear weighting significantly reduces local and global POS structure particularly for larger window sizes. For SGNS, using hyperbolic weighting improves POS clustering, similar to using subword information.

Finally, we find that closed-class POS tokens affect open-class POS formations in complex and language-specific ways. Deleting large quantities of closed-class tokens especially harms adjective, noun, and verb clustering. High frequency downsampling dampens but does not eliminate this effect. Additionally, removing specific closed classes has varying effects on open-class POS formations, varying by language. Our methods provide a framework for capturing the interactions of POS types within embeddings across languages. We hope that this method may not only provide information about how embeddings encode linguistic information, but also serve as a tool for measuring and comparing the properties of natural languages.

5

COMPUTATIONAL CUT-UPS: THE INFLUENCE OF DADA

This chapter is based on joint work with David Mimno.

> To Make A Dadaist Poem:
>
> *Take a newspaper.*
>
> *Take a pair of scissors.*
>
> *Choose an article from the newspaper about the same length as you want your poem.*
>
> *Cut out the article.*
>
> *Then carefully cut out each word of the article and put them in a bag.*
>
> *Shake gently.*
>
> *Then take out each word, one after another.*
>
> *Copy them conscientiously in the order drawn.*
>
> *The poem will be like you.*
>
> *And look! You are an infinitely original writer with a charming sensibility, yet beyond the understand of the vulgar.*
>
> — *Tristan Tzara [1963]*

5.1 INTRODUCTION

Can a work of art that has been deformed beyond recognition nevertheless be recognizable? The idea of a cut-up poem is distinctively Dada in its playful reductiveness and, simultaneously, shockingly relevant. Today the cut-up is a foundation of modern textual analysis in the form of the "bag of word" assumption. Search engines, spam filters, and social media recommenders all rely on the assumption that the information carried by the words themselves is sufficient, and that the order in which words appear is irrelevant and burdensome. The bag-of-words assumption reduces the need for intelligence. All that is required is conscientiousness, which computers have in limitless quantities.

In this work we study a deformative "reading" of Dada, not using scissors but a modern computational image-processing method known as a convolutional neural network (CNN). We assure the reader that CNNs are both charming and quite beyond the understanding of the vulgar. CNNs operate by passing images through multiple layers of pattern detectors. The output of a given layer becomes the input of the next layer. For example, the output of the first layer might identify the presence of lines or edges at different angles, while the output of the second layer might identify the presence of pairs of lines that form angles. At the top layer, the output might identify specific things—dog breeds, dishwashers, doormats.[1]

Instead of physical newspapers, we cut up digitized avant-garde periodicals from Princeton's Blue Mountain Project.[2] Our initial corpus contains more than 2,500 issues from thirty-six different journals—over 60,000 pages in total. We deform page-level images into computational cut-ups using a CNN. We then use

[1] See `http://image-net.org/challenges/LSVRC/2014/browse-synsets` for a sample list of object classes used in object detection and image classification tasks.

[2] The transcripts contain human-generated metadata which describe the editorial content within a periodical and their corresponding page locations. See `http://bluemountain.princeton.edu`.

statistical classification to determine which visual features are captured by these cut-ups. Finally, we "read" these computational cut-ups to determine whether such reductive analyses are sufficient to separate Dada from other modernist movements. Our goal is not necessarily to get the "right" answer, but rather to use computation to provide an alien, defamiliarized perspective that can call into question the boundaries between categories.

5.2 CREATING AND READING COMPUTATIONAL CUT-UPS

A computational cut-up is a mathematical representation of an object: a list of numbers that collectively preserve information about the original object. Each value in this list corresponds to a computational feature. In a text model, a feature might correspond to the number of occurrences of a particular word in a document, but image features are more abstract and less apparent. Accordingly, we do not choose these features by hand—we ask a CNN.

CNNs are powerful tools for analyzing images. Although the output of the final layer of a CNN will identify the categories of the object that it was trained to recognize, the output of the next-to-last layer has been shown to produce powerful, high-level visual features. These features are generic enough that they can be used by other image analysis systems [Razavian et al., 2014]. By using these features as our computational cut-ups, we will in essence be asking what CNNs "see" when they look at Dada and more broadly the avant-garde.

Generating a computational cut-up involves two steps. We first shrink the input image to a small 224-by-224 pixel square so that it can be fed as input to our CNN. Fine-grained details are lost through this deformation, but major elements such as layout, headers, and illustrations are generally preserved. As seen in Figure 5.1, the

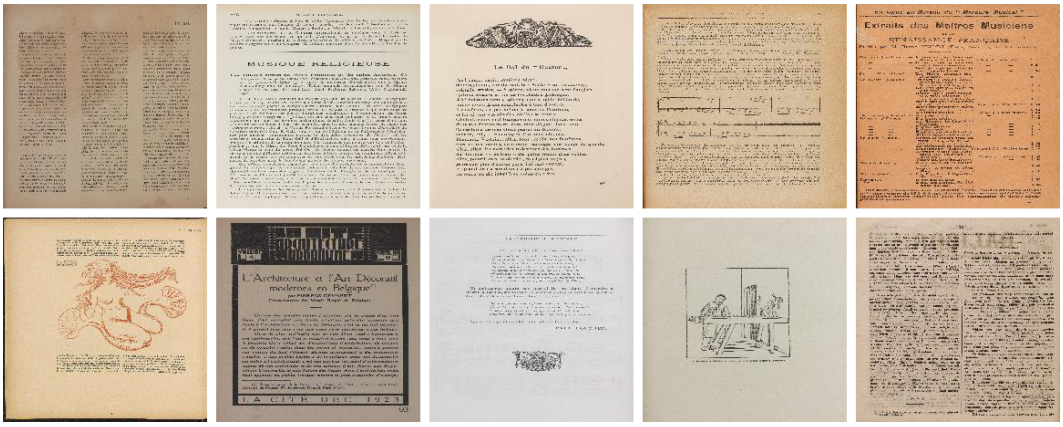

Figure 5.1: CNN input images for ten randomly sampled pages.

images fed to the CNN remain recognizable but are similar to viewing the page from the far side of a room.

Having created a shrunken version of the original page, we pass it to our CNN as input and extract a computational cut-up from the CNN's internal features.[3] These features are not readily interpretable to the human eye, but they correspond to high-level concepts such as human faces, flowers, and fields of grass.[4] For each feature, we extract a number representing the feature's measured presence within an image—a large value indicates the feature is strongly detected, while a value near zero indicates its absence. As a result, our computational cut-ups are lists of 2,048 numbers.

We evaluate the information encoded within the computational cut-ups by how well they perform on a series of binary classification tasks (e.g. Dada or not-Dada). We measure the degree to which a classifier can distinguish cut-ups with label A from cut-ups with label B. For each computational cut-up c with label A or B, we train a Naïve Bayes classifier on all other cut-ups with labels A and B and use it to predict the label for cut-up c [Broadwell et al., 2017]. The classifier consists of the

[3] We use the ResNet50 model pretrained on ImageNet, which is available through Keras.

[4] See `http://yosinski.com/deepvis#toolbox` for more information on visualizing neural network features.

mean and variance of each of the 2,048 CNN features, for each label. If the feature values of c look more similar to the typical feature values of label A, we predict A, and vice versa. The accuracy of these predictions will indicate how well the cut-ups differentiate the two label classes.

In addition to the simple question of whether a classifier is guessing correctly, we are also interested in how confidently the system makes its predictions. We therefore also measure classifier confidence for each prediction. By examining the corresponding page images for the best and worst predictions for each label, we can better understand the visual features being associated with each label.

5.3 PROOF OF CONCEPT: SEEING MUSIC

Before testing whether a CNN can recognize Dada, we verify that it is capable of performing a simpler task: identifying music within periodicals. It is fairly easy for a person to tell the difference between pages of musical scores and pages containing text and images, but how well will our CNN fare? If our computational cut-ups do not distinguish between musical scores and paintings, it would be hard to trust their capability to distinguish Dada from Cubism.

Detecting music within our corpus is a relevant task, not only because music is an avant-garde art form, but because the Blue Mountain corpus has a substantial number of music journals. The five periodicals *La Chronique musicale*, *Dalibor*, *Le Mercure* (*S.I.M.*), *Niederrheinische Musik-Zeitung*, and *La Revue musicale* are represented in the corpus by 1,405 issues and 27,791 pages. It is safe to assume that the majority of pages containing music will come from these five journals. Using the TEI-encoded transcriptions for each periodical issue, we identify 3,450 pages

containing music.[5] Only ninety-one of these pages come from the thirty-one other periodicals.

We find that computational cut-ups are useful for recognizing pages containing music. The classifier makes mistakes that a human might not, but in ways that provide intuition about what it "sees." The classifier correctly labels 67% of the 3,450 pages with music as "Music" and 96% of the 55,007 pages without music "Not-Music." For each prediction, we can measure our classifier's confidence in terms of how much more likely it thinks a page should be labeled as "Music" rather than "Not-Music." Confidence scores with large magnitudes indicate a more confident classification, while a score's sign indicates its assigned label type. So, a large, positive confidence score indicates that the classifier is very confident that a page be labeled "Music." In Figure 5.2 we see that our classifier tends to be more confident when it labels a page as "Not-Music," even when it is wrong. This difference suggests that the cut-ups may better describe features associated with non-music page elements than music page elements.

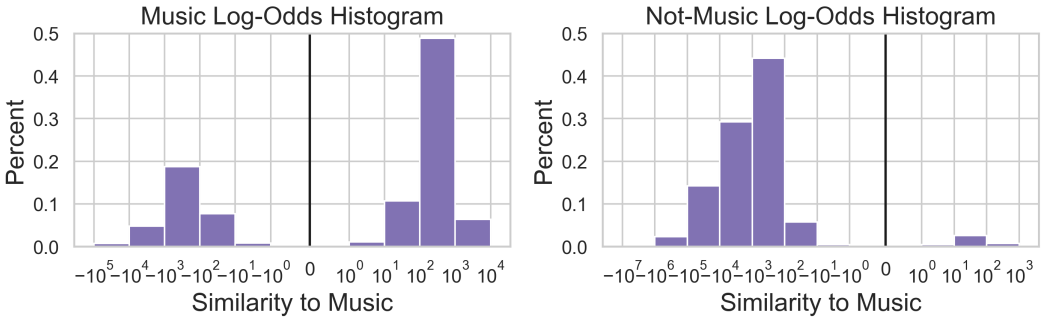

Figure 5.2: Histograms of prediction confidence for pages containing music (left) and pages without music (right). The classifier is more confident labeling pages as "Not-Music" no matter what the actual page type is.

[5]We consider content marked as "Music" to represent musical content within a page. See `https://github.com/cwulfman/bluemountain-transcriptions`. Our research used transcripts accessed in May 2017.

To understand where the classifier goes wrong we compare the pages that are most confidently classified and misclassified for each label. In Figures 5.3 and 5.4, we see that pages of sheet music are most confidently recognized as "Music" and tables are most confidently misclassified as "Music." These images share two prominent features: prominent horizontal lines and rectangular blank spaces. Given that the actual musical notes are poorly preserved in the deformed CNN inputs, it is reasonable that these are not the dominant visual features associated with pages containing music.

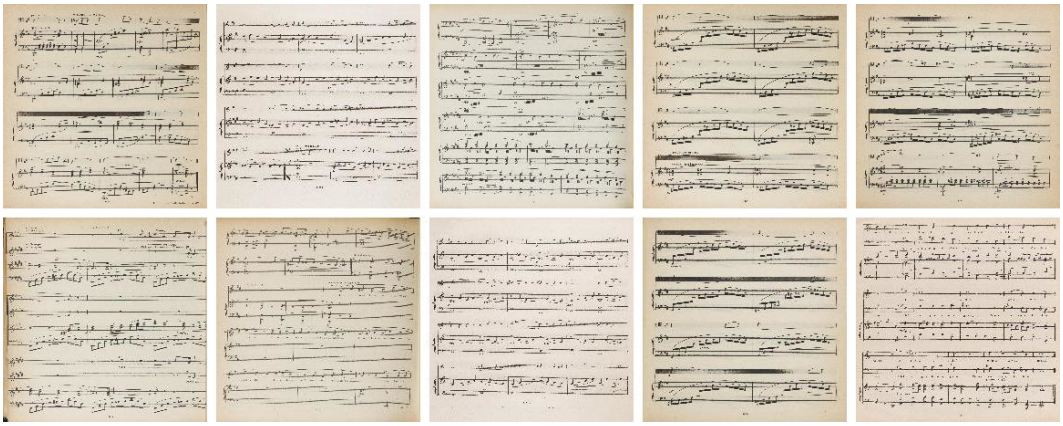

Figure 5.3: Ten pages most confidently, and correctly, classified as "Music."

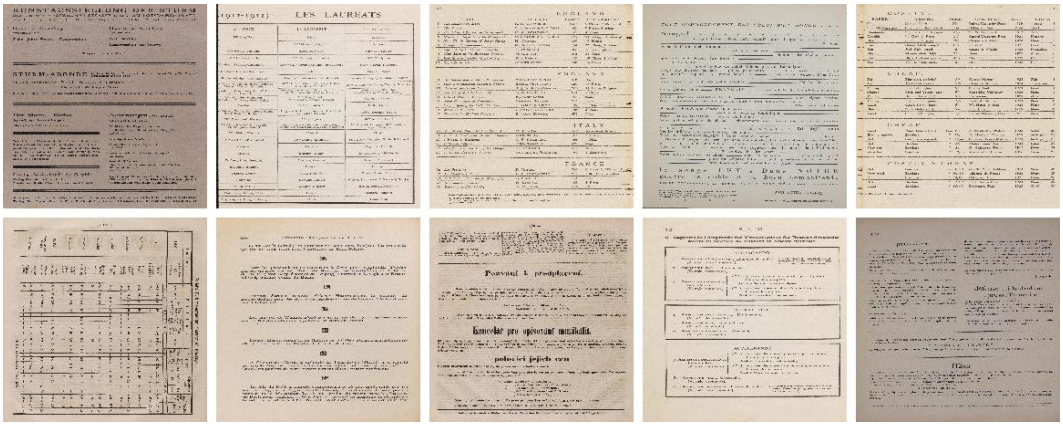

Figure 5.4: Ten pages most confidently misclassified as "Music."

Figure 5.5: Ten music-containing pages most confidently misclassified as "Not-Music."

Turning to the "Not-Music" label, we find color and pictures are the dominant visual features associated with pages without music. In Figure 5.5, we see that the top ten pages most confidently misclassified as "Not-Music" all contain pictures. Moreover, these pictures take up as much space within the page if not more than the musical elements. Many of these pages also include text.

Perhaps the most interesting of these confident "Not-Music" misclassifications is the bottom-right page in Figure 5.5, a scaled down image of a medieval folio. The rescaled CNN input image hardly looks like music, and, in a way, it is not. But looking at the original image in Figure 5.6, we see it does contain music, even though it looks nothing like modern musical notation. Additionally, the music is being seen through another medium: a picture, which could be misleading the classifier to the "Not-Music" label.

An effective, but potentially misleading feature learned by the classifier is that saturated color indicates no music. While sheet music is generally white, page coloring can vary due to paper and scanning quality. We want to verify that the non-color features perform well without the color cue, and see if the presence of pictures within a page remains the dominant "Not-Music" feature.

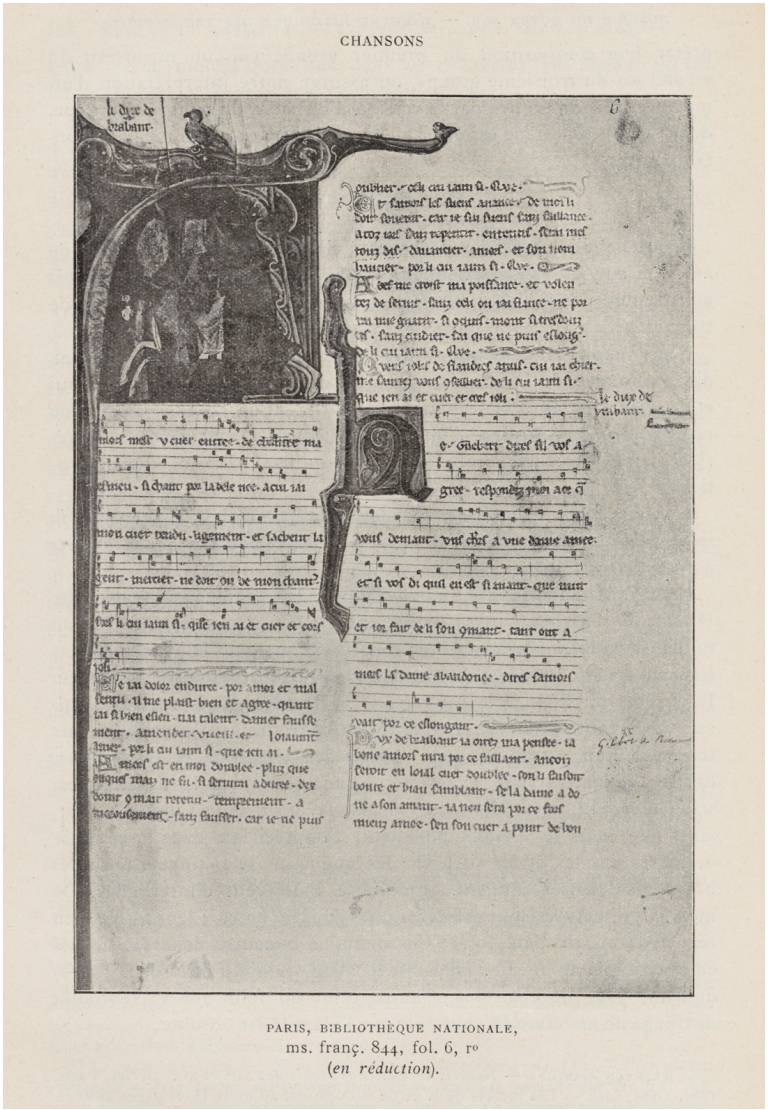

Figure 5.6: This medieval folio is confidently misclassified as "Not-Music."

We found that removing color from CNN inputs had little effect on classification performance: 66% of pages with music and 97% of pages without music were correctly labeled. Additionally, the most confidently classified and misclassified images remained largely the same for each scenario except for correctly classified pages without music. This is what we had hoped to observe. It indicates that the classifier relies on features other than color. By removing color we also confirmed that the presence of pictures is an important feature for pages without music. As seen in Figure 5.8, pages containing pictures—both illustrations and photographs—are considered the least musical.

Figure 5.7: Ten non-music pages most confidently classified as "Not-Music."

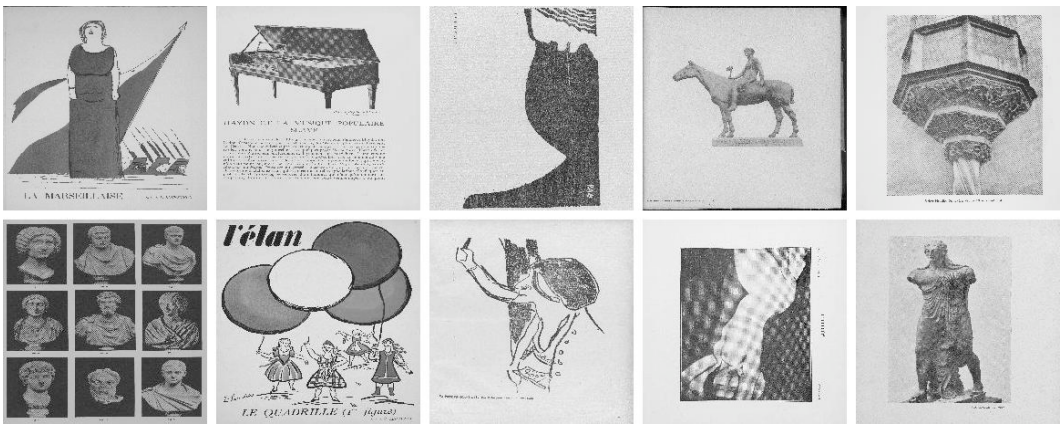

Figure 5.8: Ten grayscale pages correctly and most confidently classified as "Not-Music."

Through this exercise, we have shown that computational cut-ups are able to encode visual features that are useful for recognizing pages containing music. Pages with music tend to have regular horizontal lines and rectangular white space, while pages without music tend to contain pictures and be in color. These patterns are fairly primitive, but there is power in this simplicity that echoes the power of word counts to capture abstract textual concepts. Having established our analysis process, we begin our search for what makes Dada Dada.

5.4 DISTINGUISHING DADA

For our reading of Dada we begin with the question of whether we can distinguish "Dada" from "Not-Dada." We define labels at the periodical level: for the purposes of this study, *Dada*, *291*, *Proverbe*, and *Le coeur à barbe* are "Dada" and all other periodicals are "Not-Dada." We acknowledge that this is a particularly coarse-grained perspective. A number of periodicals may feature works of Dada artists in specific issues, and these four periodicals might not always feature Dada artists, but these mistakes should have little effect on our classifier given the volume of actual "Not-Dada" material.

We exclude the five music journals from our analysis. Their sheer volume in the Blue Mountain Project would likely drown out the visual features that we are most interested in finding. Moreover, we would like to avoid learning the naïve feature that Dada does not contain sheet music, and hopefully uncover more interesting distinctive features. After this exclusion, we have 32,642 pages labeled "Not-Dada" and 182 pages labeled "Dada."

We find that computational cut-ups are not perfect at distinguishing "Dada" from "Not-Dada," but they are better than random. The classifier correctly labels

63% of the Dada pages and 86% of the not-Dada pages. In Figure 5.9, we see that the classifier is, as with music, more confident about its "Not-Dada" predictions. We speculate that other avant-garde movements may have visual signals that are easier to identify than Dada.

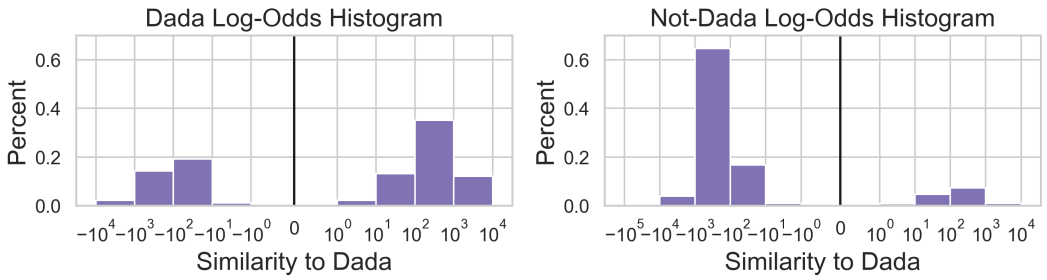

Figure 5.9: Histograms of prediction confidence for Dada (left) and not-Dada (right) pages. The classifier is more confident labeling pages as "Not Dada" no matter what the actual page type is.

What then does the classifier "see"? When examining the classifier's most confident successes and mistakes in Figures 5.10–5.13, we find that the low-level features associated with Dada are high contrast, prominent edges, and the color red. In comparison, graded texture and photographs are considered not-Dada. From these low-level features, we see that abstract human forms are generally associated with Dada, while more realistic forms are not.

Given the prominence of red in "Dada" labeled pages, we were concerned that our results were overly dependent on this simple variable, and not able to generalize to shape or texture. We therefore reran the same analysis on grayscale images to measure the overall effect of color. The classifier's accuracy worsens for both label groups with resulting accuracies of 56% for "Dada" and 84% for "Not-Dada." Since this degradation is relatively small, we conclude that color is an important feature for distinguishing Dada, but not the only feature. We find that contrast, edge sharpness, and texture all remain prominent features for classification in grayscale.

