big data and cognitive computing

MDPI

Article

Archetype-Based Modeling and Search of Social Media

Brent D. Davis [1,*], Kamran Sedig [1,2] and Daniel J. Lizotte [1,3]

1 Department of Computer Science, Western University, London, ON N6A 5B7, Canada
2 Faculty of Information & Media Studies, Western University, London, ON N6A 5B9, Canada
3 Department of Biostatistics & Epidemiology, Schulich School of Medicine and Dentistry, Western University, London, ON N6A 5C1, Canada
* Correspondence: bdavis56@uwo.ca

Received: 29 June 2019; Accepted: 22 July 2019; Published: 24 July 2019

Abstract: Existing keyword-based search techniques suffer from limitations owing to unknown, mismatched, and obscure vocabulary. These challenges are particularly prevalent in social media, where slang, jargon, and memetics are abundant. We develop a new technique, Archetype-Based Modeling and Search, that can mitigate these challenges as they are encountered in social media. This technique learns to identify new relevant documents based on a specified set of archetypes from which both vocabulary and relevance information are extracted. We present a case study from the social media data from Reddit, by using authors from /r/Opiates to characterize discourse around opioid use and to find additional relevant authors on this topic.

Keywords: information retrieval; modeling; social media; big data; machine learning; deep learning; natural language processing; word embeddings; opioids

1. Introduction

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

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

the complications of performing pharmacovigilance in social media is that stigmatized populations engaging in illicit substance use face risks ranging from social repercussions to imprisonment.

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

Reddit, a popular social media site, is composed of anonymous, ever-evolving, linguistically difficult-to-analyze online communities of users who share a common interest. Since the site's creation in 2005, Reddit has rapidly increased in size; users have generated over a billion posts and comments in a single month as recently as 2018. These posts contain unstructured text which is valuable to researchers wanting to understand different phenomena on social media. Online communities exist on Reddit as "subreddits", where users can post material and interact with one another. These subreddits have a label that describes each community's focus. Examples include /r/Happy, /r/CasualConversation, and /r/pics, where online participants discuss happiness, facilitate casual conversation, and share photography, respectively. For researchers studying stigmatized topics, there is population-level data in communities such as /r/SuicideWatch, /r/Depression, and /r/Anxiety. Content from these online communities have been used to prototype systems that detect comments containing suicidal ideation [5]. Studying these communities (e.g., Suicide Watch) provides insight into the language surrounding stigmatized topics.

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

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

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

- We establish ABMS for retrieving documents of interest in domains where the language of discourse is not well-understood. This is accomplished by using tailored representations of the language of discourse and by searching using archetypes rather than keywords.

- We provide a concrete example of how ABMS can be applied to big data in social media in order to retrieve authors of interest.
- We explain how the ABMS technique may be extended by incorporating emerging AI methodologies, and we discuss its generalizability to additional domains.

2. Background

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

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

To capture semantic information within documents and, in turn, assess similarity of a document to the archetypes, ABMS uses word embeddings, such as those produced by GloVe [13] or word2vec [14], as a foundation to capture semantic similarity between words. Such embeddings map words to points in a vector space. This is done such that the points are nearby if the corresponding words are semantically similar, where this similarity is learned from a language corpus. An advantage of using embeddings in the manner we do is that the search technique can be tailored to specific domains by using word embeddings that are trained using domain-specific language. For example, ABMS can use word embeddings from multiple languages and from different sources specialized in understanding domain vocabulary. Existing efforts to create repositories for word embeddings [15] support the performance and generality of ABMS by providing a menu of "pre-made" embeddings useful in a variety of domains. Based on its words [16], these word embeddings are then used with a machine learning technique that learns a vector representation for each document in the collection. Having defined vector-based representations for each of the documents, we may develop a machine learning classification model [17] in order to capture the patterns that distinguish the archetypes from "control" documents that are known not to be of interest. This model can then be applied to other documents to identify those that are most likely to be of interest.