It is perhaps unsurprising that color would play a role in distinguishing periodical groups, since page color is influenced by both content and printing method. If a journal has a distinctive page coloring, then it can easily be distinguished from other periodicals by this color alone. This feature can both cause pages with ambiguous content to be correctly identified and pages with otherwise highly similar content to be easily distinguished because of differences in color palette.

We find that our classifier is most confident when labeling pages containing images, and least effective for text-only pages. This result suggests that the classifier is less certain about how text-only pages relate to Dada. Although the CNN does not appear to be able to distinguish between journals based on pages of text, we should not conclude that there are not typographic or layout features that could distinguish them, since these features may simply not be preserved in downscaled pages. We will keep this concept in mind as we analyze our results at the periodical series level.

With these intuitions about which features appear Dadaesque according to the CNN's deformative viewing, we can measure Dada at the level of an entire periodical series. Which periodicals "fool" the classifier, and therefore question

Figure 5.10: Ten Dada pages most confidently classified as "Dada."

Figure 5.11: Top 150 not-Dada pages most confidently misclassified as "Dada."

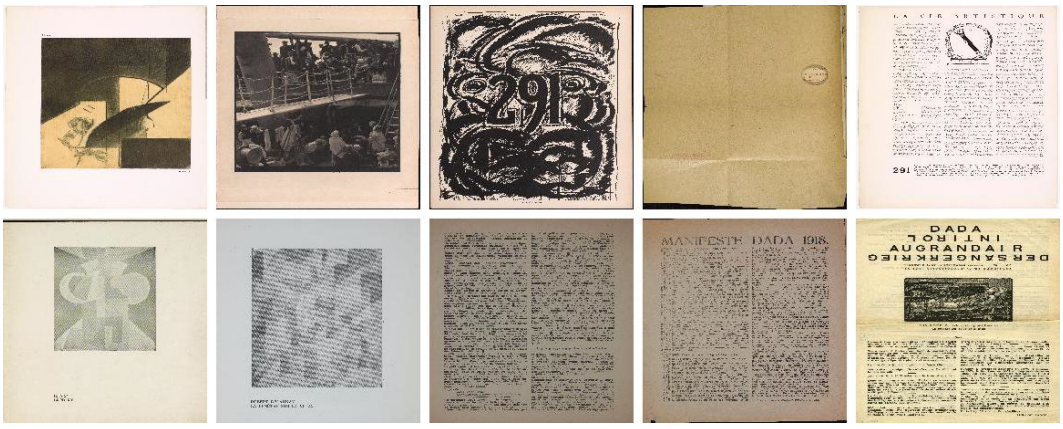

Figure 5.12: Ten Dada pages most confidently misclassified as "Not-Dada."

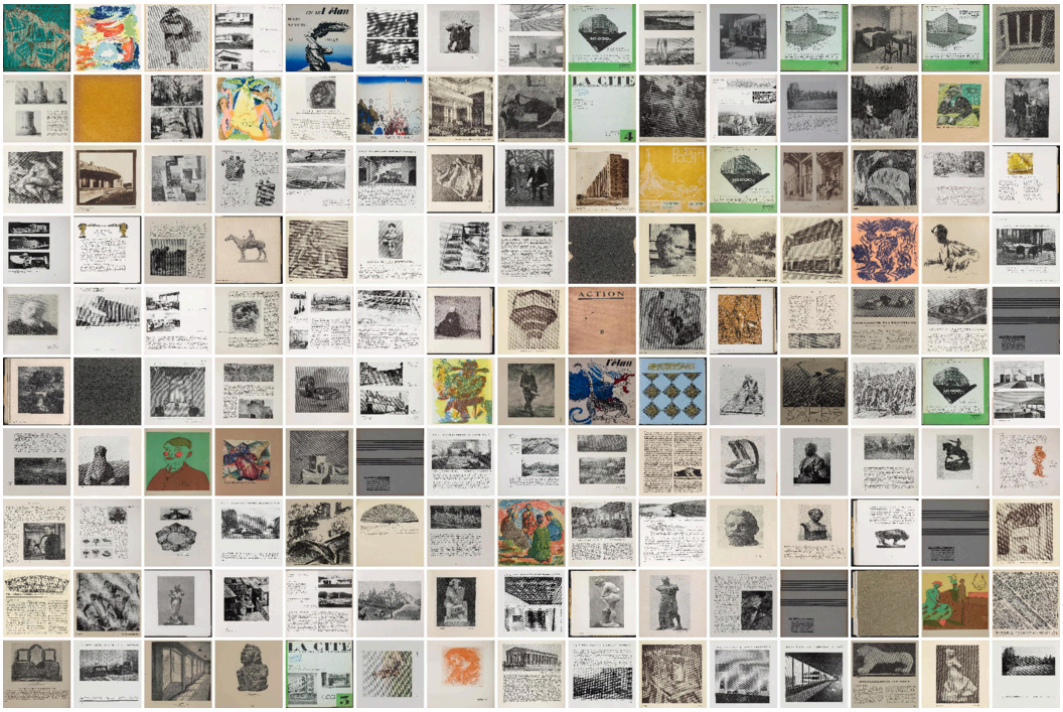

Figure 5.13: Top 150 not-Dada pages most confidently classified as "Not-Dada."

the (somewhat arbitrary) boundaries that we have constructed? We measure a periodical's closeness to Dada by the proportion of its pages classified as "Dada." As seen in Table 5.1, we find that the journals *Le coeur à barbe*, *Dada*, *291*, and *L'élan* are the most Dadaesque journals with over half of each periodical's pages classified as "Dada" for both color and grayscale images. Notably, *L'élan* is not a Dada periodical; it is a cubist war journal. In fact, we find a number of cubism and cubism-influenced journals among the most Dada-like not Dada, namely *Klingen*, *Der Sturm*, and *SIC*. We find this to be a positive result given cubism's influence on Dada, particularly Dada art.

We were surprised to find that the two Dada-related journals *Secession* and *Nord-Sud*, as well as the Surrealist journal *Surréalisme* were found to be not Dada-like. While *Nord-Sud*'s Dada proportion improves with grayscaling, it is still far below fifty percent. Examining these journals, we find that they are predominantly

composed of text. Given our suspicions that our classifier has difficulty correctly

Periodical	Total Pages	% Dada (Color)	% Dada (Grayscale)
Le coeur à barbe	8	100.00	75.00
Dada	110	70.91	60.91
291	42	59.52	57.14
Proverbe	22	18.18	18.18
L'élan	129	58.91	56.59
Klingen	727	26.55	31.09
Veshch Gegenstand Objet	64	25.00	29.69
Der Sturm	4649	22.54	24.71
La cité	4766	18.90	21.74
SIC	458	16.16	19.43
Ver sacrum	1928	16.08	21.27
Umělecký měsíčník	933	15.01	18.44
Bruno's Weekly	1219	14.68	15.42
Volné směry	1797	14.47	17.25
Zeit-Echo	756	12.70	17.06
New Numbers	222	11.71	10.36
Poesia	1603	9.61	10.61
Action	911	9.55	14.16
Broom	1751	9.42	10.34
Sturm-Bühne	32	9.38	21.88
Entretiens politiques & littéraires	2764	9.37	8.54
Ultra	156	7.69	11.54
The Mask	3980	7.66	9.92
Nord-Sud	246	7.32	14.63
The Glebe	940	7.02	9.68
Surréalisme	16	6.25	6.25
Secession	217	5.07	5.53
Nowa sztuka	76	3.95	10.53
The Signature	96	3.13	9.38
Pan	2136	3.09	8.19
East & West	70	1.43	0.00

Table 5.1: Periodical-level Dada-like proportions for all pages.

classifying text-only content, we narrow our measurements to pages containing images.

We identify pages containing images using the periodical TEI transcripts similar to our identification of pages containing music. If a page contains content marked as an illustration, then we consider it an "illustrated page."[6] We note that these annotations make different judgments on what content constitutes illustrated content and, as a result, smaller, more decorative illustrations are included inconsistently. Nevertheless, this labeling is sufficient to allow us to demonstrate patterns; future work could specifically analyze the images within journals. We find around one-third of the pages are illustrated, with a very uneven distribution across periodicals.

As we narrow our focus to "illustrated" pages, the Dada-like page proportions increase across journals as shown in Table 5.2. Encouragingly, the Dada-like journals found in the overall page set remain Dada-like. Now, *Surréalisme*, *Secession*, and *Nord-Sud* have much higher Dada percentages, although these three journals are represented by very few pages. Moreover, the more illustrated journals *Der Sturm* and *SIC* have a high Dada page proportion.

5.5 CONCLUSION

A deformative technique such as the cut-up poem seeks meaning in the visible features of language, while playfully ignoring the original concepts and intentions of the text. By "reading" data art with Convolutional Neural Networks (CNNs), we can take the same approach, isolating ourselves from concepts and intentions, and accessing only visual features. The CNN can only describe and distinguish,

[6] We consider content marked as "Illustration" or "ComplexIllustration" to represent images within a page.

not define. We find that CNNs indeed enable a deformative viewing of modernist journals. This in itself is not surprising: a tool that analyzes images analyzed images.

Periodical	"Illustrated" Pages	% Dada (Color)	% Dada (Grayscale)
Le coeur à barbe	1	100.00	100.00
Dada	61	85.25	85.25
291	31	74.19	67.74
Proverbe	3	33.33	33.33
Surréalisme	1	100.00	100.00
Secession	7	100.00	85.71
Der Sturm	1034	74.76	79.40
SIC	83	71.08	78.31
L'élan	87	66.67	73.56
Broom	238	52.94	60.92
Nord-Sud	4	50.00	100.00
Poesia	136	47.06	46.32
Bruno's Weekly	317	45.11	44.79
Ultra	23	39.13	56.52
Zeit-Echo	213	38.97	47.89
Klingen	464	36.85	44.83
La cité	1784	32.23	40.30
Veshch Gegenstand Objet	23	30.43	34.78
Umělecký měsíčník	395	27.85	34.49
Action	222	25.23	40.99
Ver sacrum	1415	16.11	22.83
The Mask	1500	14.20	17.73
Volné směry	1286	12.60	18.04
Pan	1041	5.76	15.37
East & West	1	0.00	0.00
Entretiens politiques & littéraires	1	0.00	0.00
Nowa sztuka	1	0.00	0.00
The Glebe	0	N/A	N/A
New Numbers	0	N/A	N/A
The Signature	0	N/A	N/A
Sturm-Bühne	0	N/A	N/A

Table 5.2: Periodical-level Dada-like proportions for "illustrated" pages.

What is critical from a scholarly perspective is whether this deformative reading provides a perspective that is both distinct from and complementary to human reading. Can a tool designed for identifying dogs be repurposed for exploring the avant-garde? Can it see Dada among the rest?

From the perspective of a computational cut-up of Dada journal pages, we find that pages of Dada journals can be distinguished from pages of non-Dada journals with a degree of accuracy that exceeds random chance. This suggests that there is substance behind the name of Dada. The internal state of neural networks is notoriously inscrutable, but by sorting pages by predicted Dada-ness, we can start to infer how the machine "sees" Dada both from its successes and its mistakes. Dada is characterized (not defined) by red hues, sharp and prominent edges, and high contrast. These features are simplistic but can be combined to form more complex structures such as schematic-like figures and abstract human forms. The pages that the CNN thinks really ought to be Dada show the porous boundaries of the category: cubism thus appears to be measurably the closest movement to Dada at the level of simple visual features.

This characterization of Dada is both alien and familiar. It is produced by an alien gaze, by a machine trained to identify image content ranging from specific dog breeds to microwaves and guillotines. From this machine-view, we gain abstract, but at times unfamiliar, features that nonetheless reflect human concepts. In the case of Dada, the CNN directs our attention to the presence of abstract forms and schematic drawings, and strongly away from photography and more realistic representations of the body. Is this a machine's way of separating art from anti-art?

A potential shortcoming or strength of this machine reading is its illiteracy. The CNN was not trained to read human language; moreover downscaling images makes text largely illegible if not invisible. This prevents the CNN from cheating by associating Dada with the name of the movement or artists associated with

it. Instead, it must find visual cues that are significant to Dada journals alone. In all likelihood, this causes the CNN to fixate on particular artists and their styles. Clearly, it will take art at face value and not read into the intent of the artist. However, this deconstruction of art echoes the effects of Dada itself.

As with all scholarship, but particularly data-driven scholarship, our analysis is limited by the scope of the collection. This reading primarily focused on the pictures contained within a page, and on the journals present. The former is a shortcoming of the CNN, while the latter is of the data itself. The CNN is unlikely to associate the poems of Tristan Tzara or the readymades of Marcel Duchamp as Dada because its attention is focused away from texts and photographs. Similarly, it cannot associate other avant-garde art with Dada that it never sees. Despite these shortcomings, we come upon an interesting and believable finding. From the perspective of the CNN and this collection, Dada is most similar to cubism. Unfortunately, potential connections to surrealism could not be observed because the surrealist journals we included had no pictures.

The idea of viewing art with computers necessarily implies a reductive and even ludic perspective. The choice of Dada as a testbed for this approach is quite deliberate, and one we hope fits with the spirit of the movement. The characteristics of Dada learned by the CNN may simply be artifacts of printing choices, and almost certainly "miss the point" at a conceptual level. But they also force us to recognize the visible, structural characteristics of Dada art, and more importantly, point us to the potential connections and influences of the movement outside Dada proper. The CNN-based classifier is like Dada, but has its own sensibility. Perhaps in explaining its successes and puzzling over its mistakes we may ourselves become infinitely original.

Part II

HOW CAN WE DIRECT WHAT MODELS LEARN?

6

AUTHORLESS TOPIC MODELS: BIASING MODELS AWAY FROM KNOWN STRUCTURE

This chapter is based on joint work with David Mimno.

6.1 INTRODUCTION

Unsupervised semantic models are a popular and useful method for inferring low-dimensional representations of large text collections. Examples of such models include latent semantic analysis [Deerwester et al., 1990] and word embeddings [Bengio et al., 2003], but for this work we will focus on statistical topic models [Hofmann, 1999; Blei et al., 2003], which are used to infer word distributions that correspond to recognizable themes. In practice, collections are often constructed by combining documents from multiple sources, which may have distinctive style and vocabulary. This heterogeneity of sources leads to a serious but rarely studied problem: the strongest, most prominent patterns in a collection may simply repeat the known structure of the corpus. Instead of finding informative, cross-cutting themes, models simply repeat the distinctive vocabulary of the individual sources. The model in this case is "correct" in that it has detected the strongest dimensions of variation, but it tells us nothing we did not already know.

As a motivating example, we focus on models trained on novels, where it is known that inferred topics are often simply names of characters and settings [Jockers, 2013]. The words *Harry*, *Ron*, and *Hermione* look to the algorithm like the

basis of an ideal topic because they occur very frequently together but not in other contexts. But this topic only tells us which books within a larger corpus are part of the *Harry Potter* series; themes like friendship, adolescence, and magic remain hidden. This phenomenon is not limited to fiction: we also include a case study of opinions from US state supreme courts. Unlike examples from fiction, Maine and Utah both exist in the same universe, but exhibit specific regional term use.

We begin by demonstrating that the problem of overly source-specific topics is both substantial and measurable. We present three metrics that provide related but distinct views of source specificity. These metrics are orthogonal to existing metrics of topic semantic quality: uselessly source-specific topics are often still highly coherent and meaningful. These metrics are also inversely related to commonly-used document classification evaluations. Learning 20 newsgroup-specific topics from 20 Newsgroups may be informative as an evaluation, but in practice users are rarely unaware of such structure.

Finally, we present a simple but effective method for reducing the prevalence of source-specific topics. This method relies on probabilistically subsampling words that correlate with known source metadata, and is related to subsampling methods that have been highly effective in word embeddings [Mikolov et al., 2013b; Levy et al., 2015]. The best of the proposed methods substantially reduces source-specific topics, increases topic differentiation without increasing model complexity, and improves topic stability.

6.2 RELATED WORK

The common assumption of prior work on metadata-aware topic modeling has been that metadata provides valuable hints that can be used to improve topics. Several

methods use document metadata to influence document-level topic distributions. The author-topic model [Rosen-Zvi et al., 2004], relational topic model [Chang and Blei, 2009], and labeled LDA [Ramage et al., 2009] extend LDA by directly incorporating a particular type of metadata (e.g. author information, document links, user-generated tags) into the model. Others, like factorial LDA [Paul and Dredze, 2012], Dirichlet-multinomial regression topic models [Mimno and McCallum, 2008], and structural topic models [Roberts et al., 2014] incorporate more general categories of metadata. All of these aim to *increase* dependence between topics and metadata. In contrast, our goal is to make topics *independent* of specified metadata.

Other research makes topic-word distributions sensitive to document-level metadata. The special words with background model [Chemudugunta et al., 2006] incorporates document-specific word distributions into LDA, while cross-collection LDA [Paul, 2009] incorporates collection level word distributions. The topic-aspect model [Paul and Girju, 2010] extends LDA to include a mixture of aspects of documents such that aspects affect all topics similarly. Although these models may be able to sequester author-specific words, there is no reason to expect that those words will not also drag along general, cross-cutting words.

In this paper we focus on ways to explicitly identify words that bias topics towards a specific metadata tag and modify the input corpus for an algorithm to reduce their effect. Researchers have often dismissed this sort of data curation as unprincipled and heuristic "preprocessing." More recent work [Denny and Spirling, 2017; Boyd-Graber et al., 2014] emphasizes that *meta-algorithms* for data preparation can greatly affect the intrinsic model quality and human interpretability of topic models.

Corpus	Authors	Docs	Types	Avg Len
SCI-FI	245	327K	132K	153
COURTS	50	52K	89K	1039

Table 6.1: Corpus statistics for the number of authors, documents, and word types, as well as average document length. Document and word type counts are listed in thousands (K).

6.3 COLLECTIONS AND MODELS

We collected two real-world corpora that combine text from multiple distinct sources: science fiction novels and U.S. state supreme court opinions.[1]

SCIENCE FICTION (SCI-FI). We selected 1206 science fiction novels by 245 authors based on award nominations and curated book lists hosted on Worlds Without End.[2] We consider each author as a source, and treat collaborations as distinct sources. We augmented the corpus with other established authors to increase the diversity of author gender and ethnicity. The novels span from the early 1800s to the present day. Most of these works are currently protected by copyright, so rather than full text we obtained page-level word frequency statistics from the HathiTrust Research Center's Extracted Features Dataset [Capitanu et al., 2016]. This data indicates, for example, that page 227 of *Dune* contains one instance of the word *storm* as a noun. Following previous work [Jockers, 2013] we divide volume-length works into page-level segments, omitting headers and footers.

U.S. STATE SUPREME COURTS (COURTS). Each U.S. state has a supreme court that decides appeals for decisions made by lower state courts. In this collection each document is a court opinion, written by the court after the completion of a

[1] Code and data is available at `https://github.com/laurejt/authorless-tms`.

[2] `https://www.worldswithoutend.com/lists.asp`

case, summarizes the case and judgment. We treat each state court as a source, expecting that courts use geographically specific language (e.g. Colorado, Denver, Colo., Boulder) that is not relevant to the legal content of opinions. We examine court opinions for all 50 state supreme courts for cases filed from 2012 through 2016.[3]

DATA PREPARATION. We apply the same initial treatment to both corpora. Tokens are three or more letter characters with possible internal punctuation (excluding em- and en-dashes). Words are lower-cased. To deal with globally frequent terms, we remove words used by more than 25% of documents in a corpus. To reduce the computational burden of a large vocabulary, we remove words occurring in fewer than five documents. We remove all documents with fewer than 20 tokens. This process removes 706 pages and 9192 court opinions from our starting science fiction and state courts corpora.

We train LDA models using Mallet [McCallum, 2002] with hyperparameter optimization occurring every 20 intervals after the first 50. We set the number of topics to be on the same order as the number of sources, so for SCI-FI we use $K \in [125, 250, 375]$ and for COURTS we use $K \in [25, 50, 75]$.

6.4 EVALUATING TOPIC-AUTHOR CORRELATION

We introduce three ways to measure the source-specificity of topics. For concreteness we will use the terms "source" and "author" interchangeably, but a document's source could be any categorical variable. We want to identify topics that are used by relatively few authors, and more specifically topics whose "meaning" is unduly influenced by the contributions of relatively few authors.

[3] https://www.courtlistener.com

Given a collection of D documents written by A authors such that each document d is written by a single author a, we train an LDA topic model with K topics. Then for each word token i in document d we have both a word type w_{id} and a posterior distribution over its token-level topic assignment z_{di}. For clarity of presentation we can assume a single topic assignment for each token and view the corpus as a data table with three columns: word type w, topic z, and author a. By summing over rows of this table we can define marginal count variables for authors $N(a)$ and topics $N(k)$ as well as joint count variables for the count of a word in a topic $N(w,k)$, a topic in an author $N(k,a)$, and a word in a topic in an author $N(w,k,a)$. A maximum likelihood estimate of the probability of word w given topic k is $P(w \mid k) = \frac{N(w,k)}{N(k)}$.[4]

We note that these statistics must be defined at the token level. As in Mimno and Blei [2011] we are looking for violations of the assumption that $\Pr(w \mid k) = \Pr(w \mid d,k)$. Gibbs sampling algorithms typically preserve token-level information in the form of sampling states, but EM-based algorithms often preserve only document-topic distributions θ_d and topic-word distributions ϕ_k. We can estimate the posterior distribution over topic assignments for each token in document d with word type w as $\Pr(z \mid d,k) \propto \sum_k \phi_k(w)\theta_d(k)$, and generate sparse representations by sampling from this distribution.

AUTHOR ENTROPY. We begin by measuring a topic's author diversity—how evenly its tokens are spread across authors—using the conditional entropy of authors given a topic (Eq. 6.1). Topics whose tokens are largely concentrated within a few authors will have low entropy, while topics more evenly spread across many authors will have high entropy. With asymmetric hyperparameter optimization we

[4]We do not use Dirichlet smoothing for the purposes of this work for simplicity and to make more reliable comparisons across varying vocabulary sizes. Results using smoothing are similar.

find that the most frequent topics (large α_k) have high author entropy, but topics with high author entropy can have a wide range of frequencies: topics can be both rare and well-distributed.

$$H(A \mid k) = \sum_a \Pr(a \mid k) \log_2 \Pr(a \mid k) = \sum_a \frac{N(a,k)}{N(k)} \log_2 \frac{N(a,k)}{N(k)} \tag{6.1}$$

While author entropy provides a general sense of author diversity, it does not take into account the expression of topics by authors. Content-based evaluation is especially important because many collections are not well balanced across authors. The fact that a topic is not balanced across authors does not necessarily imply that it is problematic. A novel about the voyages of a ship captain may contain a large proportion of words about sea travel and ships, while a novel that contains one minor character who is a ship captain may contain a small proportion of the same language, used in the same way. We therefore need to be able to distinguish two cases: first, a topic that is consistent across authors but that is used at different rates by different authors, and second, a topic that is not only used at different rates but has different contents across authors. In the first case we can accurately use a topic to "stand for" a particular concept of interest, while in the second case we would get a false impression of the contents of documents, because the expression of the topic in the minority authors differs from the topic as a whole.

To differentiate expected author imbalance from pathological cases, we calculate Jensen-Shannon divergence between a topic's word distribution as estimated from the full collection $\Pr(w|k)$ and two distributions that have been transformed to reduce the influence of the most prominent authors. If the topic has low author correlation then there will be little divergence between the original distribution and its transformation. This method mimics a technique for identifying "junk" topics by AlSumait et al. [2009].