The archetype-based approach and specialized representations used by ABMS are particularly useful in the context of social media because of their properties of discourse, which contains population-enriched language with vocabulary, slang, and memetics that are unique to it. Language analysis among social media users shows that different groups use different vocabularies to communicate [18]. These linguistic differences reflect traits of the individuals in these groups. However, there is an ongoing evolution of memes and slang in social media, causing vocabulary and meanings to change [7]. For instance, the nature of pharmaceutical drug chatter on social media has been documented to have changed over time [19]. The variability of words, the evolution of their meanings, and the difficulty of maintaining knowledge of their specifics is referred to as the 'unknown vocabulary problem' [20]. Keyword search techniques are hampered by the unknown vocabulary problem because they rely on users to have a priori and in-depth knowledge of the discourse within communities and how that discourse evolves. Making sense of, and adapting to, the evolving language of social media

in order to use keyword-based search approaches requires considerable time and effort on the part of the user that is often not practical to expend.

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

3. Archetype-Based Modeling and Search

ABMS is an exploratory modeling and search technique that solicits documents, rather than keywords, from a user in order to retrieve additional documents of interest. These solicited documents consist of archetypes, which are themselves documents of interest, and controls, which are documents not of interest. ABMS is especially useful when the specific vocabulary and patterns of discourse that distinguish archetypes from controls are unknown to the user. The key technology underlying ABMS is the development of representations for words and documents that capture these unknowns in a way that enables the retrieval of additional relevant documents. These representations map from the space of words or documents to the space of vectors with a fixed dimension, allowing for the learning of models that distinguish archetypes from controls. Once constructed, these models may be applied to any document, including those that are neither archetypes nor controls, thereby enabling retrieval of new relevant documents. The main steps of ABMS are as follows:

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

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

We describe an algorithm for our ABMS technique below (Algorithm 1).

Figure 1 depicts the ABMS algorithm above, showing the extraction of the archetypes and controls from the document representation corpus and finishing with a ranked list of search results. The user identifies a collection of documents that are of interest (archetypes) and a collection of documents which are not of interest (controls).