MINUS MAJOR AUTHOR. The first transformed distribution M (Eq. 6.2) recalculates the probability of words based on all documents *except* those written by the majority author. If a topic is consistent across authors then the presence or absence of its largest author contribution (labeled a_{major}) should have little effect on the topic's word distribution. The larger the resulting divergence, the more influence the major author has over the topic. Unlike author entropy, this technique does not inherently favor balanced distributions of authors; a very author-imbalanced (low entropy) topic can still have a low minus major author divergence if the dominating author's contribution agrees with the remaining topic tokens.

$$\Pr(w \mid M_k) = \Pr(w \mid \neg a_{major}, k) = \frac{N(w,k) - N(w, a_{major}, k)}{N(k) - N(a_{major}, k)} \tag{6.2}$$

BALANCED AUTHORS. The second transformed distribution B (Eq. 6.3) treats the contribution of each author equally, no matter how many words in that topic the author produces. The minus-major metric is most sensitive to the case where a single author dominates a topic, but does not handle the case where a small group of authors dominates. Using the balanced transformation we measure the similarity of each author contribution. The larger the resulting divergence between the original and transformed word distributions, the larger the variance in contributing author token usage.

$$\Pr(w \mid B_k) \propto \sum_a \Pr(w \mid k, a) = \sum_a \frac{N(w,k,a)}{N(k,a)} \tag{6.3}$$

We check the validity of our metrics by evaluating topic models trained on SCI-FI for a wide range of topic sizes (125–1000). As seen in Figure 6.1, all three measures produce bimodal distributions for all topic sizes, combining highly author-specific topics and more general cross-cutting ones. The proportion of cross-cutting topics

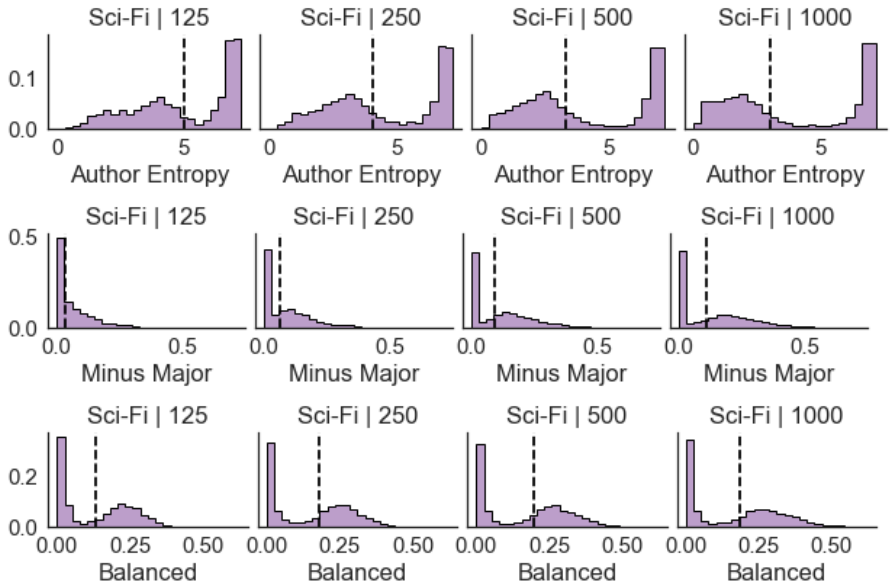

Figure 6.1: Author entropy, minus major author divergence, and balanced author divergence for topics in topic models trained on Sci-Fi. Dashed lines indicate medians. Increasing the number of topics in a model does not reduce the proportion of author-specific topics.

remains fairly constant across topic sizes: for all of these models, over 50% of topics fall in the source-specific range. We emphasize that source-specific topics are not necessarily "bad". If the structure of the corpus were not known, these topics would provide a highly useful and coherent insight into that structure. But if, as is typical, the structure *is* known, more than half of the statistical capacity of these models is wasted learning distributions that simply reiterate known structure, regardless of the number of topics.

While all three measurements produce similarly shaped distributions, they do not always agree in detail. Table 6.2 shows example topics that provide intuition for these differences. At the extremes, Topic A is a general, cross-cutting topic while Topic G is dramatically author-specific. While all three metrics score well for Topics A and B, in Topic B the word *paul* seems out of place, but it is common enough in several authors that its word-level author entropy is not low. Topics E and G

Topic	Entropy	Minus Major	Balanced	Top Words
A	6.79	0.00067	0.017	school professor work university years research science students student college
B	6.67	0.0047	0.032	doctor paul hospital nurse patient medical patients doctors room ward bed drugs
C	5.44	0.043	0.17	jack emma **malenfant** trip janet michael ing **wireman leonard nemoto** sally **jeannine**
D	5.31	0.027	0.13	sand **pirx** mars desert roger dust rock **bass** dunes crater martian **jeffries kirov** dune
E	3.42	0.080	0.16	robot robots andrew human **cully** susan **calvin** brain being **powell donovan** law
F	2.32	0.067	0.083	old night yes cried town last men **rocket** god years hands house upon stood wind boy
G	0.28	0.35	0.32	**f'lar lessa weyr robinton hold dragon f'nor lord dragons benden rider bronze harper**

Table 6.2: Topics from a 250-topic model trained on Sci-Fi and their corresponding measures of author entropy, minus major author, and balanced authors. Underlined values indicate poor quality scores and bolded terms indicate word types with low author diversity within the topic.

both score poorly in all three metrics, and both are highly specific to single authors (Isaac Asimov and Anne McCaffrey). But while G is clearly and exclusively names and settings, E contains the common terms *robot*, *robots*, and *human*, and could be confused for a general topic on artificial intelligence.

The metrics are also enlightening when they disagree. Topic C has high author entropy, but only because it mixes highly author-specific words from several different authors. Since each author's contribution differs from the others it scores poorly on the two content-based metrics. Topic D is partially about Mars, but also contains author-specific character names from stories set on Mars. No single author dominates, but the contributions of each author look different. Topic F is so highly correlated with Ray Bradbury that its entropy is low and it looks different when his contribution is removed, but its words are sufficiently general that Bradbury's use of the topic is close to the other authors' (minimal) use.

6.5 CONTEXTUAL PROBABILISTIC SUBSAMPLING

In this section we present interventions that predict the effect of words and contexts, and modify an input corpus to reduce the number of overly author-specific topics in resulting models. We hypothesize that this problem is due to *burstiness* [Doyle and Elkan, n.d.]: words that are globally rare, but locally frequent. Dampening the author-specificity of individual word types may reduce their connection to document sources. We therefore evaluate context-specific subsampling prior to modeling, with parameters based on tail probabilities of word-specific parametric models.

In selecting this particular approach we follow three design principles that we believe maximize use in actual practice. First, we want interventions to be minimal and have the least possible disruption to current work processes. We therefore choose to focus on meta-algorithms for data preparation that are compatible with but independent from existing, widely implemented inference algorithms. Second, we want any user-specified parameter choices to be simple and intuitive. Although we find that entropy is a useful diagnostic metric, information theoretic metrics such as mutual information are difficult for non-experts to interpret correctly, and critical values can differ widely across collections and dimensionalities. Third, we want both the choice of interventions and the effects of interventions to be transparent to users. We initially considered methods such as adversarially trained autoencoders, but we find that directly subsampling words is much faster, simpler, and easier to explain.

IDENTIFYING AUTHOR SPECIFIC TERMS. The simplest way to find author-specific terms is to find terms unique to an author. The SCI-FI collection contains an

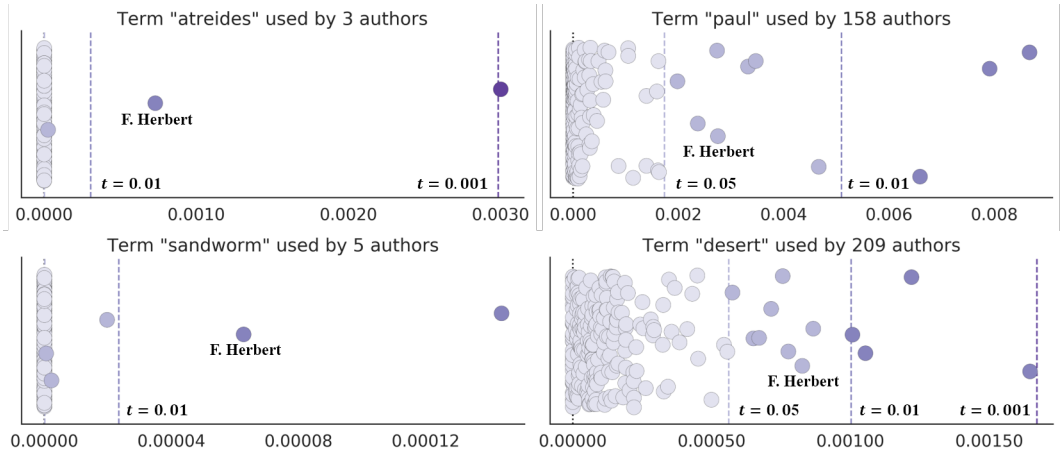

Figure 6.2: Reasonable threshold values t flag both rare words (left) and common words being used in author-specific ways (right). Each point represents the relative frequency of a term (x-axis) for an author (y-axis) in SCI-FI.

unusual number of author-specific coinages, but words used by many authors can still be highly correlated with a particular author. We therefore estimate parametric distributions for each term and compare author-specific term proportions to this distribution. For each word type w, we calculate the sample mean $\bar{x}_w$ and variance s_w^2 and construct a gamma distribution Γ_w with shape $k = \bar{x}_w^2 / s_w^2$ and rate $\theta = s_w^2 / \bar{x}_w$. Similar to a significance test, given a user-specified probability threshold t we can define a critical term proportion value under Γ_w

$$\Pr[\Gamma_w \leq f_w^*] \leq 1 - t. \tag{6.4}$$

A word w is thus considered too specific to an author a if a's usage is too unlikely to occur according to Γ_w. Specifically, this occurs when the frequency $f_{w,a}$ is larger than the cutoff frequency f_w^* defined in Eq. 3. This method satisfies our design goals of simplicity and transparency: the threshold is intuitive and can be adjusted to change how aggressively words are flagged for curation.

Figure 6.2 shows two character names and nouns from Frank Herbert's *Dune*, where one name and noun are rare and the others are frequent. We see that the rare words *atreides* and *sandworm* are significant to Frank Herbert for $t = 0.01$: there is essentially no "normal" level of use of these words in other authors. Herbert also uses the more common terms *paul* and *desert* more than expected, but to a lesser extreme.

DETERMINING STOP RATES. How we choose to dampen author-specific words is as important as how we detect them. If we globally removed these words using a traditional stoplist, we would lose a substantial portion of the vocabulary. A more sophisticated approach is to construct a stoplist for each author. In this case, words are only removed from contexts in which they are statistically overrepresented. For rare terms, where there is no middle ground between significant use and no use at all, this contextual treatment is effectively the same as a traditional stoplist. But for a word with more widespread use, that word would disappear only from contexts with abnormally high usage.

While this technique avoids erasing the majority of a collection's vocabulary, it leads to a paradoxical situation where a word that is thematically central to a work occurs *less* frequently in that work than in other works. Entirely removing *desert* from Frank Herbert or *robot* from Isaac Asimov would reduce the model's ability to identify relevant themes.

To find a middle ground, we use probabilistic subsampling to reduce outlier author use to something more in line with the collection's overall usage. We use the same threshold t to set subsampling rates. For a word type w and author a the probability of stopping a token of type w in a is

$$\Pr(\text{Stop } w \text{ in } a) = 1 - f_w^* / f_{w,a}. \tag{6.5}$$

The threshold t specifies when an author's use of a word is too extreme for our model Γ_w. If we reduce these outlier frequencies to their corresponding cutoff frequencies f^*_w, they will be set to the largest below-threshold frequency dictated by Γ_w. We construct our subsampling rates such that in expectation new author frequencies will equal their corresponding threshold frequency from the original distribution.[5]

6.6 RESULTS

Unless otherwise noted, we refer to models with a topic size of 250 for SCI-FI and 50 for COURTS, and set the hyperparameter t of context-based methods to 0.05. We refer to a treatment with no intervention beyond standard stopword removal as **None**. We compare these models to three classes of curation methods, each with varying parameters. **AF-[n]** removes all terms that are used by at most n authors. **C-[t]** removes any term from author a's context whose frequency $f_{w,a}$ exceeds significance threshold t with respect to distribution Γ_w. **CP-[t]** subsamples terms according to Eq. 6.5. We train 10 runs with random initializations for each parameter setting.

SUBSAMPLING REDUCES TOPIC-METADATA CORRELATION. We begin by measuring how well the curation techniques reduce the formation of author-correlated topics. We find that removing words with low author frequency has little effect, while contextual methods greatly reduce the formation of "bad" topics according to all three measures. As expected, the value of the threshold t affects performance of the context-based methods. In Figure 6.3, we see that lowest values of t are ineffective; $t = 0.001$ is hardly distinguishable from NONE and $t = 0.005$ is

[5] Iteratively reevaluating Γ_w leads to an unstable "race to the bottom."

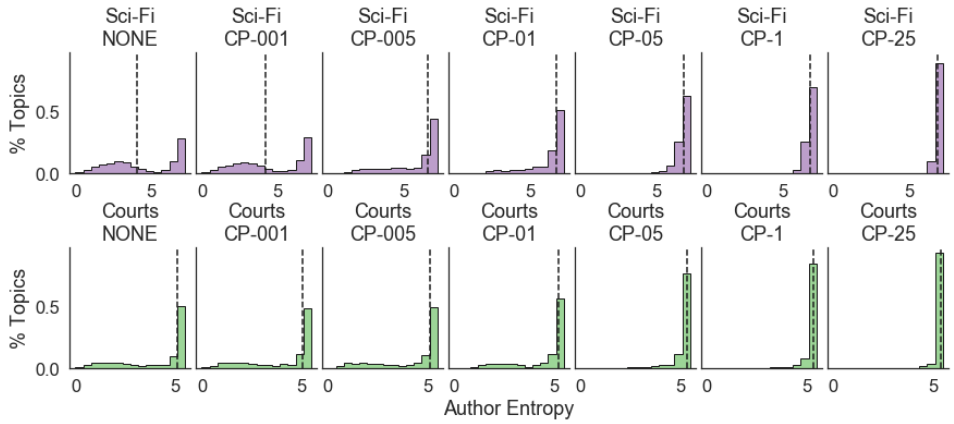

Figure 6.3: Increasing the threshold t for contextual probabilistic (CP) subsampling results in more topics with high dispersion over authors.

on par with low author frequency stoplists. We observe that settings of $t \geq 0.05$ perform very well, and choose this value as a default in our public code release.

Subsampling before inference does more than change the appearance of topics, it changes the content of the inferred topics. To test whether subsampling after inference has the same effect we construct ten additional models by *post hoc* stopping the 250-topic trained models for NONE-treated Sci-Fi to match token-for-token the CP-05 curated versions. We find that *post-hoc* removal has little effect on topic-metadata correlation; over twenty percent of topics are dominated by a single author with the worst having 96.4% of tokens contributed by one author.

SEMANTIC QUALITY IS PRESERVED. We define author-specificity as a property orthogonal to model quality: there is nothing fundamentally wrong with a topic full of character names outside the context of specific user needs. But ideally in reducing the prevalence of overly author-specific topics we would replace them with equally meaningful ones. We measure semantic quality of topics using Mimno et al. [2011]'s topic coherence metric as reported by Mallet. This metric measures

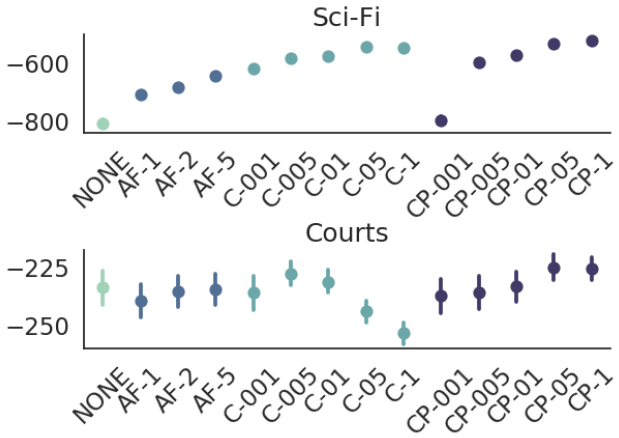

Figure 6.4: Contextual probabilistic subsampling improves mean topic coherence for Sci-Fi despite the removal of frequent words. Coherence degrades under context curation for Courts.

the tendency for the most probable (top) words of a topic to cooccur. A topic k with m top words $w_{k,1}, \ldots, w_{k,m}$ has topic coherence

$$\sum_i \sum_{j<i} \log \frac{D(w_i, w_j) + \beta}{D(w_i)}, \tag{6.6}$$

where D represents the number of documents containing a word or word pair and β is the LDA hyperparameter for topic-word smoothing. Large negative values indicate that the top words of a topic seldom cooccur, while values close to zero indicate that the top words frequently coccur.

We find that despite substantial changes in topic content, corpus modification has no consistent effect on the semantic quality of topics. In Figure 6.4, we find that all curation methods except CP-001 have significantly higher mean topic coherence than None for Sci-Fi. Contextual methods with $t \geq 0.05$ have the highest coherence. For Courts, topic coherence is maintained across treatments, except for the most aggressive interventions C-05 and C-1.

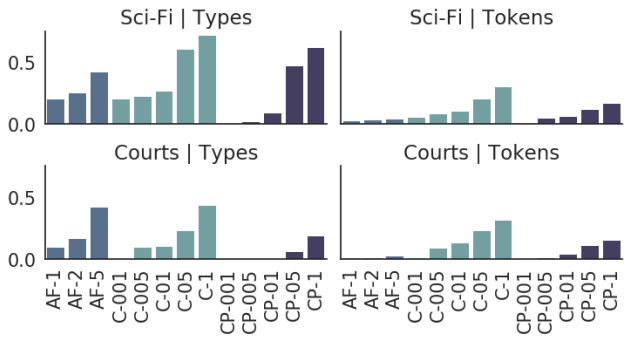

Figure 6.5: Proportional loss of removed word types and tokens. Contextual probabilistic subsampling does substantially less damage than contextual curation.

CORPUS DAMAGE IS REDUCED. All things being equal, we want to modify the input collection minimally, both in terms of vocabulary and actual document content. Figure 6.5 confirms that contextual curation has the highest type and token loss across corpora, because it completely removes all instances of a word type in a context. This may partially explain the dramatic loss of model quality for these specific treatments.

Contextual probabilistic subsampling removes more tokens than author frequency cut-offs, but better preserves the vocabulary. For thresholds $t \leq 0.01$, contextual probabilistic subsampling removes fewer word types than any of the author frequency cut-off methods. However, there is less agreement across corpora for $t \geq 0.05$. For SCI-FI, these methods remove more types than AF-5, while the reverse is true for COURTS. This discrepancy might arise from differences in the relative size of collection sources—some authors write more than others, some courts issue more opinions—and vocabulary use.

SUBSAMPLING AFFECTS MORE THAN NAMES. Character names are the most prominent motivating example for this work, so it is reasonable to ask whether named-entity tagging or even a simple part-of-speech (POS) filter would be suf-

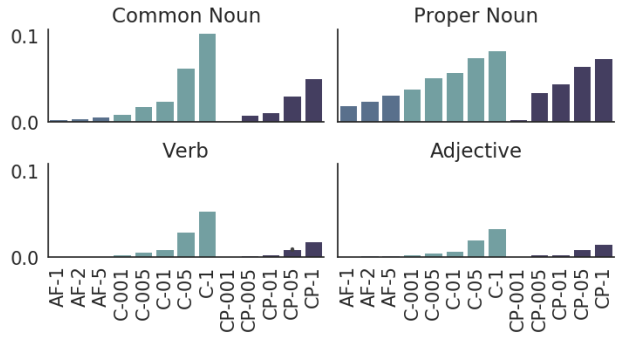

Figure 6.6: Proportion of Sci-Fi tokens removed across part-of-speech groups. Contextual methods remove tokens from all groups.

ficient. To check whether we are just removing proper nouns, we compare the frequency of four general POS categories: common nouns, proper nouns, verbs, and adjectives. These make up 37%, 10%, 27%, 13% of all tokens respectively in Sci-Fi. Figure 6.6 shows the proportion of tokens removed from each category for each curation method. Unsurprisingly, proper nouns make up a large proportion in all cases, but contextual methods also remove substantial numbers of tokens across all word groups.

SUBSAMPLING INCREASES STABILITY AND SPECIFICITY. We find that removing author-specific terms using contextual probabilistic subsampling greatly mitigates the formation of author-correlated topics, but what do these models learn instead? Are they augmenting the set of uncorrelated topics found within the untreated models, or are they perhaps identifying entirely new structure? More importantly, what are the characteristics of the newly formed or persisting author-correlated topics? To answer these questions, we perform pairwise comparisons of topic-word distributions from different models using Jensen-Shannon divergence to find the most likely of topic correspondences. By linking these topics together, we can gain a sense of which topics persist across treatments, which are refined or

split, and which are lost entirely. We focus on SCI-FI since it has larger models, but we will highlight similar analysis for COURTS.

Before making any inter-treatment comparisons, we examine topic stability internally within each treatment. We define stability as the average similarity between a topic and its nearest equivalent from each of the nine other trained models for a treatment. More formally, the stability of topic k_i from the ith instance of a model is

$$\text{Stability}(k_i) = 1 - \frac{1}{9}\sum_{j \neq i} \min_{k_j} JSD(P(w \mid k_i), P(w \mid k_j)) \tag{6.7}$$

where JSD is Jensen-Shannon divergence. A topic stability close to one implies that a topic persists across runs, while a value close to zero implies that a topic is *ephemeral*—observed once and unlikely to be seen again across random initializations.

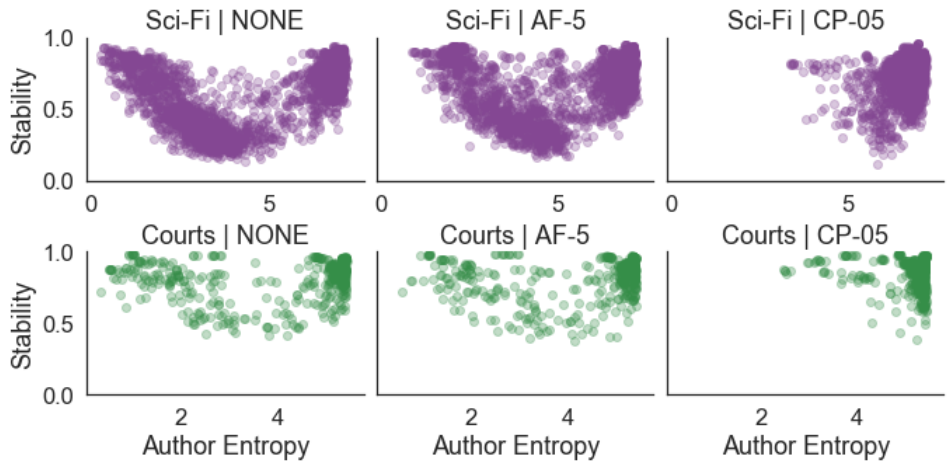

Figure 6.7: Topic Stability and Entropy for SCI-FI ($K = 250$) and COURTS ($K = 50$). AF-5 has little effect. Many of the low-entropy topics avoided by CP-05 are highly unstable.

High stability does not imply author-specificity. In Figure 6.7, we see that the most stable topics tend to have either maximal or minimal author entropy, while

the most unstable topics have middling values. The unstable topics tend to capture a mixture of disjoint structures as we saw in topics C and D from Table 6.2. This also occurs (but to a lesser extent) in COURTS with topics containing many distinct regional terms (S1: *s.w oklahoma tenn kan ind n.e indiana app tennessee o.s*) or containing a mixture of a general and state-specific concept (S2: *school wyo miss wyoming mississippi ann education students hill student*). Thus, the most stable topics are the most apparent by being very context specific or most cross-cutting.

Now that we have evaluated the stability of topics under the baseline NONE treatment, we can use minimum divergence to align those topics with topics trained under the CP-05 subsampling treatment. Unstable NONE topics are generally very distant from their nearest CP-05 counterparts. Of our example topics in Table 6.2, C and D are the most unstable at 0.39 and 0.42 respectively. Topic C diverges heavily (0.87) from its closest CP-05 match, while aspects of D are echoed in its nearest match *sand desert rock mountains mountain dust land surface plain water* (0.53). COURTS topic S2 is also more distantly associated (0.63) with an education/administration topic: *board school commission administrative agency plan department board's education regulations*. Over 95% of NONE topics with high stability and high author entropy are linked to a CP-05 topic with divergence less than 0.5. Topic A has a close match (*professor university college student students research school science work years*) at 0.23. A appears to have become more specific in CP-05 by splitting into two additional topics that echo other aspects, namely teaching children (0.54) and scientific research (0.55).

The case of stable, low entropy NONE topics is harder to interpret. While half of these topics are far from their CP-05 match, 16% have divergences of less than 0.4. Topic G matches well to *lord hold between master queen star enough turns high good* (0.3) which is both very stable and CP-05's lowest author entropy topic (64.1%

from Anne McCaffrey).[6] While these topics have not been prevented entirely, they have been largely mitigated.

The topics in CP-05 that are the most dissimilar from topics within NONE demonstrate that this treatment adds differentiation. We find that overall 50% of CP-05 topics have a large divergence with the NONE topics. Some of these divergent topics consist of names, but these groupings might indicate regional or temporal naming patterns. In other cases, we encounter new and interesting topics such as an authentic robots topic (*machine robot machines robots human mechanical metal brain men built*), which matches to both a general computer topic and example topic E (Asimov). We also find a new topic on magic and witchcraft (*magic ghost demon evil witch demons power spell magician ghosts*) whose closest match is a general religion topic *god gods religion world religious ancient temple people faith these*. In fact, the term *witch* never appears as a top-20 term for any topic within the 250-topic NONE models. These topics may appear for NONE when we increase the topic size to 1000 topics, but at the cost of a much larger model and with no guarantee against intruding character names.

SUBSAMPLING PRODUCES CROSS-CUTTING TOPICS. While our topics score well quantitatively, how humanly interpretable and useful are the resulting topics? Are they actually cross-cutting in nature? We address these questions by more closely examining topics generated by the CP-05 subsampling treatment. We can explore the collection by sorting authors and individual novels within topics.

The highest frequency topics from the *NONE* treatment are largely preserved by *CP-05*. These topics by their nature are very cross-cutting and filled with frequent, general words. Despite this extreme generality they can provide a way to analyze

[6] The topic is composed of common words used in specific ways: a *hold* is a fortified settlement, dragons teleport by going *between*.

passages representing high-level discourse concepts such as inquiry (*why asked ask answer question want questions should does because*) and the description of events and time (*during such most these course because happened effect period result*).

The mid-frequency topics are more concretely thematic in nature. We find a topic describing empire, politics, and history (*empire world power people war new government history political under*) which is associated with Doris Lessing's *Canopus in Argos* series, Isaac Asimov's *Foundation* series, and Kim Stanley Robinsons's *The Years of Rice and Salt*. In line with the science fiction genre, these novels focus on expansive future and alternative histories. We also find a topic on language (*language words english speak word understand spoke speech languages talk*). The most prominent authors in the topic—Robert A. Heinlein, Robert Silverberg, and Poul Anderson—are among the five most prolific authors in Sci-Fi, which suggests the generality of the topic. Notably the most prominent volumes are by none of these authors: *Babel-17* by Samuel R. Delany, *Native Tongue* by Suzette Haden Elgin, and *Changing Planes* by Ursula K. Le Guin. All three include the social and political language as a major plot point. These three works are fundamentally tied confirming that this topic embodies a cross-cutting linguistic theme.

Looking more closely at the lower frequency robots topic (*machine robot machines robots human mechanical metal brain men built*), we find that it is both topically cohesive and cross-cutting. The five most-represented authors all have works heavily related to artificial intelligence: Isaac Asimov, Robert Silverberg, Stanisław Lem, Clifford D. Simak, and Philip K. Dick. The most-represented volumes tell a similar story with *Men and machines* by Robert Silverberg, *The complete robot* by Isaac Asimov, and *The Humanoids* by Jack Williamson holding the top three ranks. Reassuringly, there are well-represented novels by less-represented authors such as *The Starchild Trilogy* by Fredrick Pohl and Jack Williamson. The low frequency of this topic is surprising given the presence in the collection of robot-related novels,

especially works by Isaac Asimov. This discrepancy revealed that an Asimov-specific topic (*human being law might must such without may robot beings*) has persisted. Many authors receive a non-negligible token representation, but Asimov's token count is still a factor of ten larger than the second most prominent author (Robert A. Heinlein).

6.7 CONCLUSION

We present a formal definition of the problem of overly source-specific topics, three evaluation metrics to measure the degree of source-specificity, and a simple text curation meta-algorithm that dramatically reduces the number of source-specific topics. This approach has immediate practical application for the many collections that combine multiple distinct sources, but it also has important theoretical implications.

We view this work as a preliminary step towards predictive theories of latent semantics, beyond purely descriptive models. Despite ample practical evidence that interventions such as stoplist curation can have significant effects, most previous work has focused on algorithms for identifying a single "optimal" low-dimensional semantic representation. Our results indicate that there are potentially many interventions in text collections that each have distinct but predictable effects on the results of algorithms. Just as biologists use multiple stains to view different aspects of microorganisms using the same microscope, users of text mining algorithms should be able to choose multiple distinct text treatments, each with its own predictable effects, to meet distinct user needs.

7

SETTING THE STAGE FOR MAGICAL GEMS: CONSTRUCTING USEFUL COMPUTATIONAL CUT-UPS.

7.1 INTRODUCTION

With recent advancements in computer vision, massive collections of visual material are not only accessible but also computationally analyzable. But how well do these methods extend to studying material artifacts that are inherently three dimensional in nature? In Chapter 5, I introduced a framework for studying avant-garde periodicals using computational cut-ups—extracted features from pretrained convolutional neural networks (CNNs)—but these materials are inherently two-dimensional. How much might perspective and framing matter given that the underlying models are trained on ImageNet [Deng et al., 2009], a large collection of images gathered from the internet? ImageNet model representations are known to be useful for a wide range of tasks [Razavian et al., 2014], but not all [Kornblith et al., 2019]. In this chapter I will focus on building useful image representations for studying "magical gems"—a modern category of engraved gemstones from the Roman Imperial period (chiefly 1st c. BCE through 4th c. CE).

"Magical gems" are small, physical objects that are primarily museal in nature. Very few of these objects have known ancient contexts, but tend to have long post-antique ones. Since the Renaissance they have been collected by European elites who acquired them from art markets and surface finds, but they have also survived through reuse such as their incorporation into medieval relics and

post-antique jewelry. Because very few (less than 1%) of these objects come from documented archaeological contexts, they have primarily been studied through their iconography and inscriptions. Their categorization has derived from their deviation from typical Graeco-Roman images and texts rather than a known, unified ancient use. In fact, from the small set of known contexts, these objects seem more varied than homogeneous; these engraved gems have been found in sites from across the Roman empire—from Thetford (England) to Rome to Pergamon (Turkey)—and their contexts suggest differing uses some mundane, others ritualistic [Barrett, forthcoming].

Because of their long presence in museums and private collections, magical gems have been copied and disseminated through many mediums including drawings and photographs, as well as plasters casts and impressions. Unsurprisingly, the majority of these representations prioritize preserving iconography and inscriptions over materiality. Black and white photographs provide higher contrast to better highlight fine etchings, but no longer preserve gemstone color; similarly, impressions can improve the legibility of inscriptions but erase the gem's material form (e.g. color, banding, translucency) entirely. While any digital representation of any these mediums has flaws, collectively they provide a more complete rendering of the original, physical object. By building a vector space that productively links similar objects across mediums, we might reexamine the category of magical gems through an alien lens, one that has no awareness of what imagery and text is "standard" versus "strange." Such a method would not only be useful for studying this niche group of objects, but also for studying the broader range of artifacts in archives and museums. Moreover, these methods provide the potential to explore relations of objects across medium types (e.g. paintings, sculptures, jewelry).

Unfortunately, and perhaps unsurprisingly, extracted neural features are far from a panacea. At the best of times, it is difficult to interpret what exactly

these representations are capturing. For example, if we revisit the "Not-Dada" classification results seen in Figures 5.12 and 5.13, a large number of these pages are misaligned such that page edges are visible. Without further analysis it is difficult to tell whether this is problematic or coincidental. If page-level alignment is being captured and utilized by our classifiers, then we have a problem reminiscent to that described in Chapter 6. Page misalignment can indicate concepts we would like to ignore about the physical volumes and periodicals. They might indicate the relative size of the volumes via the difficulty of scanning fully flattened pages. Even worse, they might simply indicate who scanned which volumes. So, we need a method that can eliminate or at the very least dampen the encoding of specific information within our computational cut-ups.

Given the nature of our computational cut-ups, the methods of Chapter 6 are not immediately applicable. The dimensions of these vectors are neither directly interpretable nor generated from content that can be easily subdivided into meaningful components. So, instead of focusing on modifying the original images, I will focus on interventions applied to the extracted feature space directly. In other words, we want to remove or at the very least heavily dampen the presence of some known learned structure from an existing embedding so that our subsequent uses of this space (e.g. classification, clustering) do not depend on this unwanted structure. This problem is closely related to the notions of model bias and fairness. Through the course of this chapter I will apply methods from the model (de)bias and fairness literature in order to produce useful computational cut-ups for studying magical gems.

7.2 BACKGROUND

In this section I will provide a brief overview of the varying definitions of magical gems as a category and summarize past statistical analyses of said category.

7.2.1 *What is magic?*

Before going further, it is worth unpacking the meaning of the term "magical."[1] Magic has long been defined with respect to and in opposition of religion. Within this framing, magic is unlike religion because it focuses on the individual, positions practitioners as having power over the supernatural, and is generally anti-social with selfish goals in conflict with those of society. However, these conceptions of magic and religion reflect modern, etic, viewpoints rather than ancient, emic, ones.[2]. In early definitions, magic is considered a "bastard sister of science" [Frazer, 1911–1915] that falsely claims that through proper knowledge and precise action individuals can achieve power over their environment. The perspective that magic is false derives from a modern and particularly Protestant perspective. In opposition to this narrative, Tambiah [1990] influentially argued that magical acts, and ritual acts more broadly, are performative and effective but are dependent on their audience and the social framework from which this audience operates. Now, scholars tend to view magic and religion as etic concepts that while modern in conception provide a useful framework to operate within. After all, as H. S. Versnel [1991b] aptly said "you cannot talk about magic without using the term magic."

That being said, the ancient Greek and Roman concepts of *mageia* and *magia* (respectively) are not entirely incomparable with modern conceptions of magic. These

[1] See Collins [2008, Ch. 1] for an overview of the history of magic within anthropology.
[2] On definitions of magic see Versnel [1991b]; on definitions of religion see Smith [1998]

emic terms covered a wide range of practices including love charms, binding spells, controlling the weather, and raising the dead. In both cases, they were typically marked as abnormal, problematic, and even dangerous. In Rome, many harmful magical practices were illegal. Nonetheless, many of these practices especially binding curses (in the form of curse tablets) were very common throughout the Roman empire. Still, healing and protective practices were not generally conceptualized in this way since they were not meant to cause harm but rather prevent or remedy it. Generally, these "positive" magical acts were not banned or criminalized and were often used by the very same people who might condemn other forms of magic. Moreover, given that emic definitions of magic focused on stigma and illegitimacy, the term was most often applied to criticize and delegitimize others; one's own practices would rarely be considered magical.

In any case, the working etic definitions of magic consistently include two categories of object: the Graeco-Egyptian magical papyri (both *Papyri Graecae Magicae* and *Papyri Demoticae Magicae*; see Betz, 1992) and curse tablets (*katedesmoi*, *defixiones*; see Gager, 1992). The former is a modern collection of extant papyri from Graeco-Roman Egypt that contain spells, rituals, and hymns. The latter are thin, inscribed tablets, predominantly made of lead, that were typically folded, often pierced with nails, and usually deposited under ground (or water). Unlike the magical papyri, curse tablets have been recovered from depositions throughout the Graeco-Roman world. Despite these geographical differences, the texts of these "magical" artifacts have much in common. They both use similar formulaic language, magical names (*voces magicae* and *logoi*), and magical signs (*characteres*). Additionally, these object groups both invoke a level of secrecy; curse tablets are literally buried and a number of magical papyri explicitly instruct keeping the described ritual or text itself secret.

While the magical function of both the Graeco-Egyptian magical papyri and curse tablets may seem readily apparent, both categorizations have their limitations. In the case of the magical papyri it is worth considering *who* wrote these texts. Since these papyri are written in a mixture of Greek, Demotic, and Coptic, they require authors with literacy in both Greek, the common language of Graeco-Roman Egypt, and the Egyptian scripts of Demotic and Coptic. This has led to the growing argument that these texts were written by Egyptian priests who often served as local ritual specialists [Frankfurter, 1997]. This emphasizes the potential disconnect between emic and etic conceptions of magic, since in the case of the magical papyri there might be little to no difference between "magical" and "religious" function.

Curse tablets are much more numerous, widespread geographically, and varied in form than the magical papyri. Perhaps unsurprisingly, this diversity results in subcategories that vary across the "magic"-"religion" spectrum. Some subgenres such as those to bind opponents in theatrical and athletic competitions (see Gager, 1992, p. 42–77) or legal opposition (see Gager, 1992, p. 116–150) easily fall under etic definitions of magic; their purposes are individualistic and harmful. However, there are also groups whose connections with magic are much more tenuous. For example, curse tablets that invoke pleas of justice and revenge (see Versnel, 1991a and Gager, 1992, p. 175–199) more closely resemble prayers than curses. In fact, Versnel [1991a] refers to this group of tablets as "judicial prayers" and rejects their categorization as *defixiones* entirely. Their inscriptions depict supplications to deities to redress some wrong on the behalf of the invoker. These pleas do not fit the typical magical narrative because the power dynamics between the practitioner and the supernatural is inverted. These pleas typically explain *why* the targets need to be punished and might provide a transactional incentive. For example, in the case of stolen property, the stolen goods are customarily dedicated to the god being invoked. Suffice it to say that magic as a category is both modern and subjective.

While it is useful as a broad definition, it does not establish ancient conception or function and its labelled instances should be (re)examined.

7.2.2 *What are magical gems?*

An engraved gemstone is considered a "magical gem" based on its iconography and inscriptions. A gem is "magical" if its engravings contain "magical names (*voces magicae* and *logoi*), magical signs (*characteres*), and non-standard iconographic types (e.g. Chnoubis or the Anguipede scheme)" [Nagy, 2011, p. 88]. While these artifacts only need to possess one of the above features to qualify for category membership, they also tend to have the following two physical features: "(i) they are generally engraved on both sides, rather than one; (ii) they are usually inscribed with Greek text that is not in retrograde (that is, the gemstone was not used as a signet ring to seal documents)" [Faraone, 2018, p. 16].

This art historic categorization of these artifacts stems from early modern collecting practices. Since the beginning of their modern collecting histories, gemstones have been organized by their iconography. By the 17th century, the defining characteristics of magical gems developed because of their clear divergence from the classical Greek canon. This alien quality was attributed initially to the heretical, "un-classical" Gnostic tradition [Nagy, 2012]; at this time, they were known as "Gnostic gems." This distinction from classical Greek iconography led scholars to spurn these gemstones as ugly and lesser quality compared with their non-Gnostic, non-magical counterparts. For these scholars, magical gems depicted the sharp cultural decline occurring in the Graeco-Roman world which therefore made them unworthy of study [Gordon, 2011]. Johann Joachim Winckelmann [1764, p. 59–60] explicitly dismissed magical gems as unworthy of inclusion in the study of ancient

(i.e. classical Greek) art: "sind nicht würdig, in Absicht der Kunst, in Betrachtung gezogen zu werden."

The modern aesthetic of Winkelmann's time which preferred classical statuary to be a bare white rather than painted [Potts, 2000], also preferred to separate these lesser gemstones from their classical counterparts. Adolf Furtwängler [1900] excluded magical gems from his study of ancient gems and even had the majority of these objects removed from the Berlin Museum's Antiquarium to the Ägyptisches Museum [Gordon, 2011]. Similarly, the magical gems of the British Museum were moved to medieval antiquities departments [Gordon, 2008]. In fact, until the mid-1970s, the British Museum's collection of magical gems was spread over more than four different departments [Gordon, 2002].

The current definition of magical gems as "engraved stone[s] used in a magical manner" [Nagy, 2012, p. 89] formed in the 20th century. In 1914, Armand Delatte [1914, p. 21–22] rejected magical gems' association with Gnosticism for Graeco-Egyptian magic:

> En réalité, ce monuments n'ont aucun rapport spécial avec le Gnosticisme: ce sont simplement des amulettes qu'on doit attribuer a l'époque d'efflorescence des doctrines et des pratiques de la magie gréco-égyptienne (du 1er au IV s. ap. J. C.).
>
> In reality, these monuments have no special relationship with Gnosticism: they are simply amulets that must be attributed to the blooming period of doctrines and practices of Graeco-Egyptian magic (1st to 4th c. CE).

Delatte's connection to Graeco-Egyptian magic derived from the icongraphic and textual similarities shared by these gemstones and the Graeco-Egyptian magical papyri as well as surviving curse tablets. This position was developed further in

Campbell Bonner [1950]'s catalogue of magical amulets, chiefly Graeco-Egyptian, and Delatte and Derchain [1964]'s catalogue of magical Graeco-Eygptian gems. In these works, visual representations of gems are photographs of casts which erased many aspects of their materiality but made their engravings more visible. While the category of magical gems underwent changes in name and interpretation, its defining characteristics did not fundamentally change. However, this newly formed connection with magical papyri and curse tablets did shape scholarly opinions of magical gems. Many have assumed that these gemstones must have been predominantly produced in Alexandria [Gordon, 2011] and that as magical objects they must have been owned by the poor and uneducated [Nagy, 2014].

However, recent work on magical gems has brought into question the "magical" function of these objects. Unlike the magical papyri, these engraved gemstones are not primarily textual. Likewise, unlike curse tablets, magical gems seldom have known proveniences—excavational findspots—that could point to their "magical" function.[3] Moreover, as material objects belonging to the broader category of engraved gemstones, it is worth considering how much their textual/iconographic features—versus other physical/material ones—reflect form rather than function. A number of scholars, such as Árpád M. Nagy and Christopher Faraone, stress the modern bias inherent in the term of "magical" by considering it interchangeable with the adjectives "ugly" [Nagy, 2011, p. 75], "nonsense," and "weird" [Faraone, 2018, p. 5]. Clearly, this viewpoint is bolstered by strangeness being a key reason for the gemstones' initial separation from their classical brethren, but how much do they actually have in common with other objects within the Graeco-Roman "magical" genre?

[3]See Barrett [forthcoming] for a study of the few magical gems with known archaeological contexts.

Of the three iconographic features of magical gems, only the first two fully align with the magical papyri and curse tablets. While magical names and signs are prevalent among both magical gems and other magical media, the "non-standard iconographic types" are not nearly as consistent. The magical papyri contain fifteen recipes involving gems and rings all of which can be matched with at least one surviving gem [Nagy, 2015, p. 218].[4] However, the quintessential images of Chnoubis and the Anguipede (see Figure 7.1) are almost entirely absent (see Nagy, 2015 and Vitellozzi, 2018). Moreover, texts on magical and medicinal amulets suggest that even normal, "canonical" imagery can have protective properties. For example, engraved representations of Poseidon might prevent ship wrecks [Bonner, 1950, p. 14] or cure eye diseases [Nagy, 2012, p. 85], while representations of the twins Apollo and Artemis could help ensure a safe childbirth [Nagy, 2012, p. 85]. Furthermore, these texts also emphasize the importance of a gem's material for its function and power (see Mastrocinque. Some even suggest that the material rather than the iconography gives an amulet its magical or healing power. For example, Galen (*De simpl.* 10.19) argues that the healing properties of green jasper amulets is due to their material rather than their engravings of a radiate serpent (i.e. Chnoubis). All of this supports the view that "magical" for magical gems has more to do with being "ugly" and "weird" than etic magical functions.

While the figure of Chnoubis—a radiated or haloed lion-headed snake—is seldom seen outside of magical gems, it is attested in ancient medical texts as an image of healing, especially for the stomach [Dasen and Nagy, 2012]. For example, the following passage from the lapidary of Socrates and Dionysius could easily describe the scheme we see in Figures 7.1a and 7.1b:

> Engrave on it, then, a serpent coil with the upper part of a lion and rays.
> If worn this stone completely prevents pain in the stomach; rather you

[4] See Vitellozzi [2018] for a recent comparison of magical texts and magical gems.

(a) Chnoubis (b) Chnoubis (c) Anguipede (d) Anguipede

Figure 7.1: Example Chnoubis (CBd-2350) and Anguipede (CBd-1367) iconographies. Each gemstone face is presented in two forms—an image of the gem and an alternative representation (impression and cast)–to more clearly present the iconographies. Sources: (a), (b) courtesy of the Getty's Open Content Program; (c) © American Numismatic Society; (d) from Bonner [1950, Pl. VIII].

> will easily digest however many foods you make use of. Let the wearer not set this aside. [trans. Faraone, 2011, p. 51]

In several cases, this healing function is further supported by the gems themselves through the additional inscriptions they bear. These inscriptions can be both detailed (e.g. "Keep Proclos's stomach healthy" CBd-2943) and simple (e.g. "Digest, digest!" CBd-1041).[5]

(a) CBd-210 (b) CBd-2366

Figure 7.2: Example representations of Chnoubis signs both with the figure Chnoubis (a) and without (b). Sources: (a) Gem Impressions Collection, Cornell University Library, (b) courtesy of the Getty's Open Content Program.

[5]See Dasen and Nagy [2012] for the full range of inscriptions (Appendix 8), as well as a discussion of the broader range of textual evidence on the Chnoubis of magical gems.

The figure of Chnoubis is typically accompanied by an inscription of the name *Chnoubis* as well as a specific magical sign—a series of S's, usually three, crossed through by a horizontal line (e.g. Figure 7.2)—known as the "Chnoubis sign." The name always co-occurs with the figure or sign of Chnoubis, while the sign can occur without either. In fact, there is textual evidence that the sign had its own healing value (see Dasen and Nagy, 2012). Unlike the visual motif, both the name and sign of Chnoubis can be loosely connected with the magical papyri. While the name is not directly seen in other magical mediums, possible spelling variations such as *Chnouph* and *Chnoub* are present in the magical papyri (see Shandruk, 2016 esp. Table 8). Similarly, the Chnoubis sign has been connected with a recipe from the magical papyri (*PGM* IV. 12267-1264; see Betz, 1992 and Dasen and Nagy, 2012) that mentions a phylactery (i.e. amulet) engraved with a similar magical symbol as part of a ritual to drive out daimons (i.e. spirits). The symbol is a single crossed S, which while visually similar, does not itself appear on magical gems (see Shandruk, 2016, §2.4).