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

Algorithm 1 Archetype-Based Modeling and Search

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

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

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

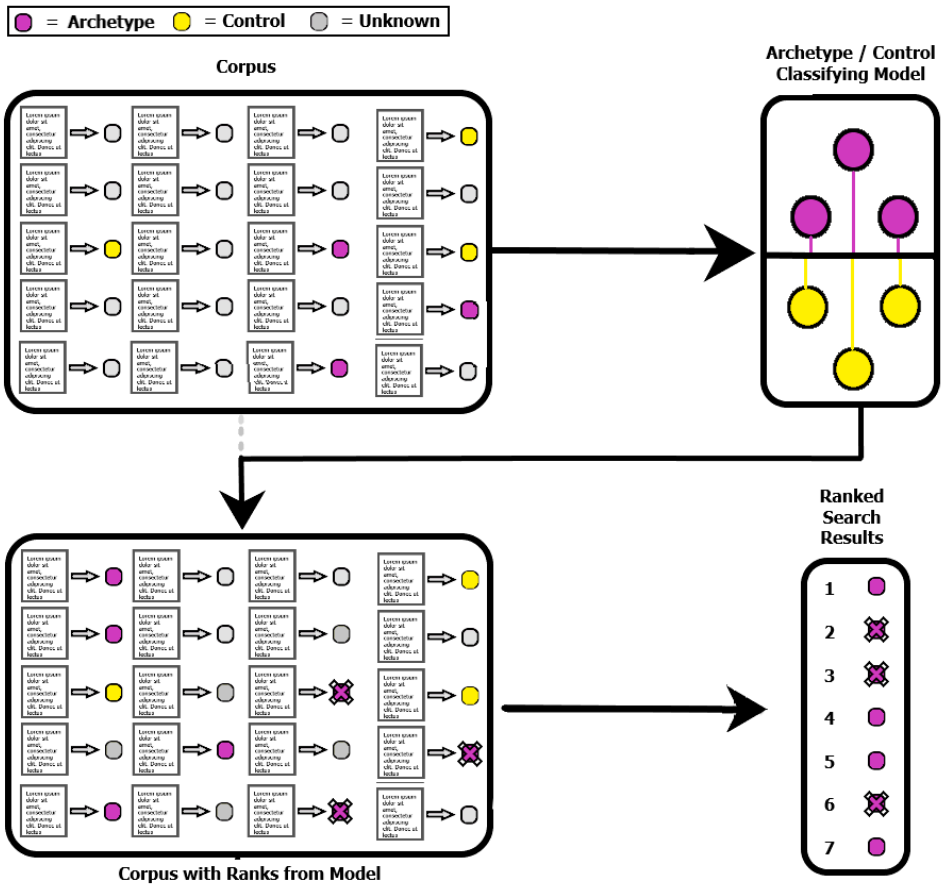

Figure 1. This overview shows the process behind an Archetype-Based Modeling and Search performed on a collection of documents. A word representation is learned, and then using the word representation each document is assigned a vector-based representation. A classification model is learned to distinguish representations of archetypes from representations of controls. The remaining unlabeled documents are then ranked by the model. The resulting scores are sorted and form a ranking for the search results.

4. Case Study: Opioid Dialogue on Social Media

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

4.1. Case Study Methods

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

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

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

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

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

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

Models derived from the Twitter embedding and the /r/Opiates embedding have dimensionality 100 (chosen by the Stanford team) and 300 (typical default setting for GloVe), respectively.

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

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

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

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

4.2. Case Study Results

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

4.2.1. Modeling and Classification

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

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

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

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

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

By taking the decision direction—the vector orthogonal to the separating hyperplane found by the SVM—and comparing it to the original word vectors from the embedding that was used to construct the document representations and classification model, we can see which words have a vector that is most similar to the decision direction. These words are the ones that exert the most influence on the decision value in the positive direction, meaning that their presence in a document increases that document's affinity with the archetypes. We visualize the 200 words most aligned to the decision direction for both the Twitter and /r/Opiates embeddings in Figures 2 and 3 respectively.

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

The vocabulary associated with the decision direction also offers insights into the discourse of /r/Opiates authors. While 'opioid' makes an appearance in the top 200, numerous other drugs are present. A slang form of fentanyl, 'fent', can be observed. Cocaine makes an appearance, as does impulsivity. 'Slinging' makes an appearance, which can be a term used to describe the selling of drugs. Some words defy these intuitive explanations, such as 'bookcase' and 'germaphobe', but their

discovery here allows keyword-base retrieval of these posts so that they can be examined further to determine why these words appear to be important.

We now consider how these observed differences in the models translate into differences in which authors are assigned highest affinity. The decision values for each author assigned by the linear SVMs are shown in Figure 4 for the Twitter embedding and Figure 5 for the /r/Opiates embedding. By comparing the peaks in Figures 4 and 5, we can see that the location of the peaks in decision value, which correspond to the authors with strongest affinity, vary depending on which embedding is used. This is important because it implies that the choice of embedding could have a significant impact on which documents are retrieved by the search process.

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

Figure 2. A word cloud generated from the most aligned words to the decision direction separating author vectors from /r/Opiates and /r/CasualConversation. This was generated using the restricted Stanford Twitter GloVe embedding with a 20 thousand size vocabulary.

Figure 3. A word cloud generated from the most aligned words to the decision direction separating author vectors from /r/Opiates and /r/CasualConversation. This was generated using the GloVe embedding based on /r/Opiates authors, with a 78 thousand size vocabulary.