If we accept the notion that the figure of Chnoubis has a healing function, this does not inherently imply a magical one. Recalling the definitions of magic discussed in §7.2.2, a healing function is far from a dangerous one and not inherently in opposition to society or religion. We can only reach a "magical" position from a modern perspective wherein we connect the use of Chnoubis gems with other "superstitious" rituals such as charms or amulets that are often operating within religious contexts (e.g. representations of saints).

Unlike the image of Chnoubis, it is more difficult to ascribe an ancient function to the Anguipede figure. The rooster-headed, snake-legged figure with an armored human torso typically holding a shield in its left hand and a whip in its right is not described in ancient texts. The typical motif does not appear in the surviving magical papyri and curse tablets, although there is a possible variant figure in

each.[6] Despite this lack of antique history, the figure has an established post-antique one. The motif is referenced in medieval and Renaissance lapidaries—treatises on the properties of stones—where it could be used both as an amulet for warriors and protection against poisons and hemorrhages (see Nagy, 2014). Additionally, (antique) Anguipede gems underwent less magical post-antique reuse as jewelry, seals, and even as part of a medieval reliquary. There is also some evidence of more "normal" (i.e. non-magical) ancient usage of Anguipede gems as seals; of the tens of thousands of preserved sealings from a civic archives building in Zeugma (Turkey), two depict the Anguipede motif (CBd-1753, 1754; see Barrett, forthcoming).

Given the lack of supporting textual sources and appearances in other magical mediums, scholars have focused instead on the inscriptions that most often accompany the Anguipede figure. It is heavily associated with two magical names: Iao, which is often inscribed within the Anguipede's shield, and Abraxas (sometimes Abrasax). These religious names, Jewish and Gnostic respectively, are frequently invoked in the magical papyri and curse tablets. The association between these names and Anguipede gems is so pronounced that the Anguipede figure has often been referred to as Abrasax or Iao Abrasax [Bonner, 1950, p. 123]. The problem with this association is that it is not exclusive. Walter Shandruk [2016] points out that *Abraxas* is more correlated with *Iao* than the Anguipede motif itself. However, this could be accounted for by Nagy [2019]'s theory that the Anguipede represents an iconographic representation of the name of the God of Israel. That being said, this does not explain why such a popular figure in magical gems is not seen elsewhere. It is clearly a strange figure, not belonging to the classical Greek canon, but it is far from evident that it served a wholly "magical" function.

[6] A human-headed anguipede occurs on a lead scroll from Corinth [Wiseman, 2016]. A rooster headed, human bodied figure without the eponymous snake legs is depicted within a love spell in the magical papyri (*PGM* XXVI. 69-101).

Even if the "non-standard iconographic types" belong to the Graeco-Roman magical repertoire, it is not clear that the primary ancient uses of magical gems align with "magical" practices. Magical gems, unlike curse tablets and the magical papyri, have evidence of more mundane and visible uses as seals and jewelry. While there is limited evidence of magical gems used as seals, at least in some cases the non-canonical imagery of these objects could be used as an individual's signature (e.g. CBd-1753, 1754; see Barrett, forthcoming). In much larger supply, are magical gems with evidence of ancient settings in rings and pendants. Some surviving gems have remained in their ancient metal mounts (e.g. CBd-14, 617, 1021), while many others bear shell-like chip marks which may indicate that these gemstones were forcibly removed from their original ancient mountings (e.g. CBd-387, 1132, 1554) [Nagy, 2015]. These magical gems, like antique gems more broadly, served a social purpose: they were worn as jewelry and held the capacity to serve as markers of wealth and status (see Barrett, forthcoming). The visibility of these gems as seals and jewelry conflicts with the general view of magic as private and secretive. While it is possible that some gems may have been used in a more secretive way, such as by having the stranger, magical faces hidden from view, the material evidence cannot uphold this notion for the entire category.

It is worth noting that the use of magical gems as jewelry does not remove the possibility that these stones granted power to their wearer, just that their power is not necessarily derived from the same "magical" sources as the magical papyri and curse tablets. However, we must consider, as with any visible imagery, the possibility that aesthetics might play a larger role than magical or religious beliefs. Some wearers might select these iconographies and materials for their "magical" properties, but others might choose them for their beauty or social implications. At the very least, we must consider the value of medium from which many of these stones were made and what the social statements that these materials alone provide.

After all, the material values of these gemstones have played no small role in their continual circulation, both in terms of post-antique collecting and post-antique reuse (see Nagy, 2014). Moreover, these objects were circulated throughout the Graeco-Roman world and while they might bear a ritual or magical function in one context, they might serve an entirely different purpose for people in another (see Barrett, forthcoming).

In summary, magical gems have been categorized because of their weirdness, not their etic or emic magicalness. This becomes even more apparent if we consider Nagy [2012, p. 90]'s diagram for categorizing talismans, jewelry used in magical manners, in Figure 7.3. Magical gems only cover gemstones deviating from classical imagery. This should come as no surprise given that the 17th century category has hardly changed in definition, only in interpretation. So, we should treat "magical gems" as little more than a name and keep in mind its functional meaning is about as helpful as the categorization between bugs and insects: "many bugs are insects, while some insects are bugs" [Nagy, 2012, p. 89].

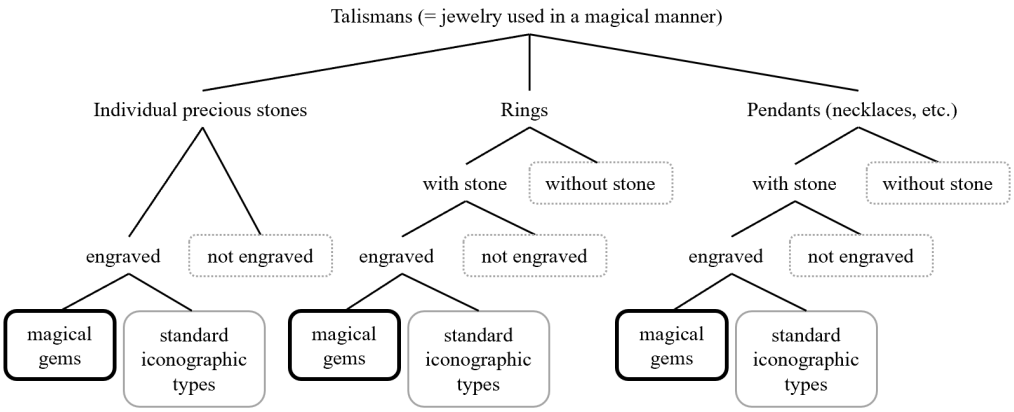

Figure 7.3: A diagram for identifying magical gems from other forms of talismans [adapted from Nagy, 2012, p. 90]

7.2.3 *Past and future analyses of magical gems*

In recent years, there have been extensive efforts to fully catalog the set of extant magical gems residing in public and private collections (e.g. Mastrocinque, 2003–7 and Michel, 2004). With thousands of gemstones categorized as magical, the use of statistical analysis has become more inviting [Nagy, 2011]. These studies have aimed to quantitatively characterize various properties of magical gems: color and material [Mastrocinque, 2011], decorations and magical signs [Dzwiza, 2019], Anguipede iconography [Nagy, 2019], and the general cooccurrence of attributes [Shandruk, 2016]. While these studies all critique some aspect of magical gems as a category, they all rely on preexisting categorizations and limit their studies to magical gems themselves.

To my knowledge, Shandruk's study of magical gems is the first to use a computational approach. He uses network analysis to better understand how different gem attributes—namely, "material, color, iconography, and inscription" [Shandruk, 2016, p. 3]—relate to one another. Surprisingly, this network does not directly contain magical gems. Instead, its nodes represent gem attributes that are linked by their degree of cooccurrence across magical gems. Using cluster analysis, Shandruk identifies well-connected groups of features. These feature sets provide a new, empirically driven organization of magical gems that can be further subdivided by analyzing the gems belonging to a specific attribute cluster. Unlike previous taxonomies of magical gems (i.e. Bonner, 1950 and Michel, 2004), Shandruk's attribute clusters are inherently more flexible allowing gems to belong to whichever clusters their attributes dictate.

Computational cut-ups provide a new way to explore and question the boundaries of magical gems as a category removed from the wider canon of engraved

gemstones. With an appropriate embedding space, we can observe how gems—both magical and canonical—tend to cluster within this space. Do magical gems actually form their own distinct subgroups? Or is there a more complex relationship, one that might depend on how these objects were collocated and catalogued? Ideally, such an analysis would use multiple sets of computational cut-ups that each emphasize and de-emphasize particular aspects such as color, iconography, and size. These embeddings act as different perspectives with different prioritizations of gem features; in some, the category of magical gems might be wholly recognizable, while in others magical and non-magical gems are indistinguishable.

However, this proposal is not feasible without useful latent vector spaces. Off-the-shelf extracted features will encode structures that we will often want to remove such as representation medium (e.g. drawing, cast, photograph), levels of preservation, and cataloging artifacts (e.g. catalogue numbers). This chapter investigates how the computational cut-ups of *magical gem* representations can be transformed into more useful forms. I leave the larger study of Graeco-Roman engraved gemstones—both canonical and magical—to future work.

7.3 DATA

I use the Campbell Bonner Magical Gems Database (CBd) to build a working collection of gem-level images and metadata.[7] CBd is an online database that intends to make the entire collection of magical gems publicly available. It currently contains over 3,200 entries of predominantly magical gems. Each entry is represented by a series of structured text fields, a free-form textual description, and typically one or more images of the object. The images represent a variety of mediums including

[7] http://cbd.mfab.hu/

photographs—both color and monochrome—of the object, its cast, or its impression, as well as drawings of the object.

Although CBd primarily contains magical gems, it contains a number of objects that do not necessarily fit within this category. The database also covers gems with similar, but not explicitly magical, iconography (e.g. certain Egyptian motifs); a number of amulet gems included for their appearance in particular scholarly texts or their restricted access; and a few votive gems that bear inscriptions describing their function (e.g. CBd-8 [2017]: "Ophelimus offered it, after a dream that a divinity had shown him"). But CBd also contains post-antique gems (e.g. CBd-3, 365) as well as objects, both antique and post-antique, that are not gems.[8] In general, dating engraved gemstones is difficult and relies primarily on style. Moreover, newer gems can copy or at least echo ancient ones even as they take on new meanings. This is clearly the case for modern casts and impressions, but should also be considered for antique gems as well; the present can always reflect the past.[9] In any case, this additional heterogeneity does not pose an issue for my task of building more useful computational cut-ups. Useful vector representations should be able to incorporate these objects since they are related to the category of magical gems if not inherently part of said category. However, their differences should be taken into consideration for any analysis that engages with these embedding spaces. What objects should be included or excluded will depend greatly on the particular question being operationalized. For simplicity, I will refer to the artifacts within my working collection as "gems" going forward.

[8]CBd contains a small number of related non-gem objects such as an antique sheet of papyri describing an amulet for fever that might have been worn as an amulet (CBd-9) and a post-antique bronze statuette of an Anguipede (CBd-1022).

[9]See [Nagy, 2019] §4.3, 5.

7.3.1 *Images.*

By including all CBd entries with images, the working collection contains a 3,202 gems. For these objects, CBd has a total of 10,629 images, but some of them contain multiple views (e.g. images of the obverse and reverse) of a gem which may also be included separately. Similarly, other images include multiple physical representations (e.g. gem and its impression) in the same photograph. Without intervention, the resulting embedding space will be dominated by whether or not an image contains multiple views or representations. So, I split these images such that the final ones contain only a single view or representation of a gem; any resulting duplicates are discarded. This splitting process is semi-automated and operates at the highest available image resolution. It results in 12,014 distinct images. The median number of images per gemstone is two. At the extremes, 607 gems are represented by a single image while 61 are represented by more than ten.

While I did regularize the set of working images to contain a single view of a gem, there is unsurprisingly more variability which I am not accounting for. In particular, I do not control for the background of the image or maximally crop each image. Image backgrounds vary widely: sometimes black or white, other times multi-toned or textured. Additionally, images may only represent a detail (e.g. close-up of a iconographic element or inscription) rather than a full view of a gem. Moreover, these images come in a variety of dimensions which will result in different degrees of deformations when preparing them as input for deep learning image models.

Label	Criteria
Magical Names	Are there any magical names inscribed on this gem?
Characteres	Are there any *characteres* inscribed on this gem?
Anguipede (rough)	Is the Anguipede figure depicted on this gem?
Anguipede (clean)	Is the Anguipede figure depicted within this image?
Chnoubis (rough)	Is Chnoubis depicted on this gem?
Chnoubis (clean)	Is Chnoubis depicted within this image?
Drawing	Is this image a drawing of a gem?
Photograph	Is this image a photograph of a gem?
Simulacrum	Is this image a photograph of a cast or impression of a gem?

Table 7.1: Working labels and their definitions. The top six labels represent wanted structure, while the bottom three represent unwanted structure.

7.3.2 *Metadata.*

I build two general groups of binary labels related to (i) the characteristics of magical gems and (ii) the mediums depicted in each image. For the purposes of this chapter, these label groups represent wanted and unwanted structure respectively, with the ultimate goal being to identify transformations that remove the unwanted structures from computational cut-ups with out compromising the encoding of wanted structure. The first group is constructed using the wealth of structured information provided by CBd. The database provides collection, iconographic, and philological information about each gem. From this information, I build six labels that correspond to the three criteria associated with magical gems: magical names, signs, and iconography. For the second group, I build three binary labels by hand relying primarily on visual inspection.

Label	# Images	% Images	# Gems	% Gems
Magical Names	8576	71.4%	2096	65.5%
Characteres	3256	27.1%	787	24.6%
Anguipede (rough)	1559	13.0%	388	12.1%
Anguipede (clean)	829	6.9%	387	12.1%
Chnoubis (rough)	1326	11.0%	295	9.2%
Chnoubis (clean)	695	5.8%	292	9.1%
Drawing	2983	24.8%	869	27.1%
Photograph	8023	66.8%	2922	91.3%
Simulacrum	1008	8.4%	641	20.0%

Table 7.2: Label statistics for the number and proportion of positively labeled images and gems. Magical Names has the largest positive label class in terms of images, while Photograph has the largest in terms of gems.

MAGICAL NAMES. The "Magical Names" binary label indicates whether or not a gem is inscribed with a magical name. Magical Names labels are automatically generated based on whether a gem's CBd entry contains non-empty "Divine Names & Voces" or "Logoi" fields. Note these labels are coarse-grained in nature since *all* images of a gem will be the same, whether or not the inscription in question is visible within the image.

MAGICAL SIGNS. The "Characteres" binary label indicates whether or not a gem is engraved with a magical sign (i.e. *characteres*). Unlike magical names, there is no direct corresponding field within a gem's CBd entry. Instead, there is an invisible but searchable tag (i.e. Keyword: Characteres). Like "Magical Names" labels, "Characteres" labels operate at the gem, rather than image, level.

MAGICAL ICONOGRAPHY. I build four binary labels representing whether the Anguipede or Chnoubis iconographies are engraved on a gem. For each iconographic scheme, I construct two separate labels: one operating at the gem-level ("rough") and one at the image-level ("clean"). Like "Magical Signs" labels,

"Anguipede (rough)" and "Chnoubis (rough)" labels are built automatically using search-level tags. These tags collectively account for the variant types of each motif. For my purposes, I include all variations of the Anguipede and Chnoubis figures such as the lion-headed Anguipede and human-headed Chnoubis types. Then, I manually refine these "rough" labels into "clean" labels such that the images marked as having Anguipede or Chnoubis schemes do in fact depict these figures. Note that images containing Anguipede or Chnoubis motifs that are overlooked by "Anguipede (rough)" and "Chnoubis (rough)" are also missed by their refined counterparts.

MEDIUM. In addition to labels representing magical characteristics, I also construct three binary labels associated with the medium depicted in each image. The "Drawing," "Photograph," and "Simulacrum" labels correspond respectively to a drawing of a gem, a photograph of a gem,[10] and a photograph of a "simulacrum"—a physical, partial copy of a gem in the form of an impression or cast (both ancient and modern). Mediums cannot simply be ignored, since no one medium represents all gemstones. Moreover, these different mediums emphasize (and hide) different details of an object. For example, a shared iconographic element might only be recognized when comparing different mediums since color, lighting, and contrast can have a huge effect. In order for computational cut-ups to be useful, they must facilitate comparison across medium.

7.3.3 *Computational Cut-Ups.*

The construction of computational cut-ups follows the method outlined in Chapter 5. I create two sets of cut-ups; the first uses the original (color) images as input

[10] This includes photographs of non-gem objects (see note 8).

(RGB), while the second uses grayscaled versions as input (GS). All images are reduced to 224-by-224 pixel squares then passed as input to a ResNet-50 neural network [He et al., 2016] pretrained on ImageNet available through Keras [Chollet et al., 2015]. As a reminder, no textual information (e.g. transcriptions or textual descriptions) is used in the creation of these computational cut-ups. While inscriptions can be visually prominent features and thus likely incorporated by the underlying image model, the model is by no means literate.

7.4 ANALYSIS

My analysis of computational cut-ups for magical gems will be broken into two key stages. First, I must establish what information is captured by the initial computational cut-ups without any transformative intervention. I will measure how well characteristics of magical gems and depicted mediums, as represented by my working set of labels, are captured by the vector representations of the color and grayscale cut-ups. I will confirm that these cut-ups encode both structures of interest (i.e. characteristics of magical gems) and unwanted structure (i.e. medium). Then, with this framework of wanted and unwanted structure in place, I will investigate possible methods for transforming computational cut-ups into more useful forms. I will study which transformative interventions meaningfully dampen the encoding of unwanted structure without compromising the encoding of wanted structure.

Before going any further, it is worth discussing which of the working labels we should expect to be captured by computational cut-ups. Of all the labels, those relating to medium should have the best chance of being encoded within the vector spaces of computational cut-ups. Like music in Chapter 5, medium—especially

drawings—can easily be identified by both human experts and non-experts alike. Moreover, these categories can be identified using fairly simple and global features such as color range and relative contrast. Since such features are captured in the earliest layers of CNNs, we should expect computational cut-ups to encode these features in a way that is easily detected by a simple classifier. Although all mediums should be well-captured, I expect that the Drawing label will be the easiest to predict. It is easy to distinguish drawings from photographs in general, whereas distinguishing photographs of gems from photographs of casts and impressions requires a bit more nuance.

When it comes to the characteristics of magical gems, I expect that the iconographic features will be better captured than the textual ones. Although I believe that both have the possibility of being encoded, the Anguipede and Chnoubis iconographies are more visually coherent and dominant than the collective textual inscriptions of magical names and signs. While there is variation within each of the iconographic schemes, their visual diversity is smaller than the wide range of magical names and signs that are included within the Magical Names and Characteres labels. However, this reasoning does not account for the effects of label quality. While gem-level labels are much easier to acquire, they provide a much poorer description of the underlying visible characteristic than image-level labels and make the classification task inherently harder. Instead of identifying the presence (or absence) of a *visible* characteristic within an image, the task is now one of identifying whether an image shows a *gem* that possesses such a characteristic. Essentially, this means gem-level labels include many false positives from the image-level point of view. Nonetheless, these gem-level labels might still be predictable because of cooccurring characteristics. All of that being said, we should expect the Anguipede (clean) and Chnoubis (clean) labels to be captured by computational cut-ups but to a lesser degree than the three medium labels. In

contrast, the gem-level versions of these labels—Anguipede (rough) and Chnoubis (rough)—should have a much weaker encoding within the cut-ups. Similarly, the Magical Names and Characteres labels should have the weakest and possibly undetectable representation within the cut-ups.

7.4.1 *What structures are captured initially?*

I find that the depicted medium of an image is a dominant aspect of the computational cut-ups. We can see this prominence by visualizing the cut-ups' latent vector space using UMAP [McInnes et al., 2018] a dimensionality reduction method that preserves the global structure of the input data. Figure 7.4 shows that for both sets of cut-ups, images tend to cluster by medium type. Drawings are well separated from the photographs of gems and simulacra. Although photographs of gems and simulacra are intermixed, there appear to be some distinct groups.

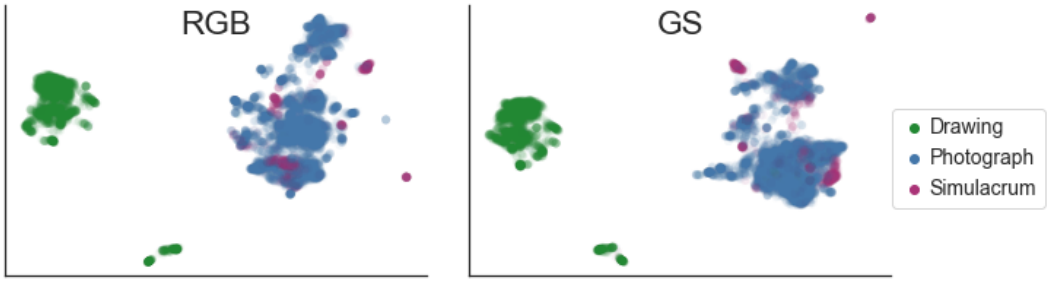

Figure 7.4: Two dimensional UMAP projections of computational cut-ups. Drawings are well-separated from photographs and simulacra.

In order to gain a better sense of which visual aspects are most prominent within the computational cut-ups, I examine how individual cut-ups tend to cluster within the overall embedding space using spherical k-means clustering, as was introduced in Chapter 3. This algorithm identifies k clusters as defined by their center points

(i.e. centroids) with a computational cut-up belonging to the cluster whose center is nearest to its vector. Just as a topic can be represented by its top words, as we saw in Chapters 3 and 6, a cluster can be represented by its top cut-ups, the cut-ups closest to the cluster's centroid. As seen in Figure 7.5, cut-ups do cluster by medium but this is more likely a byproduct of their tendency to cluster by gem color, material translucency, object shape, and background color. Notably, the final cluster shown in Figure 7.5 has top cut-ups of both photographs and simulacra. Rather than strictly representing photographs of gems or simulacra, it appears to be representing photographs of more fragmentary objects. While it is clear that the medium-related features dominate the representational space of computational cut-ups, cut-ups also cluster by a gem's engravings but this occurs more as a secondary grouping for a particular set of medium-oriented features. For example, of the drawings with white rather than off-white backgrounds there are separate clusters for rectangular gem faces (cluster 6), circular gem faces covered in textual inscriptions (cluster 3), circular gem faces that are predominantly unengraved (cluster 9), and circular gem faces with iconographic engravings (clusters 7 and 14). This kind of secondary clustering is more prominent for large numbers of clusters. For 100 clusters, cut-ups continue to cluster by medium-related characteristics, but also by engraving type.

This visualization of the computational cut-up embedding space also highlights additional encoded structures that might hamper the effort to computationally study magical gems. While image background is related to the depicted medium of an image, it also speaks to the collecting histories of the objects themselves. Photographs are typically taken within, if not by, the institution that houses the depicted object. So while backgrounds are not strictly unique at the collection or institution level, there will be detectable correlations. For example, all of the top images for the 20th cluster in Figure 7.5 have the same flat gray background and

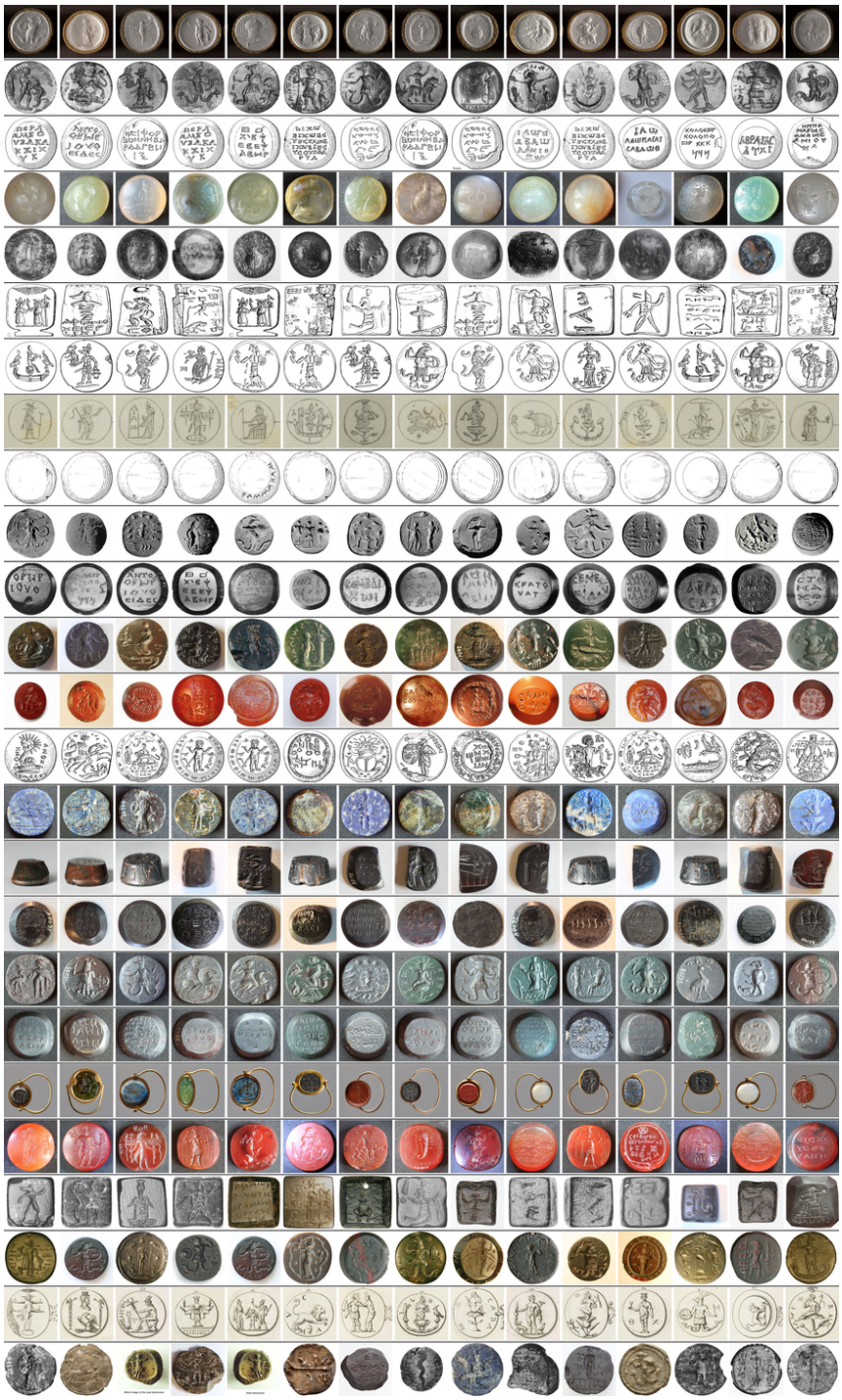

Figure 7.5: The computational cut-ups of gems cluster by medium, shape, color, and background. Each row represents the top 15 images of a cluster produced by the spherical k-means algorithm with $k = 25$.

are all housed within the Kunsthistorisches Museum in Vienna, Austria. Likewise, the top images for clusters 15, 18, 19, and 21 all possess near-identical dark gray, textured backgrounds and are housed within the British Museum. Similarly, while object shape typically reflects a gem's general shape (e.g. circular, rectangular), it can also include a gem's setting. There is a clear grouping of rings within the 25-cluster setting and I observe multiple ring and pendant clusters for 100-cluster setting. Suffice it to say that visually inspecting the clustering tendencies of computational cut-ups provides additional insights into which visual aspects are being captured and can call attention to more problematic structures without needing labels.

To more directly determine how well a given label's underlying structure is captured by computational cut-ups, I use the method introduced in Chapter 5 albeit with slight modification. I split the images into five groups of similar size such that all images of a gem are contained within the same group. Then, I obtain predictions for each group by training a classifier, in this case a linear support vector machine, on the images from the other four groups. By averaging the group-level results, I can compute an overall score for each label. Instead of reporting accuracy, I report balanced accuracy—the average proportion of correct predictions across label classes—to account for the variation of class imbalance across labels. So, a simple baseline classifier that always makes the same prediction will have a balanced accuracy score of 50% for any binary label.

As expected, the classifier results indicate that image medium can easily be predicted from computational cut-ups while "magical" characteristics are more difficult to identify. Surprisingly, there is no significant difference in performance between the original and grayscale cut-ups. Of the medium types, drawings are the easiest to identify while simulacra are the most difficult. Looking at the overall predictions, Drawing and Photograph label classes have very similar accuracy

within 0.01 of each other, while Simulacrum's class accuracies differ by about 0.1. This larger difference can be in part explained by Simulacrum's larger class imbalance; there are much fewer images of simulacra than drawings or photographs of gems. However, the difficulty of identifying photographs of simulacra also contributes to this difference. Distinguishing a gem cast or impression from an actual gem can require a close attention to material details and additional contextual knowledge.

Unsurprisingly, classification performance is much worse for the labels corresponding to magical gem features and especially bad for gem-level labels. While all of the gem-level labels have average balanced accuracy scores of over 50%, these scores are less significant given that there tends to be a large difference between class-level accuracies. As expected, the Magical Names and Characteres labels are the least predictable. The difficulty of the task, especially with the added noise of gem-level labels, is highlighted by the relatively poor performance of the classifiers during training. These simple classifiers are expected to achieve high accuracy during training, but of course at the cost of naively over-fitting to the training data. However, these two textual labels only achieve balanced accuracies nearing 85% during training which is far below the 97+% training scores of the other labels. So, not only are the trained classifiers not learning discriminative features that generalize to the test data, but they are unable to learn sufficiently discriminative features during training. In contrast, classifiers for the gem-level Anguipede and Chnoubis labels just perform poorly at test time. Promisingly, the image-level labels for the iconographic features perform much better with balanced accuracy scores above 70%. Examining the class-level accuracies directly, I find that negative labels are much easier to predict than positive labels. For both labels and cut-ups, negative class accuracy reaches 95%, while positive classes have accuracies between 50% and 60%. The Anguipede (clean) label has a positive class accuracy of 58% for

both color and grayscale cut-ups, which is an encouraging sign that computational cut-ups are encoding structures related to the Anguipede iconographic scheme. The Chnoubis (clean) label has much weaker and more variable positive class accuracy across cut-ups with 50% for the color cut-ups and 54% for the grayscale ones. While this difference in performance is not significant, it does suggest that there are other structures, such as color, that are negatively impacting classification, especially given the higher observed variability across runs.

The most confident correct predictions (i.e. true positives and true negatives) and incorrect predictions (i.e. false negatives and false positives) for the Anguipede (clean) and Chnoubis (clean) classifiers provide additional support that computational cut-ups encode structures of these iconographic schemes that are independent from medium-related ones. Promisingly, multiple mediums are represented within these top (mis)predictions as we can see in Figures 7.6 and 7.7. In fact, the top two true positives for both labels are not photographs of gems but instead drawings for the Anguipede figure and simulacra—in this case gem impressions—for the Chnoubis figure. Many of the top true positives for the Anguipede iconographic scheme are drawings and grayscale images which is a bit unexpected given that the computational cut-ups are also encoding color. In contrast, the Chnoubis figure's top true positives include more color photographs of gems. Nonetheless, both capture potentially noteworthy characteristics associated with each scheme. Many of the most confident true positives for the Anguipede label correspond to gems with broken edges. This damage might indicate that Anguipede engraved gems are made from more fragile material, are more likely to be have been set within jewelry and subsequently removed, and in some cases are intentionally broken (see Nagy, 2019, p. 186). For the Chnoubis label, the most confident positive predictions are of light, semitransparent materials while the most confident negative predictions are not. This suggests that Chnoubis figures

Label	RGB (%)	GS (%)
Magical Names	59.78 ± 1.13	59.51 ± 1.69
Characteres	57.15 ± 1.63	56.52 ± 1.75
Anguipede (rough)	62.86 ± 1.44	63.42 ± 1.34
Anguipede (clean)	76.45 ± 1.47	76.76 ± 1.27
Chnoubis (rough)	64.73 ± 0.36	65.80 ± 2.49
Chnoubis (clean)	72.68 ± 1.18	74.59 ± 2.37
Drawing	99.80 ± 0.04	99.86 ± 0.06
Photograph	97.79 ± 0.32	97.41 ± 0.34
Simulacrum	92.99 ± 1.96	92.32 ± 1.82

Table 7.3: Mean and standard deviation of balanced accuracy scores for the initial computational cut-ups. Object medium is easily predicted from computational cut-ups, while "magical" characteristics are more difficult to identify.

Figure 7.6: Four most confident true positive (top) and true negative (bottom) classifications for image-level Anguipede (left) and Chnoubis (right) labels. Multiple mediums are represented within these predictions.

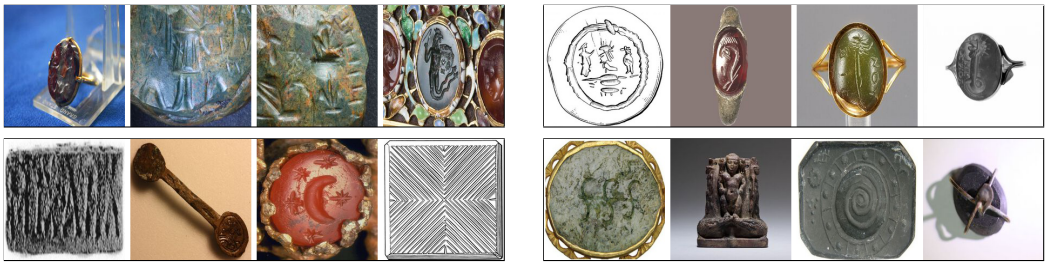

Figure 7.7: Four most confident false negative (top) and false positive (bottom) classifications for image-level Anguipede (left) and Chnoubis (right) labels. The false positives are less interpretable than the false negatives.

tend to be engraved on light, semitransparent stone which matches past analysis of Chnoubis gems [Mastrocinque, 2011] and emphasizes the importance of gem material.

The top negative predictions for both labels show that gemstones in ring and pendant settings are not associated with the iconographic schemes by the classifiers. The classifiers also tend to overlook Anguipede and Chnoubis figures when they are not the dominant, central element of the visible gem face as well as when they vary from the more typical versions of the iconographic scheme. For correct identification, the legs of the Anguipede figure need to be ess-shaped snakes that collectively form a *w*-like shape as in the top true positives in Figure 7.6. Likewise, the Chnoubis figure's snake body must also possess ess-shaped curves if not full loops. Funnily enough, this apparent discriminative aspect of the Chnoubis figure makes the classifier more likely to mistake the Chnoubis sign as a Chnoubis figure as can be seen in Figure 7.7. This particular trait of the Chnoubis classifiers helped uncover mislabelings within the Chnoubis labels where images (and gems) with the Chnoubis sign but not the Chnoubis figure were incorrectly labeled. However, while the false negatives are generally interpretable, the false positives as a whole tend to be more incomprehensible. While these classification issues are unlikely to go away entirely with the dampening of medium-related structures, such transformations could improve our ability to interpret a classifier's predictions.

7.4.2 *Removing unwanted structure*

Given that gem medium is a dominant feature of both the color and grayscale cut-ups, it might be highly correlated with the most significant dimensions identified by singular value decomposition (SVD). If there are, it should be possible to

dampen the presence of this structure by subtracting out these dimensions. This is a simplified version of Bolukbasi et al. [2016]'s formative method for debiasing word embeddings. I do not apply Bolukbasi et al. [2016]'s method directly because it requires vector pairs in order to find a biased subspace.

To identify correlated SVD dimensions, I measure the cosine similarity between the left singular vectors produced by SVD (i.e. SVD dimensions) and the binary vectors representing each label with the ith element corresponding to the ith image's label (i.e. 0 for the negative class and 1 for the positive class). I find that the top two SVD dimensions heavily correlate with the Photograph and Drawing labels for both color and grayscale cut-ups with similarity scores above 0.8. In contrast, the Simulacrum label does not correlated strongly with any particular dimension.

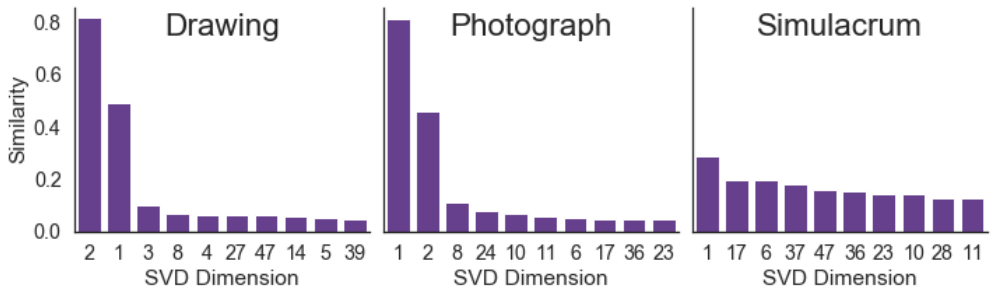

Figure 7.8: The most similar SVD dimensions for RGB cut-ups and medium labels. The Drawing and Photograph labels have high similarity with the first two dimensions, while the Simulacrum label does not have high similarity with any specific dimension.

Turning to the wanted, "magical" structures, I find that neither the Anguipede nor Chnoubis image-level labels are highly correlated with the SVD dimensions. Their highest similarities are with the first dimension but with scores below 0.27. In contrast, the Magical Names label is very similar to the first SVD dimension with similarities above 0.83 and the Characteres label also has some similarity with this dimension with scores above 0.50. This relationship can be partially explained

by how these labels overlap with the Photograph label: 71.8% of photographic images of gems correspond to gems engraved with magical names, while only 27.6% correspond to gems inscribed with *characteres*, 6.7% depict the Anguipede figure, and 5.6% depict the Chnoubis figure. Given that the Magical Names and Characteres labels operate at the gem- rather than image-level and are not easily visible to a linear classifier, I find it reasonable to go ahead and remove the first and second SVD dimensions that heavily correlate with Photograph and Drawing labels.

Label	RGB (%)	GS (%)
Magical Names	59.75 ± 1.33	59.57 ± 1.62
Characteres	57.23 ± 1.56	56.47 ± 1.68
Anguipede (clean)	75.71 ± 1.81	76.69 ± 1.11
Chnoubis (clean)	72.21 ± 2.45	74.50 ± 3.06
Drawing	$\mathbf{38.43 \pm 1.27}$	$\mathbf{41.08 \pm 1.72}$
Photograph	$\mathbf{56.13 \pm 1.58}$	$\mathbf{56.36 \pm 0.79}$
Simulacrum	92.82 ± 1.63	92.36 ± 1.56

Table 7.4: Mean and standard deviation of balanced accuracy scores for modified cut-ups with the first two SVD dimensions removed. Bold values indicate statistically significant drops in performance. Removing these two dimensions has made it harder to identify drawings and photographs of gems, but not other structure.

Subtracting out the top two SVD dimensions from the computational cut-ups significantly decreases classifier performance for Drawing and Photograph labels, but has little effect on the performance of other labels. The Drawing classifiers now perform worse than random and the Photograph classifiers perform no better than the Characteres classifiers. It is somewhat surprising that the Magical Names and Characteres classifiers experience no significant loss in performance given these labels were similar to one of the removed SVD dimensions. Then again, this suggests that the observed similarities are more related to overlap with the Photograph label rather than something inherent to magical names or symbols.

While the classifiers are becoming more blind to photographs and drawings of gems, this is only with respect to linear expression. The UMAP projections in Figure 7.9 show that there are still dominant clusters of photographs and drawings of gems although these clusters are less distinct and homogeneous. While there is little motivation to remove more SVD dimensions from the transformed cut-ups, it is irrelevant because no existing SVD dimension shares high similarity with any label. While simulacra are very easy to identify, their structure is spread across many dimensions rather than concentrating among a small few.

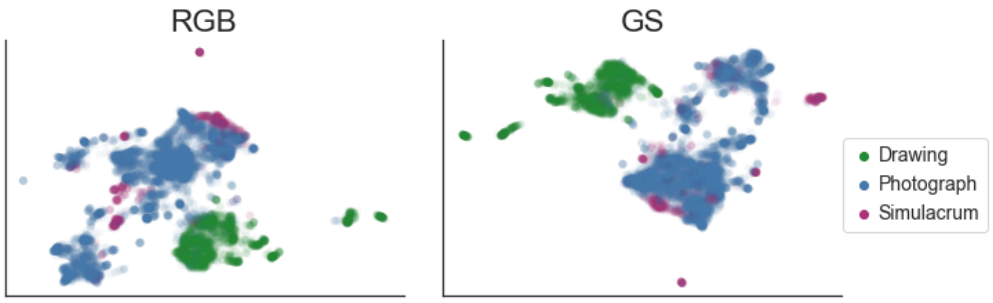

Figure 7.9: Two dimensional UMAP projections of computational cut-ups with first two SVD components removed. Drawings are not as well-separated but remain fairly distinct from photos and casts.

These results match Gonen and Goldberg [2019]'s findings that debiasing techniques such as Bolukbasi et al. [2016]'s do not fully debias embedding spaces; bias is still reflected in how the "debiased" word vectors cluster together. Unlike the fairness setting, medium-related structures do not need to be removed wholesale. Each medium usefully captures some important visual characteristics and not others. Instead, the overall goal is to dampen the prominence of medium-related structures so that other structures can become more prominent as was the case in Chapter 6. Table 7.4 shows that there is still more linear structure related to depicted medium that can be removed, particularly relating to photographs of simulacra. This leads me to apply a related, but more sophisticated and iterative

approach to removing linear structure correlated with data-level labels: Iterative Null-space Projection (INLP) recently proposed by Ravfogel et al. [2020].

For a given structure that we want to remove, INLP iteratively trains linear classifiers to predict this structure and remove the directions corresponding to the decision boundaries of the linear classifiers (i.e. its null space). This method is very easy to use since it only requires data vectors and labels corresponding to the structure to be removed; these labels do not need to be binary. Using Ravfogel et al. [2020]'s implementation,[11] I apply INLP with 10, 25, 50, and 100 iterations (i) to remove a combined three-class Medium label composed of the positive labels of the Drawing, Photograph, and Simulacrum labels. This new three-class label encompasses the entire set of working images.

INLP projections effectively remove photograph-, drawing-, and simulacrum-related structures without negatively impacting encoding of the wanted, "magical" structures. Table 7.5 shows that for color computational cut-ups all medium-related classifiers perform significantly worse than 0.5 balanced accuracy of the simple baseline, while the performance for the other labels' classifiers do not change significantly. There are no significant differences in classifier performance for grayscale cut-ups. Increasing the number of iterations of INLP makes linear classifiers more oblivious to the targeted unwanted structure. The trained classifiers balanced accuracies drop well below 50% for the test data, but also plummet to 60–75% for the training data. Of course there is a saturation point where linear models no longer recover any useful decision boundaries for predicting the blinded labels. In the extreme, continual applications of INLP will negatively affect encodings of labels of interest as seen in Table 7.5 for 100 iterations. I observe that the saturation

[11] https://github.com/shauli-ravfogel/nullspace_projection

point is reached near 50 iterations,[12] but iteration choice is fairly flexible since INLP can be continued from where the last run ended.

While INLP transformations do not significantly impact the classifier performance for the non-medium labels, they do have some effect on classifier interpretation. Over half of the four most confident (mis)predictions for the image-level Anguipede and Chnoubis labels change when comparing the classifiers before and after applying 50 iterations of INLP (see Figures 7.6 and 7.7; Figures 7.10 and 7.11). For true positives and false negatives, these changes are more of a reordering of high confidence predictions; the new and replaced images tend to rank within the top 25. Still there are a few exceptional cases. For the Chnoubis label, a gem impression depicting a radiated snake wrapped around a cylindrical base (CBd-209) has gone from a confident misprediction to a confident correct prediction. For the Anguipede label, the updated top false negatives helped identify two images of a gem face that were mislabeled (CBd-1918).[13] In contrast, the differences for true negatives and false positives speak to a more substantial shift in how images without Anguipede and Chnoubis iconographies are perceived by the classifiers. The true negatives for both labels more strongly indicate that neither iconographic scheme is associated with rings or text heavy inscriptions by the classifiers. The false positives still remain fairly opaque but suggest that Anguipede figures are being associated with damaged gems, possibly large, lower-case omegas (i.e. ω), and possibly humanoid figures with heads turned in profile. In turn, the false positives for the Chnoubis figure suggest that it is being confused with snakes, typically in ess-shaped poses, and gems with lighter, semi-transparent materials. So while INLP transformations do not make characteristics of magical gems more

[12] 50 is likely higher than necessary, but is sufficient

[13] Correcting the labels of these two images had no significant effect on classification performance.

Label	$i = 10$	$i = 25$	$i = 50$	$i = 100$
Magical Names	59.53 ± 1.08	59.56 ± 1.08	58.97 ± 1.49	57.20 ± 0.95
Characteres	57.15 ± 1.78	56.61 ± 1.95	57.14 ± 1.83	54.78 ± 1.22
Anguipede (clean)	76.06 ± 0.67	76.25 ± 1.17	75.25 ± 1.02	$\mathbf{68.49 \pm 0.61}$
Chnoubis (clean)	72.44 ± 2.64	72.19 ± 2.33	70.89 ± 2.22	67.60 ± 1.01
Drawing	$\mathbf{76.83 \pm 1.45}$	$\mathbf{36.10 \pm 1.31}$	$\mathbf{25.22 \pm 1.02}$	$\mathbf{24.30 \pm 1.24}$
Photograph	$\mathbf{56.84 \pm 1.42}$	$\mathbf{29.33 \pm 1.14}$	$\mathbf{24.97 \pm 0.65}$	$\mathbf{24.32 \pm 0.51}$
Simulacrum	$\mathbf{54.16 \pm 1.53}$	$\mathbf{38.25 \pm 2.16}$	$\mathbf{29.96 \pm 1.90}$	$\mathbf{27.25 \pm 1.33}$

Table 7.5: Mean and standard deviation of balanced accuracy scores for transformed color cut-ups using Iterative Null-space Projection (INLP) with $i \in \{10, 25, 50, 100\}$ iterations. Bold values indicate statistically significant drops in performance. INLP effectively removes medium-related structure without harming the other structures of interest.

Figure 7.10: Four most confident true positive (top) and true negatives (bottom) classifications for image-level Anguipede (left) and Chnoubis (right) labels using INLP transformed cut-ups as input ($i = 50$).