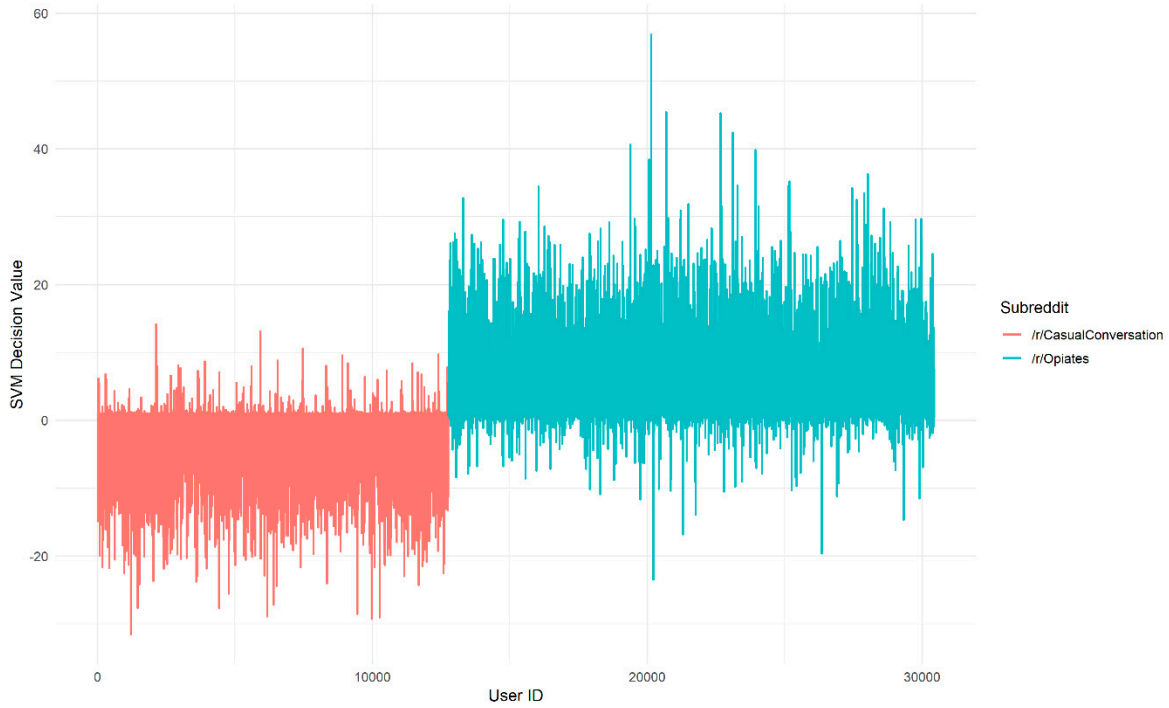

Figure 4. Decision values produced by a Linear SVM for author vectors trained with the Twitter GloVe embedding, colored by subreddit of origin. A large magnitude of decision values corresponds to a large distance from the decision boundary of the SVM. The x-axis is an arbitrary user ID assigned to an author vector from /r/CasualConversation (left) or /r/Opiates (right).