Figure 7.11: Four most confident false negatives (top) and false positives (bottom) classifications for image-level Anguipede (left) and Chnoubis (right) labels using INLP transformed cut-ups as input ($i = 50$).

prominent (at least for simple linear classifiers), they do make classifier predictions more humanly interpretable.

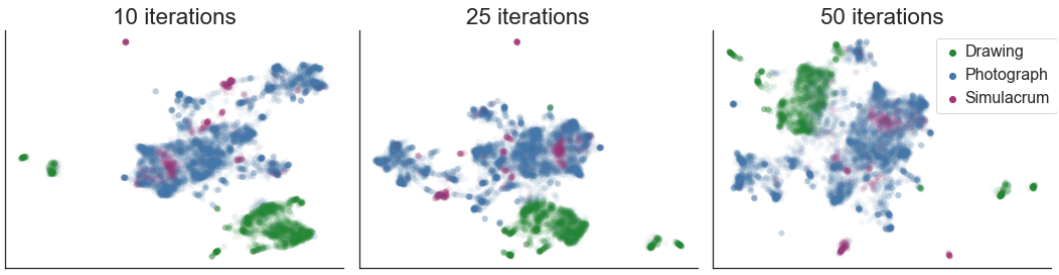

Figure 7.12: Two dimensional UMAP projections of transformed color cut-ups using INLP with $i \in \{10, 25, 50\}$. While medium-specific clusters are still present after 50 iterations of INLP, these clusters are more diffuse and more overlapping.

Figure 7.13: Three medium cross-cutting clusters identified by spherical k-means for INLP-transformed computational cut-ups ($i = 25$). INLP enables the formation of a small number of cross-cutting clusters within the embedding space of computational cut-ups.

INLP clearly removes medium-related linear structures, but the encoding of depicted mediums within computational cut-ups might not be predominantly linear. The UMAP projections of the INLP-transformed cut-ups in Figure 7.12 confirm that depicted medium is still captured by computational cut-ups if not quite as strongly. After 50 iterations of INLP, there is still distinct, medium-specific clustering but it tends to be less concentrated and more overlapping. This indicates that these transformed computational cut-ups still contain problematic non-linear

structures related to specific medium types, but that INLP did remove some of the contributing linear structures. However, the clusters of the transformed cut-ups as identified by spherical k-means indicate that persisting medium-related structures still need to be dealt with. The majority of clusters are still medium-specific, however there are some promising cross-cutting clusters. Figure 7.13 depicts three such clusters. The first two contain all three medium types within the top 5 images and appear to be capturing gem shape. Although the third cluster's top images only depict photographs of gems and casts, they seem to share a common iconography of a central humanoid figure standing in profile. Ultimately, INLP can diminish the encoding of unwanted structures from computational cut-ups but it only removes linear—and not non-linear—aspects of these structures.

7.5 RELATED WORK

The task of removing unwanted *known* structure from an embedding space is very similar to learning fair representations [Zemel et al., 2013; Edwards and Storkey, 2016; Madras et al., 2018; Quadrianto et al., 2019]. In this work, the general goal is to transform a data set X with some sensitive variables S to a new space Z such that Z is still useful for predicting a target variable Y. Typically S and Y are binary labels, but most methods can be extended to multi-class settings. Fairness can be defined in a number of ways, but one of the most common is *demographic parity* where the conditional probability of a positive outcome for Y given S is the same between the protected and unprotected classes of S.

Using methods from this literature can be useful, but there is a cost in flexibility. Most methods require the knowledge of Y *a priori*, but it may be difficult to identify and construct a specific Y for exploring the representational space of a working

collection. Ultimately, it is much easier to identify known, unwanted structures than specifying unknown, interesting properties of a collection before studying it. So, the requirement of target variable Y is overly limiting.

In one of the more recent adversarial learning methods, LAFTR [Madras et al., 2018], a target variable Y is not explicitly required, but these adversarial learning methods come with their own trade-offs. They tend to require a large amount of domain knowledge to run and adapt to a new dataset or problem setting. Additionally, these models are very sensitive to hyperparameter choices. While I believe adversarial methods have much promise, more work needs to be done for them to be a realistic option for out-of-domain users.

7.6 CONCLUSION

In this work, I have shown the potential of using computational cut-ups for studying magical gems and that these cut-ups can be made more useful by eliminating unwanted, structures from their encodings. Iterative Null-space Projection (INLP) provides a simple method for eliminating unwanted structure and only requires data-level labeling of said structure. However, this method can only fully remove linear structures. Persisting non-linear aspects will remain encoded within the representations of computational cut-ups and continue to influence the clustering of the embedding space. Nonetheless, I have demonstrated that INLP can improve the human interpretability of computational cut-up analysis. All told, this work is a promising step in enabling an image-based computational analysis of magical gems and more broadly studying large visual material collections, especially from museums and archives.

8

A SYMBIOTIC FUTURE FOR MACHINE LEARNING & THE HUMANITIES

In this dissertation I present a series of work that demonstrate the opportunities for machine learning and the humanities to help one another. Machine learning methods have the potential to be additional analytical tools for humanities scholarship that both support and expand the directions of current research. Perhaps less obvious but no less important is the converse. The humanities can provide an additional perspective on the affordances of machine learning methods. Testing, expanding, and at times discovering the limits of computational models, but also highlighting and possibly breaking model assumptions presumed and established by their creators and the broader machine learning community. Clearly, there is a larger range of interplay between the humanities and machine learning (and technology more broadly), especially as it pertains to ethics, but this dissertation focuses specifically on machine learning and humanities research.

Part I focuses on understanding what machine learning models actually learn. Chapters 3 and 4 study the geometries of the vector spaces of word representations. Chapter 3 examines how the clustering of contextualized word representations produced by BERT and other pre-trained language models can be used to identify and locate the themes of a specific collection of texts in a way similar to topic modeling. Chapter 4 investigates how the geometries of word embeddings differ across languages. Clearly, these works propose new ways to compare language models through the geometries of their representational spaces, but they are also driven by

humanities use cases. The first focuses on a new instance of the topic modeling-like work flow commonly used in the humanities, while the second provides a new way of comparing human languages from a (computational) distributional semantics lens. In contrast, the more humanities-focused Chapter 5 asks how deep learning image models sees Dada and the avant-garde. This chapter proposes a new methodological framework for studying visual humanities collections using machine learning models, specifically convolutional neural networks. In addition to providing a new tool for (digital) humanities research, it demonstrates how statistical machine learning methods can be used qualitatively. In sum, a humanities perspective can drive *why* and expand *how* we ask what machine learning models learn. The results of which benefit both communities providing more transparent and attenuated tools for humanities research as well as deeper understandings of the machine learning models themselves.

Part II uses purposeful data modification to transparently direct what models learn. Chapter 6 focuses on the textual domain with science fiction novels and U.S. state supreme court opinions, while Chapter 7 focuses on the visual, image domain with engraved magical gemstones from the Graeco-Roman world. These works focus on a problem often encountered by humanities scholars when using machine learning methods that of models predominantly—if not entirely—learning known, uninteresting aspects of a collection. While this can be reassuring that the model is learning something meaningful, meaningful does not necessarily imply useful or insightful. Chapter 6 shows that topic models learn discourses that can often correlate by known, unwanted contexts and that these problematic topics are not inherently obvious by inspection. An apparent topic on robots might be more exclusively about Isaac Asimov's *Robots* series, or an apparent topic on water rights only covers Hawaii-specific issues. This work shows that this problem can be mitigated by selectively subsampling words that are overly

context-specific. Chapter 5 seeks to apply this use of purposeful data modification to the image domain. Namely, improving the computational analysis of magical gem characteristics by mitigating the encoding of depicted medium. Ultimately, this problem is more difficult because of the limited interpretability of image vector features. Nonetheless, linear medium-related structures are identified and removed. Although the remaining non-linear structures maintain the encoding of depicted medium, its presence is diminished and the interpretability of the magical characteristics is improved. Altogether, Part II embodies how the humanities and machine learning can help one another. By addressing problems faced by humanists, we not only gain models that are more useful for humanities research, but also gain a better understanding of *which* structures models are capable of learning. Moreover, purposeful data modification shifts the focus to the transformative power of data itself. A fixed model can learn a wide range of structures, just as a microscope can view a wide range of phenomena with the right stains.

This dissertation has focused on a small slice of how machine learning and humanities research beneficially overlap. I have focused predominantly on *unsupervised* models—models that learn patterns without any guidance through labelled examples—but this is only one flavor of computational analysis used by digital humanists and computational social scientists. Similarly, I have focused on problem settings with massive quantities of data where automation is a necessity. Even though automation is needed, it is not a replacement of scholarly analysis but an aid. It should be used in conjunction with other methods of analysis, both qualitative and quantitative.

While machine learning, statistics, and technology more broadly are often pitted against the humanities, there is a path for their relationship to be mutually beneficial and symbiotic rather than antithetical and parasitic. It requires an awareness, recognition, and respect of each other's scholarship; a value of both qualitative

and quantitative research. I hope this dissertation exemplifies the exciting range of possibilities in combining these multiple perspectives. Machine learning has the potential to be a powerful tool that complements existing humanities scholarship. Likewise, humanities research and collections test, challenge, and expand the affordances of machine learning models. Machine learning helps the humanities and the humanities helps machine learning.

BIBLIOGRAPHY

Achlioptas, Dimitris (2001). "Database-Friendly Random Projections." In: PODS 2001. Association for Computing Machinery, pp. 274–281. DOI: 10.1145/375551.375608 (cit. on p. 34).

Aharoni, Roee and Yoav Goldberg (2020). "Unsupervised Domain Clusters in Pretrained Language Models." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7747–7763. DOI: 10.18653/v1/2020.acl-main.692 (cit. on p. 22).

AlSumait, Loulwah, Daniel Barbará, James Gentle, and Carlotta Domeniconi (2009). "Topic Significance Ranking of LDA Generative Models." In: *ECML PKDD 2009*. Springer Berlin Heidelberg, pp. 67–82. DOI: 10.1007/978-3-642-04180-8_22 (cit. on p. 91).

Antoniak, Maria and David Mimno (2018). "Evaluating the Stability of Embedding-based Word Similarities." In: *Transactions of the Association for Computational Linguistics* 6, pp. 107–119. DOI: 10.1162/tacl_a_00008 (cit. on p. 57).

Antoniak, Maria, David Mimno, and Karen Levy (2019). "Narrative Paths and Negotiation of Power in Birth Stories." In: *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, pp. 1–27. DOI: 10.1145/3359190 (cit. on p. 1).

Arnold, Taylor, Lauren Tilton, and Annie Berke (2019). "Visual Style in Two Network Era Sitcoms." In: *Journal of Cultural Analytics*. DOI: 10.22148/16.043 (cit. on p. 15).

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu (2013). "A Practical Algorithm for Topic

Modeling with Provable Guarantees." In: *Proceedings of the 30th International Conference on Machine Learning*. PMLR 28, pp. 280–288. URL: http://proceedings.mlr.press/v28/arora13.html (cit. on p. 20).

Assael, Yannis, Thea Sommerschield, and Jonathan Prag (2019). "Restoring ancient text using deep learning: a case study on Greek epigraphy." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, pp. 6368–6375. DOI: 10.18653/v1/D19-1668 (cit. on p. 14).

Bamman, David, Olivia Lewke, and Anya Mansoor (2020). "An Annotated Dataset of Coreference in English Literature." In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, pp. 44–54. URL: https://www.aclweb.org/anthology/2020.lrec-1.6 (cit. on p. 14).

Bansal, Mohit, Kevin Gimpel, and Karen Livescu (2014). "Tailoring Continuous Word Representations for Dependency Parsing." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 809–815. DOI: 10.3115/v1/P14-2131 (cit. on pp. 53, 54).

Barrett, Caitlín (forthcoming). "Magical Gems as Material Texts: Contextual Analysis of Greco-Roman Amulets." In: Textual Archaeology of the Ancient Near East: Are We Doing It Wrong? Ed. by Yağmur Heffron (cit. on pp. 110, 117, 121–123).

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo (2018). "Individuals, institutions, and innovation in the debates of the French Revolution." In: *Proceedings of the National Academy of Sciences* 115.18, pp. 4607–4612. DOI: 10.1073/pnas.1717729115 (cit. on p. 11).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (2003). "A Neural Probabilistic Language Model." In: *Journal of Machine Learning Research*

3, pp. 1137–1155. URL: https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf (cit. on p. 85).

Betz, Hans Dieter (1992). *The Greek Magical Papyri in Translation, Including the Demotic Spells, Volume 1*. Chicago: University of Chicago Press (cit. on pp. 113, 120).

Bischof, Jonathan and Edoardo Airoldi (2012). "Summarizing topical content with word frequency and exclusivity." In: *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, pp. 201–208. URL: https://icml.cc/Conferences/2012/papers/113.pdf (cit. on p. 28).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation." In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (cit. on pp. 10, 19, 85).

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. DOI: 10.1162/tacl_a_00051 (cit. on pp. 47, 56).

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., pp. 4349–4357. URL: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf (cit. on pp. 141, 143).

Bommasani, Rishi, Kelly Davis, and Claire Cardie (2020). "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4758–4781. DOI: 10.18653/v1/2020.acl-main.431 (cit. on p. 25).

Bonner, Campbell (1950). *Studies in magical amulets: chiefly Graeco-Egyptian*. Ann Arbor: University of Michigan Press (cit. on pp. 117–119, 121, 124).

Boot, Peter (2017). "A Database of Online Book Response and the Nature of the Literary Thriller." In: *Book of Abstracts of DH2017*. Alliance of Digital Humanities Organizations. URL: https://dh2017.adho.org/abstracts/208/208.pdf (cit. on p. 1).

Bourrier, Karen and Mike Thelwall (2020). "The Social Lives of Books: Reading Victorian Literature on Goodreads." In: *Journal of Cultural Analytics*. DOI: 10.22148/001c.12049 (cit. on p. 1).

Boyd-Graber, Jordan, Yuening Hu, and David Mimno (2017). "Applications of Topic Models." In: *Foundations and Trends in Information Retrieval* 11.2-3, pp. 143–296. DOI: 10.1561/1500000030 (cit. on pp. 20, 37).

Boyd-Graber, Jordan, David Mimno, and David Newman (2014). "Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements." In: *Handbook of Mixed Membership Models and Their Applications*, pp. 225–254. DOI: 10.1201/b17520-21 (cit. on p. 87).

Broadwell, Peter, David Mimno, and Timothy Tangherlini (2017). "The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification." In: *Journal of Cultural Analytics*. DOI: 10.22148/16.012 (cit. on p. 66).

Brown, Nicole M., Ruby Mendenhall, Michael Black, Mark Van Moer, Karen Flynn, Malaika McKee, Assata Zerai, Ismini Lourentzou, and ChengXiang Zhai (2019). "In Search of Zora/When Metadata Isn't Enough: Rescuing the Experiences of Black Women Through Statistical Modeling." In: *Journal of Library Metadata* 19.3-4, pp. 141–162. DOI: 10.1080/19386389.2019.1652967 (cit. on p. 11).

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer (1992). "Class-Based *n*-gram Models of Natural Language."

In: *Computational Linguistics* 18.4, pp. 467–480. URL: https://www.aclweb.org/anthology/J92-4003 (cit. on p. 20).

Buurma, Rachel Sagner (2015). "The fictionality of topic modeling: Machine reading Anthony Trollope's Barsetshire series." In: *Big Data & Society* 2.2. DOI: 10.1177/2053951715610591 (cit. on p. 11).

Capitanu, Boris, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, and J. Stephen Downie (2016). *The HathiTrust Research Center Extracted Feature Dataset (1.0)*. DOI: 10.13012/J8X63JT3 (cit. on p. 88).

"CBd-8" (Oct. 11, 2017). In: *The Campbell Bonner Magical Gems Database (2010-), developed at the Museum of Fine Arts, Budapest, editor-in-chief: Á. M. Nagy*. URL: http://cbd.mfab.hu/cbd/8 (cit. on p. 126).

Chandrahas, Aditya Sharma, and Partha Talukdar (2018). "Towards Understanding the Geometry of Knowledge Graph Embeddings." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 122–131. DOI: 10.18653/v1/P18-1012 (cit. on p. 51).

Chang, Jonathan and David Blei (2009). "Relational Topic Models for Document Networks." In: *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*. PMLR 5, pp. 81–88. URL: http://proceedings.mlr.press/v5/chang09a.html (cit. on p. 87).

Chemudugunta, Chaitanya, Padhraic Smyth, and Mark Steyvers (2006). "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model." In: *Advances in Neural Information Processing Systems*. Vol. 19. MIT Press, pp. 241–248. URL: https://proceedings.neurips.cc/paper/2006/file/ec47a5de1ebd60f559fee4afd739d59b-Paper.pdf (cit. on p. 87).

Chen, Danqi and Christopher Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks." In: *Proceedings of the 2014 Conference on Empirical*

Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp. 740–750. DOI: 10.3115/v1/D14-1082 (cit. on pp. 13, 41).

Chen, Wenhu, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter (2016). "Guided Alignment Training for Topic-Aware Neural Machine Translation." In: *Proceedings of AMTA 2016, Vol. 1: MT Researchers' Track*, pp. 121–134. URL: https://amtaweb.org/wp-content/uploads/2016/10/AMTA2016_Research_Proceedings_v7.pdf (cit. on p. 23).

Chiu, Jason P.C. and Eric Nichols (2016). "Named Entity Recognition with Bidirectional LSTM-CNNs." In: *Transactions of the Association for Computational Linguistics* 4, pp. 357–370. DOI: 10.1162/tacl_a_00104 (cit. on p. 13).

Chollet, François et al. (2015). *Keras*. https://keras.io (cit. on p. 131).

Collins, Derek (2008). *Magic in the Ancient Greek World*. Malden, MA: Blackwell Publishing Ltd (cit. on p. 112).

Collins, Michael (2002). "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms." In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, pp. 1–8. DOI: 10.3115/1118693.1118694 (cit. on p. 44).

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (2018). "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 2126–2136. DOI: 10.18653/v1/P18-1198 (cit. on pp. 9, 48).

Cotterell, Ryan and Hinrich Schütze (2015). "Morphological Word-Embeddings." In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for

Computational Linguistics, pp. 1287–1292. DOI: 10.3115/v1/N15-1140 (cit. on p. 41).

Dasen, Véronique and Árpád M. Nagy (2012). "Le serpent léontocéphale Chnoubis et la magie de l'époque romaine impériale." In: *Anthropozoologica* 47.1, pp. 291–314. DOI: 10.5252/az2012n1a8 (cit. on pp. 118–120).

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis." In: *Journal of the American Society for Information Science* 41.6, pp. 391–407. DOI: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9 (cit. on pp. 10, 85).

Delatte, Armand (1914). "Études sur la magie grecque IV: Amulettes inédites des Musées d'Athènes." In: *Le musée belge: revue de philologie classique* 18, pp. 21–96 (cit. on p. 116).

Delatte, Armand and Phillipe Derchain (1964). *Les intailles magiques gréco-égyptiennes*. Paris: Bibliothèque nationale (cit. on p. 117).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "ImageNet: A large-scale hierarchical image database." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on pp. 14, 109).

Denny, Matthew and Arthur Spirling (Sept. 2017). "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." In: *SSRN*. DOI: 10.2139/ssrn.2849145 (cit. on p. 87).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. DOI: 10.18653/v1/N19-1423 (cit. on pp. 13, 19, 21).

Dhillon, Inderjit S. and Dharmendra S. Modha (2001). "Concept Decompositions for Large Sparse Text Data Using Clustering." In: *Machine Learning* 42, pp. 143–175. DOI: 10.1023/A:1007612920971 (cit. on p. 25).

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei (2020). "Topic Modeling in Embedding Spaces." In: *Transactions of the Association for Computational Linguistics* 8, pp. 439–453. DOI: 10.1162/tacl_a_00325 (cit. on p. 23).

Doyle, Gabriel and Charles Elkan (n.d.). "Accounting for Burstiness in Topic Models." In: *Proceedings of the 26th International Conference on Machine Learning*. Association for Computing Machinery, pp. 281–288. DOI: https://doi.org/10.1145/1553374.1553410 (cit. on p. 95).

Dzwiza, Kirsten (2019). "Magical Signs: An Extraordinary Phenomenon or Just Business as Usual? Analysing Decoration Patterns on Magical Gems." In: *Magical Gems in Their Contexts: Proceedings of the International Workshop held at the Museum of Fine Arts Budapest, 16–18 February 2012*. Roma: "L'Erma" di Bretschneider, pp. 59–84 (cit. on p. 124).

Edwards, Harrison and Amos J. Storkey (2016). "Censoring Representations with an Adversary." In: *International Conference on Learning Representations 2016*. URL: https://arxiv.org/abs/1511.05897 (cit. on p. 148).

Endo, Yasunori and Sadaaki Miyamoto (2015). "Spherical k-means++ clustering." In: *International Conference on Modeling Decisions for Artificial Intelligence 2015*. Springer, pp. 103–114. DOI: 10.1007/978-3-319-23240-9_9 (cit. on p. 25).

Ethayarajh, Kawin (2019). "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, pp. 55–65. DOI: 10.18653/v1/D19-1006 (cit. on p. 26).

Faraone, Christopher A. (2011). "Text, Image, and Medium: The Evolution of Graeco-Roman Magical Gemstones." In: *'Gems of Heaven': Recent Research on Engraved Gemstones in Late Antiquity, c. AD 200–600*. London: The British Museum, pp. 50–61 (cit. on p. 119).

Faraone, Christopher A. (2018). *The Transformation of Greek Amulets in Roman Imperial Times*. Philadelphia: University of Pennsylvania Press (cit. on pp. 115, 117).

Firth, John R. (1957). "A synopsis of linguistic theory, 1930–1955." In: *Studies in linguistic analysis*. Oxford: Blackwell, pp. 1–32 (cit. on p. 10).

Frankfurter, David (1997). "Ritual Expertise in Roman Egypt and the Problem of the Category 'Magician'." In: *Envisioning Magic: A Princeton Seminar & Symposium*. Leiden: BRILL, pp. 115–136 (cit. on p. 114).

Frazer, James George (1911–1915). *The Golden Bough: A study in magic and religion*. 3rd ed. London: Macmillan (cit. on p. 112).

Furtwängler, Adolf (1900). *Die antiken Gemmen*. 3 vols. Leipzig: Giesecke & Devrient (cit. on p. 116).

Gager, John G. (1992). *Curse tablets and binding spells from the ancient world*. New York: Oxford University Press (cit. on pp. 113, 114).

Goldberg, Yoav (2019). "Assessing BERT's Syntactic Abilities." In: *ArXiv*. URL: https://arxiv.org/abs/1901.05287 (cit. on p. 33).

Gonen, Hila and Yoav Goldberg (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 609–614. DOI: 10.18653/v1/N19-1061 (cit. on p. 143).

Gordon, Richard (2002). "Magical amulets in the British Museum." In: *Journal of Roman Archaeology* 15, pp. 666–670. DOI: 10.1017/S1047759400014598 (cit. on p. 116).

Gordon, Richard (2008). "The power of stones: Graeco-Egyptian magical amulets." In: *Journal of Roman Archaeology* 21, pp. 713–718. DOI: 10.1017/S104775940000516X (cit. on p. 116).

Gordon, Richard (2011). "Archaeologies of magical gems." In: *'Gems of Heaven': Recent Research on Engraved Gemstones in Late Antiquity, c. AD 200–600*. London: The British Museum, pp. 34–49 (cit. on pp. 115–117).