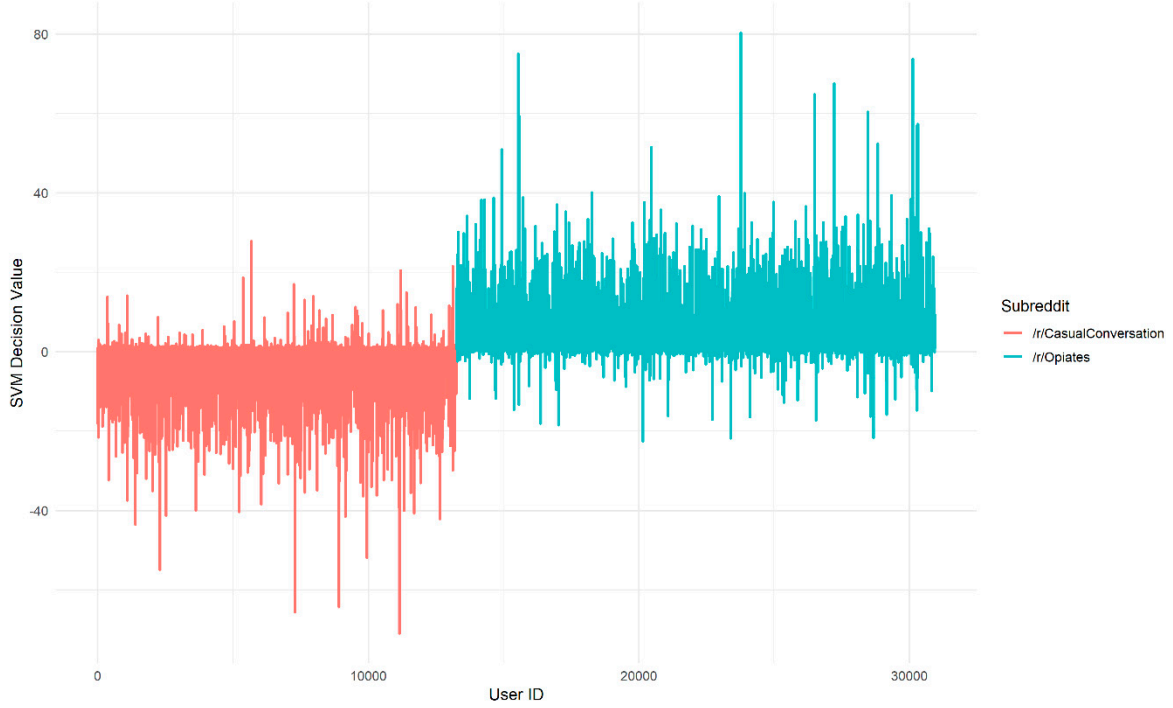

Figure 5. Decision values produced by a Linear SVM for author vectors trained with the /r/Opiates GloVe embedding, colored by subreddit of origin. A large magnitude of decision values corresponds to a large distance from the decision boundary of the SVM. The x-axis is an arbitrary user ID assigned to an author vector from /r/CasualConversation (left) or /r/Opiates (right).

4.2.2. Step 4 Results—Exploratory Search

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

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

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

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

5. Discussion

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

5.1. Scalability of ABMS

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

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

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

5.2. Generalizability of ABMS

Another characteristic of the ABMS technique is its generalizability—that is, it can be used across different social media domains, as long as there is user content in written form. For example, ABMS could be applied to Twitter data as follows. First, all the Twitter activity that contains a hashtag of interest could be collected and used to make a word embedding. This embedding could then be used to develop author representations based on each author's tweet history. Archetypes and controls could be defined in different ways—for example, archetypes could be all authors who used the hashtag, and controls could be a subset of those who did not, or who used another hashtag. Alternatively, to identify users with an affinity to a given geography, archetypes could be Twitter users who have a particular geotag included in their tweet metadata, and controls could be a subset who did not, or who have a different geotag.

ABMS could also be applied to Facebook. Facebook authors provide many possible labels from their activity on the site that could identify subgroups that use specialized language, and that could be used to define archetype/control status. For example, authors organize themselves into groups, like pages based on their interest, and even take personality quizzes, which are potentially shared with their friends. The textual activity within these groups could be collected to build a word embedding for representation learning. Authors that are in a group can be considered archetypes for performing

ABMS. Facebook authors can provide more detailed geographic information on their profiles if they so choose, including both hometown and current residence, and some authors provide information such as cell phone numbers that contain area codes which help with identifying geography. A characteristic of ABMS is that it can use any of this information to define archetypes and controls and, in turn, to enable exploratory search.

5.3. Transferability of ABMS Models Across Social Media Platforms

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

When transferring models from a source platform to a destination platform, it is important to use the same representations for words and authors in the source and the target. If author representations in the target platform are not developed using the same techniques as for the source—that is, using the same word embedding and representation training procedure—they will not work properly with the learned model. The potentially new vocabulary used in other social media can present complications for transferring models. Vocabulary irrelevant to the user's desired search results can still occur disproportionately in one group over another. If this should happen, however, adjusting some of the labels and adding more examples to the control set can help tune ABMS to search more effectively. One of the ways that the locally-trained embeddings are helpful is that, when using techniques like usr2vec, any new words not in the local word embedding's vocabulary are discarded. While this lowers the number of total words that can be used to construct the document representation, it can increase the specificity of the search.

5.4. Adaptability of ABMS to Different Tasks

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

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

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

continue to improve. As advancements are made in word and document representation learning, the ability to perform increasingly specialized search tasks will become possible.

5.5. Ethics of ABMS and Related Techniques

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

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

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

5.6. Biases in Word Embeddings and Their Effects on ABMS

Word embeddings have their own non-technical problems. An examination of the relationships found in word embeddings has shown that they can contain sexist and other biases which can be difficult to remove [33]. As our model is ultimately using the vocabulary of these subreddits to model their associations, there are important implications to its sensitivity. Any differences in language between communities—for example, the presence of Spanish or French text in /r/Opiates but not in /r/CasualConversation—are going to be detected by the model if it gives it a reasonable boost in classification performance. If a model derived from such data were used naively without adjusting for this kind of phenomenon, the model could become biased toward labeling any users engaging in French or Spanish dialogue as originating from /r/Opiates. This can be seen to some extent in the extreme decision values we noted in our results, where political leaning was given some weighting in the decision between classes, and extreme results were pulled to extremes so far from the decision boundary that their other dialogue made little contribution.

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

6. Conclusions and Future Work

We present Archetype-Based Modeling and Search, a technique for exploratory search of large corpora when keywords are not well-understood because of highly complex, evolving discourse with population-specific vocabulary. We demonstrate from our case study that this technique can model complex behaviors present in the author's representations. By performing ABMS, we can rank new

authors based on their similarities to these archetypal authors. We anticipate that the ABMS technique will find increased use as the volume and relevance of social media data increases further, providing an ever-widening window into text-rich human activity.

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

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

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

Author Contributions: Conceptualization, B.D.D.; Data curation, B.D.D. and D.J.L.; Formal analysis, B.D.D. and D.J.L.; Investigation, B.D.D.; Methodology, B.D.D.; Project administration, K.S. and D.J.L.; Resources, K.S. and D.J.L.; Software, B.D.D.; Supervision, K.S. and D.J.L.; Validation, D.J.L.; Visualization, B.D.D. and K.S.; Writing—original draft, B.D.D.; Writing—reviewing and editing, K.S. and D.J.L.

Funding: This research received no external funding.

Acknowledgments: We would like to thank Compute Canada for the computing infrastructure that made running our experiments possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stieglitz, S.; Mirbabaie, M.; Ross, B.; Neuberger, C. Social media analytics—Challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manag.* **2018**, *39*, 156–168.
2. Gruzd, A.; Jacobson, J.; Mai, P.; Dubois, E. The State of Social Media in Canada 2017. *SSRN Electron. J.* **2018**. [CrossRef]
3. Wang, Y.C.; DeSalvo, K. Timely, Granular, and Actionable: Informatics in the Public Health 3.0 Era. *Am. J. Public Health* **2018**, *108*, 930–934. [PubMed]
4. Sarker, A.; Ginn, R.; Nikfarjam, A.; O'Connor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; Gonzalez, G. Utilizing social media data for pharmacovigilance: A review. *J. Biomed. Inform.* **2015**, *54*, 202–212. [PubMed]
5. Aladağ, A.E.; Muderrisoglu, S.; Akbas, N.B.; Zahmacioglu, O.; Bingol, H.O. Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *J. Med. Internet Res.* **2018**, *20*, e215. [PubMed]
6. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.Z. Big Data for Health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1193–1208. [PubMed]
7. Rosin, G.D.; Adar, E.; Radinsky, K. Learning Word Relatedness over Time. *arXiv* **2017**, arXiv:1707.08081.
8. Marchionini, G. Exploratory search: From finding to understanding. *Commun. ACM* **2006**, *49*, 41–46.
9. ARCHETYPE N. *OED Online*; Oxford University Press: Oxford, UK, 2019.
10. Cutler, A.; Breiman, L. Archetypal Analysis. *Technometrics* **1994**, *36*, 338–347.