Grayson, Siobhán, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene (2016). "Novel2vec: Characterising 19th century fiction via word embeddings." In: *Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2016)*, pp. 68–79. URL: http://ceur-ws.org/Vol-1751/AICS_2016_paper_48.pdf (cit. on p. 13).

Hartmann, Mareike, Yova Kementchedjhieva, and Anders Søgaard (2018). "Why is unsupervised alignment of English embeddings from different algorithms so hard?" In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 582–586. DOI: 10.18653/v1/D18-1056 (cit. on pp. 48, 57).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 131).

He, Ruining and Julian McAuley (2016). "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering." In: *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)*, pp. 507–517. DOI: 10.1145/2872427.2883037 (cit. on p. 24).

Hellrich, Johannes and Udo Hahn (2016). "Bad Company—Neighborhoods in Neural Embedding Spaces Considered Harmful." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 2785–2796. URL: https://www.aclweb.org/anthology/C16-1262 (cit. on p. 57).

Heuser, Ryan (2016). *Word Vectors in the Eighteenth Century*. URL: https://ryanheuser.org/word-vectors (cit. on p. 13).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (cit. on p. 13).

Hofmann, Thomas (1999). "Probabilistic Latent Semantic Analysis." In: *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 289–296. URL: https://arxiv.org/abs/1301.6705 (cit. on p. 85).

Huffer, Damien and Shawn Graham (2018). "Fleshing out the bones: Studying the human remains trade with Tensorflow and Inception." In: *Journal of Computer Applications in Archaeology* 1.1 (cit. on p. 15).

Jockers, Matthew L. (2013). *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press (cit. on pp. 85, 88).

Kementchedjhieva, Yova, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard (2018). "Generalizing Procrustes Analysis for Better Bilingual Dictionary Induction." In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 211–220. DOI: 10.18653/v1/K18-1021 (cit. on p. 48).

Kerr, Sara J. (2017). "When Computer Science Met Austen and Edgeworth." In: *NPPSH Reflections* 1, pp. 38–52. URL: https://mural.maynoothuniversity.ie/8298/ (cit. on p. 13).

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation." In: *MT Summit*. Vol. 5, pp. 79–86. URL: http://www.mt-archive.info/MTS-2005-Koehn.pdf (cit. on p. 44).

Koo, Terry, Xavier Carreras, and Michael Collins (2008). "Simple Semi-supervised Dependency Parsing." In: *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, pp. 595–603. URL: https://www.aclweb.org/anthology/P08-1068 (cit. on p. 44).

Kornblith, Simon, Jonathon Shlens, and Quoc V. Le (2019). "Do Better ImageNet Models Transfer Better?" In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2656–2666. DOI: 10.1109/CVPR.2019.00277 (cit. on p. 109).

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). "Word translation without parallel data." In: *International Conference on Learning Representations 2018*. URL: https://arxiv.org/abs/1710.04087 (cit. on p. 48).

Lancichinetti, Andrea, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding, and Luís A. Nunes Amaral (2015). "High-Reproducibility and High-Accuracy Method for Automated Topic Classification." In: *Physical Review X* 5 (1). DOI: 10.1103/PhysRevX.5.011007 (cit. on p. 20).

Lange, Milan van and Ralf Futselaar (2018). "Debating Evil: Using Word Embeddings to Analyze Parliamentary Debates on War Criminals in The Netherlands." In: *Proceedings of the Conference on Language Technologies & Digital Humanities*, pp. 147–153. URL: https://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Lange-et-al_Debating-evil-Using-Word-Embeddings-to-Analyze-Parliamentary-Debates-on-War-Criminals-in-The-Netherlands.pdf (cit. on p. 13).

Le, Quoc and Tomas Mikolov (2014). "Distributed representations of sentences and documents." In: *Proceedings of the 31st International Conference on Machine Learning*.

PMLR 32, pp. 1188–1196. URL: http://proceedings.mlr.press/v32/le14.html (cit. on p. 12).

Levy, Omer and Yoav Goldberg (2014). "Linguistic Regularities in Sparse and Explicit Word Representations." In: *Proceedings of the 18th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 171–180. DOI: 10.3115/v1/W14-1618 (cit. on p. 53).

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). "Improving Distributional Similarity with Lessons Learned from Word Embeddings." In: *Transactions of the Association for Computational Linguistics* 3, pp. 211–225. DOI: 10.1162/tacl_a_00134 (cit. on pp. 53, 86).

Li, Ping, Trevor J. Hastie, and Kenneth W. Church (2006). "Very Sparse Random Projections." In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp. 287–296. DOI: 10.1145/1150402.1150436 (cit. on p. 34).

Lin, Chu-Cheng, Waleed Ammar, Chris Dyer, and Lori Levin (2015). "Unsupervised POS Induction with Word Embeddings." In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 1311–1316. DOI: 10.3115/v1/N15-1144 (cit. on pp. 43, 53, 54).

Lison, Pierre and Andrey Kutuzov (2017). "Redefining Context Windows for Word Embedding Models: An Experimental Study." In: *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 284–288. URL: https://www.aclweb.org/anthology/W17-0239 (cit. on pp. 55, 57).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa:

A Robustly Optimized BERT Pretraining Approach." In: *ArXiv*. URL: https://arxiv.org/abs/1907.11692 (cit. on pp. 13, 21).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9, pp. 2579–2605. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (cit. on pp. 8, 47).

Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel (2018). "Learning Adversarially Fair and Transferable Representations." In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR 80, pp. 3384–3393. URL: http://proceedings.mlr.press/v80/madras18a.html (cit. on pp. 148, 149).

Mahendran, Aravindh and Andrea Vedaldi (2015). "Understanding deep image representations by inverting them." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5188–5196. DOI: 10.1109/CVPR.2015.7299155 (cit. on p. 14).

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot (2020). "CamemBERT: a Tasty French Language Model." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645 (cit. on p. 21).

Mastrocinque, Attilio (2011). "The colours of magical gems." In: *'Gems of Heaven': Recent Research on Engraved Gemstones in Late Antiquity, c. AD 200–600*. London: The British Museum, pp. 62–68 (cit. on pp. 118, 124, 140).

Mastrocinque, Attilio, ed. (2003–7). *Sylloge Gemmarum Gnosticarum*. 2 vols. Roma: Ist. Poligrafico e Zecca dello Stato, Libreria dello Stato (cit. on p. 124).

McAuley, Julian, Christopher Targett, Qinfeng Shi, and Anton van den Hengel (2015). "Image-based Recommendations on Styles and Substitutes." In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development*

in Information Retrieval. Association for Computing Machinery, pp. 43–52. DOI: 10.1145/2766462.2767755 (cit. on p. 24).

McCallum, Andrew Kachites (2002). *Mallet: A machine learning for language toolkit*. http://mallet.cs.umass.edu (cit. on pp. 11, 26, 89).

McInnes, Leland, John Healy, and James Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. URL: https://arxiv.org/abs/1802.03426 (cit. on pp. 8, 133).

Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher (2017). *Pointer Sentinel Mixture Models*. URL: https://arxiv.org/abs/1609.07843 (cit. on p. 24).

Miao, Yishu, Edward Grefenstette, and Phil Blunsom (2017). "Discovering Discrete Latent Topics with Neural Variational Inference." In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR 70, pp. 2410–2419. URL: http://proceedings.mlr.press/v70/miao17a.html (cit. on p. 23).

Michel, Simone (2004). *Die magischen Gemmen: zu Bildern und Zauberformeln auf geschnittenen Steinen der Antike und Neuzeit*. Berlin: Akademie Verlag (cit. on p. 124).

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013a). "Efficient Estimation of Word Representations in Vector Space." In: *International Conference on Learning Representations 2013*. URL: https://arxiv.org/abs/1301.3781 (cit. on pp. 12, 20, 41).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean (2013b). "Distributed representations of words and phrases and their compositionality." In: *Advances in Neural Information Processing Systems*. Vol. 26, pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf (cit. on pp. 12, 20, 41, 58, 86).

Milli, Smitha and David Bamman (2016). "Beyond Canonical Texts: A Computational Analysis of Fanfiction." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2048–2053. DOI: 10.18653/v1/D16-1218 (cit. on p. 1).

Mimno, David (2011). "Reconstructing Pompeian Households." In: *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pp. 506–513. URL: https://arxiv.org/abs/1202.3747 (cit. on p. 11).

Mimno, David (2012). "Computational Historiography: Data Mining in a Century of Classics Journals." In: *Journal on Computing and Cultural Heritage* 5.1. DOI: 10.1145/2160165.2160168 (cit. on p. 11).

Mimno, David and David Blei (2011). "Bayesian Checking for Topic Models." In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 227–237. URL: https://www.aclweb.org/anthology/D11-1021 (cit. on p. 90).

Mimno, David and Andrew McCallum (2008). "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression." In: *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pp. 411–418. URL: https://arxiv.org/abs/1206.3278 (cit. on p. 87).

Mimno, David and Laure Thompson (2017). "The strange geometry of skip-gram with negative sampling." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2873–2878. DOI: 10.18653/v1/D17-1308 (cit. on p. 51).

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011). "Optimizing Semantic Coherence in Topic Models." In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 262–272. URL: https://www.aclweb.org/anthology/D11-1024 (cit. on pp. 27, 99).

Nagy, Árpád M. (2011). "Magical gems and classical archaeology." In: *'Gems of Heaven': Recent Research on Engraved Gemstones in Late Antiquity, c. AD 200–600*. London: The British Museum, pp. 75–81 (cit. on pp. 115, 117, 124).

Nagy, Árpád M. (2012). "Daktylios pharmakites. Magical healing gems and rings in the Graeco-Roman world." In: *Ritual Healing. Magic, Ritual and Medical Therapy from Antiquity Until the Early Modern Period*. Firenze: SISMEL edizioni del Galluzzo, pp. 71–106 (cit. on pp. 115, 116, 118, 123).

Nagy, Árpád M. (2014). "Étude sur la transmission du savoir magique. L'histoire post-antique du schéma anguipède (Ve–XVIIe siècles)." In: *Les savoirs magiques et leur transmission de l'Antiquité à la Renaissance, Florence*. Firenze: SISMEL edizioni del Galluzzo, pp. 131–155 (cit. on pp. 117, 121, 123).

Nagy, Árpád M. (2015). "Engineering ancient amulets: Magical gems of the Roman Imperial period." In: *The Materiality of Magic*. Paderborn: Wilhelm Fink, pp. 205–240 (cit. on pp. 118, 122).

Nagy, Árpád M. (2019). "Figuring out the Anguipes Gems, *bis*: A Statistical Overview." In: *Magical Gems in Their Contexts: Proceedings of the International Workshop held at the Museum of Fine Arts Budapest, 16–18 February 2012*. Roma: "L'Erma" di Bretschneider, pp. 179–216 (cit. on pp. 121, 124, 126, 138).

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018). "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206 (cit. on p. 23).

Nelson, Robert K. (2010). *Mining the Dispatch*. URL: https://dsl.richmond.edu/dispatch/ (cit. on p. 11).

Newman, David, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin (2010). "Evaluating Topic Models for Digital Libraries." In: *Proceedings of the 10th*

Annual Joint Conference on Digital Libraries. Association for Computing Machinery, pp. 215–224. DOI: 10.1145/1816123.1816156 (cit. on p. 27).

Nguyen, Dat Quoc and Anh Tuan Nguyen (2020). "PhoBERT: Pre-trained language models for Vietnamese." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 1037–1042. DOI: 10.18653/v1/2020.findings-emnlp.92 (cit. on p. 21).

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith (2013). "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 380–390. URL: https://www.aclweb.org/anthology/N13-1039 (cit. on pp. 13, 43).

Paul, Michael (2009). "Cross-Collection Topic Models: Automatically Comparing and Contrasting Text." MA thesis. University of Illinois Urbana-Champaign (cit. on p. 87).

Paul, Michael and Mark Dredze (2012). "Factorial LDA: Sparse Multi-Dimensional Text Models." In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., pp. 2582–2590. URL: https://proceedings.neurips.cc/paper/2012/file/68d13cf26c4b4f4f932e3eff990093ba-Paper.pdf (cit. on p. 87).

Paul, Michael and Roxana Girju (2010). "A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics." In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. Vol. 24, pp. 545–550. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1730 (cit. on p. 87).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine

Learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html (cit. on p. 34).

Peinelt, Nicole, Dong Nguyen, and Maria Liakata (2020). "tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7047–7055. DOI: 10.18653/v1/2020.acl-main.630 (cit. on p. 23).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162 (cit. on pp. 12, 20, 41, 47).

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202 (cit. on p. 13).

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller (2019). "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, pp. 2463–2473. DOI: 10.18653/v1/D19-1250 (cit. on p. 21).

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). "A Universal Part-of-Speech Tagset." In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA),

pp. 2089–2096. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf (cit. on p. 43).

Pierrejean, Bénédicte and Ludovic Tanguy (2018). "Predicting Word Embeddings Variability." In: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, pp. 154–159. DOI: 10.18653/v1/S18-2019 (cit. on p. 50).

Pinter, Yuval, Marc Marone, and Jacob Eisenstein (2019). "Character Eyes: Seeing Language through Character-Level Taggers." In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, pp. 95–102. DOI: 10.18653/v1/W19-4811 (cit. on p. 56).

Porter, J. D. (Sept. 2018). "Pamphlet 17: Popularity/Prestige." In: *Stanford Literary Lab Pamphlet Series*. URL: https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf (cit. on p. 1).

Potts, Alex (2000). *Flesh and the ideal: Winckelmann and the origins of art history*. New Haven: Yale University Press (cit. on p. 116).

Quadrianto, Novi, Viktoriia Sharmanska, and Oliver Thomas (2019). "Discovering Fair Representations in the Data Domain." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8219–8228. DOI: 10.1109/CVPR.2019.00842 (cit. on p. 148).

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners." In: URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (cit. on pp. 13, 21).

Raghu, Maithra, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein (2017). "SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability." In: *Advances in Neural Information Processing Sys-*

tems. Vol. 30. Curran Associates, Inc., pp. 6076–6085. URL: https://proceedings.neurips.cc/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf (cit. on p. 34).

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 248–256. URL: https://www.aclweb.org/anthology/D09-1026 (cit. on p. 87).

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg (2020). "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647 (cit. on p. 144).

Razavian, Ali Sharif, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson (2014). "CNN features off-the-shelf: an astounding baseline for recognition." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519. DOI: 10.1109/CVPRW.2014.131 (cit. on pp. 15, 65, 109).

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim (2019). "Visualizing and Measuring the Geometry of BERT." In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., pp. 8594–8603. URL: https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf (cit. on pp. 22, 32).

Al-Rfou', Rami, Bryan Perozzi, and Steven Skiena (2013). "Polyglot: Distributed Word Representations for Multilingual NLP." In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 183–192. URL: https://www.aclweb.org/anthology/W13-3520 (cit. on pp. 13, 43).

Rhody, Lisa M. (2012). "Topic Modeling and Figurative Language." In: *Journal of Digital Humanities* 2.1. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody (cit. on p. 11).

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand (2014). "Structural Topic Models for Open-Ended Survey Responses." In: *American Journal of Political Science* 58.4, pp. 1064–1082. DOI: https://doi.org/10.1111/ajps.12103 (cit. on p. 87).

Rodriguez, Paul, Alan Craig, Alison Langmead, and Christopher J. Nygren (2020). "Extracting and Analyzing Deep Learning Features for Discriminating Historical Art: Deep Learning Features and Art." In: *PEARC 2020: Practice and Experience in Advanced Research Computing*. Association for Computing Machinery, pp. 358–363. DOI: 10.1145/3311790.3399611 (cit. on p. 15).

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth (2004). "The Author-Topic Model for Authors and Documents." In: *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, pp. 487–494. URL: https://arxiv.org/abs/1207.4169 (cit. on p. 87).

Ross, David A., Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang (2008). "Incremental learning for robust visual tracking." In: *International Journal of Computer Vision* 77, pp. 125–141. DOI: 10.1007/s11263-007-0075-7 (cit. on p. 35).

Sandhaus, Evan (2008). *The New York Times Annotated Corpus LDC2008T19*. Philadelphia: Linguistic Data Consortium. DOI: 10.35111/77ba-9x74 (cit. on p. 27).

Sandy, Heather Moulaison, Heather Froehlich, Cynthia Hudson-Vitale, and Denice Adkins (2019). "Topic Modeling and Facet Analysis of an Emerging Domain: Research Data Management and Data Curation." In: *NASKO: North American Symposium on Knowledge Organization* 7, pp. 63–76. DOI: 10.7152/nasko.v7i1.15623 (cit. on p. 11).

Schachter, Paul and Timothy Shopen (2007). "Parts-of-speech systems." In: *Language Typology and Syntactic Description*. Ed. by Timothy Shopen. 2nd ed. Vol. 1. Cambridge: Cambridge University Press, pp. 1–60. DOI: 10.1017/CBO9780511619427.001 (cit. on pp. 42, 43).

Schmidt, Benjamin M. (2012). "Words Alone: Dismantling Topic Models in the Humanities." In: *Journal of Digital Humanities* 2.1. URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ (cit. on p. 11).

Schnabel, Tobias and Hinrich Schütze (2014). "FLORS: Fast and Simple Domain Adaptation for Part-of-Speech Tagging." In: *Transactions of the Association for Computational Linguistics* 2, pp. 15–26. DOI: 10.1162/tacl_a_00162 (cit. on p. 43).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162 (cit. on p. 25).

Shandruk, Walter Michael (2016). "A Computational Approach to the Study of Magical Gems." PhD thesis. University of Chicago (cit. on pp. 120, 121, 124).

Sia, Suzanna, Ayush Dalmia, and Sabrina J. Mielke (2020). "Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1728–1736. DOI: 10.18653/v1/2020.emnlp-main.135 (cit. on pp. 19–21).

Sims, Matthew, Jong Ho Park, and David Bamman (2019). "Literary Event Detection." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 3623–3634. DOI: 10.18653/v1/P19-1353 (cit. on p. 14).

Smith, Jonathan Z. (1998). "Religion, Religions, Religious." In: *Critical Terms for Religious Studies*. Ed. by Mark C. Taylor. Vol. 1998. Chicago: University of Chicago Press, pp. 269–284 (cit. on p. 112).

Srivastava, Akash and Charles Sutton (2016). "Neural variational inference for topic models." In: *Workshop on Bayesian Deep Learning, NeurIPS 2016*. URL: http://bayesiandeeplearning.org/2016/papers/BDL_27.pdf (cit. on p. 23).

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski, and David Buttler (2012). "Exploring Topic Coherence over Many Models and Many Topics." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 952–961. URL: https://www.aclweb.org/anthology/D12-1087 (cit. on p. 27).

Tambiah, Stanley Jeyaraja (1990). *Magic, Science, Religion and the Scope of Rationality*. Cambridge: Cambridge University Press (cit. on p. 112).

Tang, Gongbo, Gaoqi Rao, Dong Yu, and Endong Xun (2016). "Can We Neglect Function Words in Word Embedding?" In: *Proceedings of the 5th Conference on Natural Language Processing and Chinese Computing & the 24th International Conference on Computer Processing of Oriental Languages*. Springer, pp. 541–548. DOI: 10.1007/978-3-319-50496-4_47 (cit. on p. 57).

Tian, Yingtao, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena (2016). "On the Convergent Properties of Word Embedding Methods." In: *ArXiv*. URL: https://arxiv.org/abs/1605.03956 (cit. on p. 57).

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio (2010). "Word Representations: A Simple and General Method for Semi-Supervised Learning." In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 384–394. URL: https://www.aclweb.org/anthology/P10-1040 (cit. on pp. 13, 41).

Tzara, Tristan (1963). "Pour faire un poème dadaïste." In: *Lampisteries, précédées des Sept manifestes Dada: quelques dessins de Francis Picabia*. Paris: J. J. Pauvert, pp. 64–5 (cit. on p. 63).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on pp. 13, 21).

Versnel, H. S. (1991a). "Beyond Cursing: The Appeal to Justice in Judicial Prayers." In: *Magika Hiera: Ancient Greek Magic and Religion*. New York: Oxford University Press, pp. 60–106 (cit. on p. 114).

Versnel, H. S. (1991b). "Some Reflections on the Relationship Magic-Religion." In: *Numen* 38.2, pp. 177–197 (cit. on p. 112).

Vitellozzi, Paolo (2018). "Relations Between Magical Texts and Magical Gems." In: *Bild und Schrift auf 'magischen' Artefakten*. Berlin: De Gruyter, pp. 181–254. DOI: 10.1515/9783110604337 (cit. on p. 118).

Vulić, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen (2020). "Probing Pretrained Language Models for Lexical Semantics." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 7222–7240. DOI: 10.18653/v1/2020.emnlp-main.586 (cit. on pp. 21, 26).

Walsh, Melanie (2018). "Tweets of a Native Son: The Quotation and Recirculation of James Baldwin from Black Power to# BlackLivesMatter." In: *American Quarterly* 70.3, pp. 531–559. DOI: 10.1353/aq.2018.0034 (cit. on p. 1).

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." In: *Ad-*

vances in Neural Information Processing Systems. Vol. 32. Curran Associates, Inc., pp. 3266–3280. URL: https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf (cit. on p. 31).

Wang, Haiyan, Zhongshi He, Yongwen Huang, Dingding Chen, and Zexun Zhou (2017). "Bodhisattva head images modeling style recognition of Dazu Rock Carvings based on deep convolutional network." In: *Journal of Cultural Heritage* 27, pp. 60–71. DOI: https://doi.org/10.1016/j.culher.2017.03.006 (cit. on p. 15).

Wang, Li, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du (2018). "A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization." In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4453–4460. DOI: 10.24963/ijcai.2018/619 (cit. on p. 23).

Wattenberg, Martin, Fernanda Viégas, and Ian Johnson (2016). "How to Use t-SNE Effectively." In: *Distill*. DOI: 10.23915/distill.00002 (cit. on p. 47).

Wendlandt, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea (2018). "Factors Influencing the Surprising Instability of Word Embeddings." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, pp. 2092–2102. DOI: 10.18653/v1/N18-1190 (cit. on p. 57).

Wiedemann, Gregor, Steffen Remus, Avi Chawla, and Chris Biemann (2019). "Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings." In: *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*. URL: https://arxiv.org/abs/1909.10430 (cit. on pp. 23, 32).

Winckelmann, Johann Joachim (1764). *Geschichte der Kunst des Alterthums*. Dresden: In der Waltherischen Hof-Buchhandlung (cit. on p. 115).

Wiseman, James (2016). "The Gymnasium Area at Corinth, 1969–1970." In: *Hesperia: The Journal of the American School of Classical Studies at Athens* 41.1, pp. 1–42. URL: https://www.jstor.org/stable/147475 (cit. on p. 121).

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." In: *ArXiv*. URL: https://arxiv.org/abs/1910.03771 (cit. on p. 25).

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." In: *ArXiv*. URL: https://arxiv.org/abs/1609.08144 (cit. on p. 25).

Yosinski, Jason, Jeff Clune, Thomas Fuchs, and Hod Lipson (2015). "Understanding neural networks through deep visualization." In: *Deep Learning Workshop at the 31st International Conference on Machine Learning*. URL: https://arxiv.org/abs/1506.06579 (cit. on p. 14).

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and understanding convolutional networks." In: *European Conference on Computer Vision 2014*. Springer, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53 (cit. on p. 14).

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork (2013). "Learning Fair Representations." In: *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. PMLR 28, pp. 325–333. URL: http://proceedings.mlr.press/v28/zemel13.html (cit. on p. 148).