11. Chen, Y.; Mairal, J.; Harchaoui, Z. Fast and Robust Archetypal Analysis for Representation Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
12. Mørup, M.; Hansen, L.K. Archetypal analysis for machine learning and data mining. *Neurocomputing* **2012**, *80*, 54–63.
13. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014.
14. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*; Mit Press: Cambridge, MA, USA, 2013.
15. Fares, M.; Kutuzov, A.; Oepen, S.; Velldal, E. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, Gothenburg, Sweden, 22–24 May 2017.
16. Amir, S.; Coppersmith, G.; Carvalho, P.; Silva, M.J.; Wallace, B.C. Quantifying Mental Health from Social Media with Neural User Embeddings. *arXiv* **2017**, arXiv:1705.00335.
17. Camacho-Collados, J.; Pilehvar, M.T. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *J. Artif. Intell. Res.* **2018**, *63*, 743–788.
18. Vessey, R.; Zappavigna, M. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*; Springer: London, UK, 2015.
19. Wiley, M.T.; Jin, C.; Hristidis, V.; Esterling, K.M. Pharmaceutical drugs chatter on Online Social Networks. *J. Biomed. Inform.* **2014**, *49*, 245–254. [PubMed]
20. Furnas, G.W.; Landauer, T.K.; Gomez, L.M.; Dumais, S.T. The vocabulary problem in human-system communication. *Commun. ACM* **1987**, *30*, 964–971.
21. Selivanov, D.; Wang, Q. text2vec: Modern Text Mining Framework for R. Computer Software Manual (R Package Version 0.4.0). Available online: https://CRAN.R-project.org/package=text2vec (accessed on 14 June 2019).
22. Gaffney, D.; Matias, J.N. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLoS ONE* **2018**, *13*, e0200162.
23. Amir, S.; Wallace, B.C.; Lyu, H.; Carvalho, P.; Silva, M.J. Modelling Context with User Embeddings for Sarcasm Detection in Social Media. *arXiv* **2016**, arXiv:1607.00976.
24. Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Maintainer, A.W. *The e1071 Package*; Misc Functions of Department of Statistics: Vienna, Austria, 2005.
25. Foley, S.; Karlsen, J.R.; Putniņš, T.J. Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies? *Rev. Financ. Stud.* **2019**, *32*, 1798–1853.
26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
27. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* **2019**, arXiv:1906.08237.
28. Dai, X.; Bikdash, M.; Meyer, B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon*; IEEE: Piscataway, NJ, USA, 2017.
29. Do, T.H.; Nguyen, D.M.; Tsiligianni, E.; Cornelis, B.; Deligiannis, N. Multiview Deep Learning for Predicting Twitter Users' Location. *arXiv* **2017**, arXiv:1712.08091.
30. Ge, M.; Bangui, H.; Buhnova, B. Big Data for Internet of Things: A Survey. *Future Gener. Comput. Syst.* **2018**, *87*, 601–614.
31. Rui, Z.; Yan, Z. A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification. *IEEE Access* **2019**, *7*, 5994–6009.
32. Tariq, N.; Asim, M.; Al-Obeidat, F.; Zubair Farooqi, M.; Baker, T.; Hammoudeh, M.; Ghafir, I. The Security of Big Data in Fog-Enabled IoT Applications Including Blockchain: A Survey. *Sensors* **2019**, *19*, 1788.
33. Gonen, H.; Goldberg, Y. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv* **2019**, arXiv:1903.03862.